Apostolos Antonacopoulos ·
Subhasis Chaudhuri · Rama Chellappa ·
Cheng-Lin Liu · Saumik Bhattacharya ·
Umapada Pal (Eds.)

LNCS 15311

Pattern Recognition

27th International Conference, ICPR 2024
Kolkata, India, December 1–5, 2024
Proceedings, Part XI

Springer

MOREMEDIA

Lecture Notes in Computer Science **15311**

The series Lecture Notes in Computer Science (LNCS), including its subseries Lecture Notes in Artificial Intelligence (LNAI) and Lecture Notes in Bioinformatics (LNBI), has established itself as a medium for the publication of new developments in computer science and information technology research, teaching, and education.

LNCS enjoys close cooperation with the computer science R & D community, the series counts many renowned academics among its volume editors and paper authors, and collaborates with prestigious societies. Its mission is to serve this international community by providing an invaluable service, mainly focused on the publication of conference and workshop proceedings and postproceedings. LNCS commenced publication in 1973.

Apostolos Antonacopoulos ·
Subhasis Chaudhuri · Rama Chellappa ·
Cheng-Lin Liu · Saumik Bhattacharya ·
Umapada Pal
Editors

Pattern Recognition

27th International Conference, ICPR 2024
Kolkata, India, December 1–5, 2024
Proceedings, Part XI

Springer

Editors
Apostolos Antonacopoulos
University of Salford
Salford, Lancashire, UK

Subhasis Chaudhuri
Indian Institute of Technology Bombay
Mumbai, Maharashtra, India

Rama Chellappa
Johns Hopkins University
Baltimore, MD, USA

Cheng-Lin Liu
Chinese Academy of Sciences
Beijing, China

Saumik Bhattacharya
IIT Kharagpur
Kharagpur, West Bengal, India

Umapada Pal
Indian Statistical Institute Kolkata
Kolkata, West Bengal, India

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-031-78194-0 ISBN 978-3-031-78195-7 (eBook)
https://doi.org/10.1007/978-3-031-78195-7

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

President's Address

On behalf of the Executive Committee of the International Association for Pattern Recognition (IAPR), I am pleased to welcome you to the 27th International Conference on Pattern Recognition (ICPR 2024), the main scientific event of the IAPR.

After a completely digital ICPR in the middle of the COVID pandemic and the first hybrid version in 2022, we can now enjoy a fully back-to-normal ICPR this year. I look forward to hearing inspirational talks and keynotes, catching up with colleagues during the breaks and making new contacts in an informal way. At the same time, the conference landscape has changed. Hybrid meetings have made their entrance and will continue. It is exciting to experience how this will influence the conference. Planning for a major event like ICPR must take place over a period of several years. This means many decisions had to be made under a cloud of uncertainty, adding to the already large effort needed to produce a successful conference. It is with enormous gratitude, then, that we must thank the team of organizers for their hard work, flexibility, and creativity in organizing this ICPR. ICPR always provides a wonderful opportunity for the community to gather together. I can think of no better location than Kolkata to renew the bonds of our international research community.

Each ICPR is a bit different owing to the vision of its organizing committee. For 2024, the conference has six different tracks reflecting major themes in pattern recognition: Artificial Intelligence, Pattern Recognition and Machine Learning; Computer and Robot Vision; Image, Speech, Signal and Video Processing; Biometrics and Human Computer Interaction; Document Analysis and Recognition; and Biomedical Imaging and Bioinformatics. This reflects the richness of our field. ICPR 2024 also features two dozen workshops, seven tutorials, and 15 competitions; there is something for everyone. Many thanks to those who are leading these activities, which together add significant value to attending ICPR, whether in person or virtually. Because it is important for ICPR to be as accessible as possible to colleagues from all around the world, we are pleased that the IAPR, working with the ICPR organizers, is continuing our practice of awarding travel stipends to a number of early-career authors who demonstrate financial need. Last but not least, we are thankful to the Springer LNCS team for their effort to publish these proceedings.

Among the presentations from distinguished keynote speakers, we are looking forward to the three IAPR Prize Lectures at ICPR 2024. This year we honor the achievements of Tin Kam Ho (IBM Research) with the IAPR's most prestigious King-Sun Fu Prize "for pioneering contributions to multi-classifier systems, random decision forests, and data complexity analysis". The King-Sun Fu Prize is given in recognition of an outstanding technical contribution to the field of pattern recognition. It honors the memory of Professor King-Sun Fu who was instrumental in the founding of IAPR, served as its first president, and is widely recognized for his extensive contributions to the field of pattern recognition.

The Maria Petrou Prize is given to a living female scientist/engineer who has made substantial contributions to the field of Pattern Recognition and whose past contributions, current research activity and future potential may be regarded as a model to both aspiring and established researchers. It honours the memory of Professor Maria Petrou as a scientist of the first rank, and particularly her role as a pioneer for women researchers. This year, the Maria Petrou Prize is given to Guoying Zhao (University of Oulu), "for contributions to video analysis for facial micro-behavior recognition and remote biosignal reading (RPPG) for heart rate analysis and face anti-spoofing".

The J.K. Aggarwal Prize is given to a young scientist who has brought a substantial contribution to a field that is relevant to the IAPR community and whose research work has had a major impact on the field. Professor Aggarwal is widely recognized for his extensive contributions to the field of pattern recognition and for his participation in IAPR's activities. This year, the J.K. Aggarwal Prize goes to Xiaolong Wang (UC San Diego) "for groundbreaking contributions to advancing visual representation learning, utilizing self-supervised and attention-based models to establish fundamental frameworks for creating versatile, general-purpose pattern recognition systems".

During the conference we will also recognize 21 new IAPR Fellows selected from a field of very strong candidates. In addition, a number of Best Scientific Paper and Best Student Paper awards will be presented, along with the Best Industry Related Paper Award and the Piero Zamperoni Best Student Paper Award. Congratulations to the recipients of these very well-deserved awards!

I would like to close by again thanking everyone involved in making ICPR 2024 a tremendous success; your hard work is deeply appreciated. These thanks extend to all who chaired the various aspects of the conference and the associated workshops, my ExCo colleagues, and the IAPR Standing and Technical Committees. Linda O'Gorman, the IAPR Secretariat, deserves special recognition for her experience, historical perspective, and attention to detail when it comes to supporting many of the IAPR's most important activities. Her tasks became so numerous that she recently got support from Carolyn Buckley (layout, newsletter), Ugur Halici (ICPR matters), and Rosemary Stramka (secretariat). The IAPR website got a completely new design. Ed Sobczak has taken care of our web presence for so many years already. A big thank you to all of you!

This is, of course, the 27th ICPR conference. Knowing that ICPR is organized every two years, and that the first conference in the series (1973!) pre-dated the formal founding of the IAPR by a few years, it is also exciting to consider that we are celebrating over 50 years of ICPR and at the same time approaching the official IAPR 50th anniversary in 2028: you'll get all information you need at ICPR 2024. In the meantime, I offer my thanks and my best wishes to all who are involved in supporting the IAPR throughout the world.

September 2024

Arjan Kuijper
President of the IAPR

Preface

It is our great pleasure to welcome you to the proceedings of the 27th International Conference on Pattern Recognition (ICPR 2024), held in Kolkata, India. The city, formerly known as 'Calcutta', is the home of the fabled Indian Statistical Institute (ISI), which has been at the forefront of statistical pattern recognition for almost a century. Concepts like the Mahalanobis distance, Bhattacharyya bound, Cramer–Rao bound, and Fisher–Rao metric were invented by pioneers associated with ISI. The first ICPR (called IJCPR then) was held in 1973, and the second in 1974. Subsequently, ICPR has been held every other year. The International Association for Pattern Recognition (IAPR) was founded in 1978 and became the sponsor of the ICPR series. Over the past 50 years, ICPR has attracted huge numbers of scientists, engineers and students from all over the world and contributed to advancing research, development and applications in pattern recognition technology.

ICPR 2024 was held at the Biswa Bangla Convention Centre, one of the largest such facilities in South Asia, situated just 7 kilometers from Kolkata Airport (CCU). According to ChatGPT "Kolkata is often called the 'Cultural Capital of India'. The city has a deep connection to literature, music, theater, and art. It was home to Nobel laureate Rabindranath Tagore, and the Bengali film industry has produced globally renowned filmmakers like Satyajit Ray. The city boasts remarkable colonial architecture, with landmarks like Victoria Memorial, Howrah Bridge, and the Indian Museum (the oldest and largest museum in India). Kolkata's streets are dotted with old mansions and buildings that tell stories of its colonial past. Walking through the city can feel like stepping back into a different era. Finally, Kolkata is also known for its street food."

ICPR 2024 followed a two-round paper submission format. We received a total of 2135 papers (1501 papers in round-1 submissions, and 634 papers in round-2 submissions). Each paper, on average, received 2.84 reviews, in single-blind mode. For the first-round papers we had a rebuttal option available to authors.

In total, 945 papers (669 from round-1 and 276 from round-2) were accepted for presentation, resulting in an acceptance rate of 44.26%, which is consistent with previous ICPR events. At ICPR 2024 the papers were categorized into six tracks: Artificial Intelligence, Machine Learning for Pattern Analysis; Computer Vision and Robotic Perception; Image, Video, Speech, and Signal Analysis; Biometrics and Human-Machine Interaction; Document and Media Analysis; and Biomedical Image Analysis and Informatics.

The main conference ran over December 2–5, 2024. The main program included the presentation of 188 oral papers (19.89% of the accepted papers), 757 poster papers and 12 competition papers (out of 15 submitted). A total 10 oral sessions were held concurrently in four meeting rooms with a total of 40 oral sessions. In total 24 workshops and 7 tutorials were held on December 1, 2024.

The plenary sessions included three prize lectures and three invited presentations. The prize lectures were delivered by Tin Kam Ho (IBM Research, USA; King Sun

Fu Prize winner), Xiaolong Wang (University of California, San Diego, USA; J.K. Aggarwal Prize winner), and Guoying Zhao (University of Oulu, Finland; Maria Petrou Prize winner). The invited speakers were Timothy Hospedales (University of Edinburgh, UK), Venu Govindaraju (University at Buffalo, USA), and Shuicheng Yan (Skywork AI, Singapore).

Several best paper awards were presented in ICPR: the Piero Zamperoni Award for the best paper authored by a student, the BIRPA Best Industry Related Paper Award, and the Best Paper Awards and Best Student Paper Awards for each of the six tracks of ICPR 2024.

The organization of such a large conference would not be possible without the help of many volunteers. Our special gratitude goes to the Program Chairs (Apostolos Antonacopoulos, Subhasis Chaudhuri, Rama Chellappa and Cheng-Lin Liu), for their leadership in organizing the program. Thanks to our Publication Chairs (Ananda S. Chowdhury and Wataru Ohyama) for handling the overwhelming workload of publishing the conference proceedings. We also thank our Competition Chairs (Richard Zanibbi, Lianwen Jin and Laurence Likforman-Sulem) for arranging 12 important competitions as part of ICPR 2024. We are thankful to our Workshop Chairs (P. Shivakumara, Stephanie Schuckers, Jean-Marc Ogier and Prabir Bhattacharya) and Tutorial Chairs (B.B. Chaudhuri, Michael R. Jenkin and Guoying Zhao) for arranging the workshops and tutorials on emerging topics. ICPR 2024, for the first time, held a Doctoral Consortium. We would like to thank our Doctoral Consortium Chairs (Véronique Eglin, Dan Lopresti and Mayank Vatsa) for organizing it.

Thanks go to the Track Chairs and the meta reviewers who devoted significant time to the review process and preparation of the program. We also sincerely thank the reviewers who provided valuable feedback to the authors.

Finally, we acknowledge the work of other conference committee members, like the Organizing Chairs and Organizing Committee Members, Finance Chairs, Award Chair, Sponsorship Chairs, and Exhibition and Demonstration Chairs, Visa Chair, Publicity Chairs, and Women in ICPR Chairs, whose efforts made this event successful. We also thank our event manager Alpcord Network for their help.

We hope that all the participants found the technical program informative and enjoyed the sights, culture and cuisine of Kolkata.

October 2024

Umapada Pal
Josef Kittler
Anil Jain

Organization

General Chairs

Umapada Pal	Indian Statistical Institute, Kolkata, India
Josef Kittler	University of Surrey, UK
Anil Jain	Michigan State University, USA

Program Chairs

Apostolos Antonacopoulos	University of Salford, UK
Subhasis Chaudhuri	Indian Institute of Technology, Bombay, India
Rama Chellappa	Johns Hopkins University, USA
Cheng-Lin Liu	Institute of Automation, Chinese Academy of Sciences, China

Publication Chairs

Ananda S. Chowdhury	Jadavpur University, India
Wataru Ohyama	Tokyo Denki University, Japan

Competition Chairs

Richard Zanibbi	Rochester Institute of Technology, USA
Lianwen Jin	South China University of Technology, China
Laurence Likforman-Sulem	Télécom Paris, France

Workshop Chairs

P. Shivakumara	University of Salford, UK
Stephanie Schuckers	Clarkson University, USA
Jean-Marc Ogier	Université de la Rochelle, France
Prabir Bhattacharya	Concordia University, Canada

Tutorial Chairs

B. B. Chaudhuri	Indian Statistical Institute, Kolkata, India
Michael R. Jenkin	York University, Canada
Guoying Zhao	University of Oulu, Finland

Doctoral Consortium Chairs

Véronique Eglin	CNRS, France
Daniel P. Lopresti	Lehigh University, USA
Mayank Vatsa	Indian Institute of Technology, Jodhpur, India

Organizing Chairs

Saumik Bhattacharya	Indian Institute of Technology, Kharagpur, India
Palash Ghosal	Sikkim Manipal University, India

Organizing Committee

Santanu Phadikar	West Bengal University of Technology, India
SK Md Obaidullah	Aliah University, India
Sayantari Ghosh	National Institute of Technology Durgapur, India
Himadri Mukherjee	West Bengal State University, India
Nilamadhaba Tripathy	Clarivate Analytics, USA
Chayan Halder	West Bengal State University, India
Shibaprasad Sen	Techno Main Salt Lake, India

Finance Chairs

Kaushik Roy	West Bengal State University, India
Michael Blumenstein	University of Technology Sydney, Australia

Awards Committee Chair

Arpan Pal	Tata Consultancy Services, India

Sponsorship Chairs

P. J. Narayanan	Indian Institute of Technology, Hyderabad, India
Yasushi Yagi	Osaka University, Japan
Venu Govindaraju	University at Buffalo, USA
Alberto Bel Bimbo	Università di Firenze, Italy

Exhibition and Demonstration Chairs

Arjun Jain	FastCode AI, India
Agnimitra Biswas	National Institute of Technology, Silchar, India

International Liaison, Visa Chair

Balasubramanian Raman	Indian Institute of Technology, Roorkee, India

Publicity Chairs

Dipti Prasad Mukherjee	Indian Statistical Institute, Kolkata, India
Bob Fisher	University of Edinburgh, UK
Xiaojun Wu	Jiangnan University, China

Women in ICPR Chairs

Ingela Nystrom	Uppsala University, Sweden
Alexandra B. Albu	University of Victoria, Canada
Jing Dong	Institute of Automation, Chinese Academy of Sciences, China
Sarbani Palit	Indian Statistical Institute, Kolkata, India

Event Manager

Alpcord Network

Track Chairs – Artificial Intelligence, Machine Learning for Pattern Analysis

Larry O'Gorman	Nokia Bell Labs, USA
Dacheng Tao	University of Sydney, Australia
Petia Radeva	University of Barcelona, Spain
Susmita Mitra	Indian Statistical Institute, Kolkata, India
Jiliang Tang	Michigan State University, USA

Track Chairs – Computer and Robot Vision

C. V. Jawahar	International Institute of Information Technology (IIIT), Hyderabad, India
João Paulo Papa	São Paulo State University, Brazil
Maja Pantic	Imperial College London, UK
Gang Hua	Dolby Laboratories, USA
Junwei Han	Northwestern Polytechnical University, China

Track Chairs – Image, Speech, Signal and Video Processing

P. K. Biswas	Indian Institute of Technology, Kharagpur, India
Shang-Hong Lai	National Tsing Hua University, Taiwan
Hugo Jair Escalante	INAOE, CINVESTAV, Mexico
Sergio Escalera	Universitat de Barcelona, Spain
Prem Natarajan	University of Southern California, USA

Track Chairs – Biometrics and Human Computer Interaction

Richa Singh	Indian Institute of Technology, Jodhpur, India
Massimo Tistarelli	University of Sassari, Italy
Vishal Patel	Johns Hopkins University, USA
Wei-Shi Zheng	Sun Yat-sen University, China
Jian Wang	Snap, USA

Track Chairs – Document Analysis and Recognition

Xiang Bai	Huazhong University of Science and Technology, China
David Doermann	University at Buffalo, USA
Josep Llados	Universitat Autònoma de Barcelona, Spain
Mita Nasipuri	Jadavpur University, India

Track Chairs – Biomedical Imaging and Bioinformatics

Jayanta Mukhopadhyay	Indian Institute of Technology, Kharagpur, India
Xiaoyi Jiang	Universität Münster, Germany
Seong-Whan Lee	Korea University, Korea

Metareviewers (Conference Papers and Competition Papers)

Wael Abd-Almageed	University of Southern California, USA
Maya Aghaei	NHL Stenden University, Netherlands
Alireza Alaei	Southern Cross University, Australia
Rajagopalan N. Ambasamudram	Indian Institute of Technology, Madras, India
Suyash P. Awate	Indian Institute of Technology, Bombay, India
Inci M. Baytas	Bogazici University, Turkey
Aparna Bharati	Lehigh University, USA
Brojeshwar Bhowmick	Tata Consultancy Services, India
Jean-Christophe Burie	University of La Rochelle, France
Gustavo Carneiro	University of Surrey, UK
Chee Seng Chan	Universiti Malaya, Malaysia
Sumohana S. Channappayya	Indian Institute of Technology, Hyderabad, India
Dongdong Chen	Microsoft, USA
Shengyong Chen	Tianjin University of Technology, China
Jun Cheng	Institute for Infocomm Research, A*STAR, Singapore
Albert Clapés	University of Barcelona, Spain
Oscar Dalmau	Center for Research in Mathematics, Mexico

Tyler Derr	Vanderbilt University, USA
Abhinav Dhall	Indian Institute of Technology, Ropar, India
Bo Du	Wuhan University, China
Yuxuan Du	University of Sydney, Australia
Ayman S. El-Baz	University of Louisville, USA
Francisco Escolano	University of Alicante, Spain
Siamac Fazli	Nazarbayev University, Kazakhstan
Jianjiang Feng	Tsinghua University, China
Gernot A. Fink	TU Dortmund University, Germany
Alicia Fornes	CVC, Spain
Junbin Gao	University of Sydney, Australia
Yan Gao	Amazon, USA
Yongsheng Gao	Griffith University, Australia
Caren Han	University of Melbourne, Australia
Ran He	Institute of Automation, Chinese Academy of Sciences, China
Tin Kam Ho	IBM, USA
Di Huang	Beihang University, China
Kaizhu Huang	Duke Kunshan University, China
Donato Impedovo	University of Bari, Italy
Julio Jacques	University of Barcelona and Computer Vision Center, Spain
Lianwen Jin	South China University of Technology, China
Wei Jin	Emory University, USA
Danilo Samuel Jodas	São Paulo State University, Brazil
Manjunath V. Joshi	DA-IICT, India
Jayashree Kalpathy-Cramer	Massachusetts General Hospital, USA
Dimosthenis Karatzas	Computer Vision Centre, Spain
Hamid Karimi	Utah State University, USA
Baiying Lei	Shenzhen University, China
Guoqi Li	Chinese Academy of Sciences, and Peng Cheng Lab, China
Laurence Likforman-Sulem	Institut Polytechnique de Paris/Télécom Paris, France
Aishan Liu	Beihang University, China
Bo Liu	Bytedance, USA
Chen Liu	Clarkson University, USA
Cheng-Lin Liu	Institute of Automation, Chinese Academy of Sciences, China
Hongmin Liu	University of Science and Technology Beijing, China
Hui Liu	Michigan State University, USA

Jing Liu	Institute of Automation, Chinese Academy of Sciences, China
Li Liu	University of Oulu, Finland
Qingshan Liu	Nanjing University of Posts and Telecommunications, China
Adrian P. Lopez-Monroy	Centro de Investigacion en Matematicas AC, Mexico
Daniel P. Lopresti	Lehigh University, USA
Shijian Lu	Nanyang Technological University, Singapore
Yong Luo	Wuhan University, China
Andreas K. Maier	FAU Erlangen-Nuremberg, Germany
Davide Maltoni	University of Bologna, Italy
Hong Man	Stevens Institute of Technology, USA
Lingtong Min	Northwestern Polytechnical University, China
Paolo Napoletano	University of Milano-Bicocca, Italy
Kamal Nasrollahi	Milestone Systems, Aalborg University, Denmark
Marcos Ortega	University of A Coruña, Spain
Shivakumara Palaiahnakote	University of Salford, UK
P. Jonathon Phillips	NIST, USA
Filiberto Pla	University Jaume I, Spain
Ajit Rajwade	Indian Institute of Technology, Bombay, India
Shanmuganathan Raman	Indian Institute of Technology, Gandhinagar, India
Imran Razzak	UNSW, Australia
Beatriz Remeseiro	University of Oviedo, Spain
Gustavo Rohde	University of Virginia, USA
Partha Pratim Roy	Indian Institute of Technology, Roorkee, India
Sanjoy K. Saha	Jadavpur University, India
Joan Andreu Sánchez	Universitat Politècnica de València, Spain
Claudio F. Santos	UFSCar, Brazil
Shin'ichi Satoh	National Institute of Informatics, Japan
Stephanie Schuckers	Clarkson University, USA
Srirangaraj Setlur	University at Buffalo, SUNY, USA
Debdoot Sheet	Indian Institute of Technology, Kharagpur, India
Jun Shen	University of Wollongong, Australia
Li Shen	JD Explore Academy, China
Chen Shengyong	Zhejiang University of Technology and Tianjin University of Technology, China
Andy Song	RMIT University, Australia
Akihiro Sugimoto	National Institute of Informatics, Japan
Qianru Sun	Singapore Management University, Singapore
Arijit Sur	Indian Institute of Technology, Guwahati, India
Estefania Talavera	University of Twente, Netherlands

Wei Tang	University of Illinois at Chicago, USA
Joao M. Tavares	Universidade do Porto, Portugal
Jun Wan	NLPR, CASIA, China
Le Wang	Xi'an Jiaotong University, China
Lei Wang	Australian National University, Australia
Xiaoyang Wang	Tencent AI Lab, USA
Xinggang Wang	Huazhong University of Science and Technology, China
Xiao-Jun Wu	Jiangnan University, China
Yiding Yang	Bytedance, China
Xiwen Yao	Northwestern Polytechnical University, China
Xu-Cheng Yin	University of Science and Technology Beijing, China
Baosheng Yu	University of Sydney, Australia
Shiqi Yu	Southern University of Science and Technology, China
Xin Yuan	Westlake University, China
Yibing Zhan	JD Explore Academy, China
Jing Zhang	University of Sydney, Australia
Lefei Zhang	Wuhan University, China
Min-Ling Zhang	Southeast University, China
Wenbin Zhang	Florida International University, USA
Jiahuan Zhou	Peking University, China
Sanping Zhou	Xi'an Jiaotong University, China
Tianyi Zhou	University of Maryland, USA
Lei Zhu	Shandong Normal University, China
Pengfei Zhu	Tianjin University, China
Wangmeng Zuo	Harbin Institute of Technology, China

Reviewers (Competition Papers)

Liangcai Gao
Mingxin Huang
Lei Kang
Wenhui Liao
Yuliang Liu
Yongxin Shi
Da-Han Wang
Yang Xue
Wentao Yang
Jiaxin Zhang
Yiwu Zhong

Reviewers (Conference Papers)

Aakanksha Aakanksha
Aayush Singla
Abdul Muqeet
Abhay Yadav
Abhijeet Vijay Nandedkar
Abhimanyu Sahu
Abhinav Rajvanshi
Abhisek Ray
Abhishek Shrivastava
Abhra Chaudhuri
Aditi Roy
Adriano Simonetto
Adrien Maglo
Ahmed Abdulkadir
Ahmed Boudissa
Ahmed Hamdi
Ahmed Rida Sekkat
Ahmed Sharafeldeen
Aiman Farooq
Aishwarya Venkataramanan
Ajay Kumar
Ajay Kumar Reddy Poreddy
Ajita Rattani
Ajoy Mondal
Akbar K.
Akbar Telikani
Akshay Agarwal
Akshit Jindal
Al Zadid Sultan Bin Habib
Albert Clapés
Alceu Britto
Alejandro Peña
Alessandro Ortis
Alessia Auriemma Citarella
Alexandre Stenger
Alexandros Sopasakis
Alexia Toumpa
Ali Khan
Alik Pramanick
Alireza Alaei
Alper Yilmaz
Aman Verma
Amit Bhardwaj
Amit More
Amit Nandedkar
Amitava Chatterjee
Amos L. Abbott
Amrita Mohan
Anand Mishra
Ananda S. Chowdhury
Anastasia Zakharova
Anastasios L. Kesidis
Andras Horvath
Andre Gustavo Hochuli
André P. Kelm
Andre Wyzykowski
Andrea Bottino
Andrea Lagorio
Andrea Torsello
Andreas Fischer
Andreas K. Maier
Andreu Girbau Xalabarder
Andrew Beng Jin Teoh
Andrew Shin
Andy J. Ma
Aneesh S. Chivukula
Ángela Casado-García
Anh Quoc Nguyen
Anindya Sen
Anirban Saha
Anjali Gautam
Ankan Bhattacharyya
Ankit Jha
Anna Scius-Bertrand
Annalisa Franco
Antoine Doucet
Antonino Staiano
Antonio Fernández
Antonio Parziale
Anu Singha
Anustup Choudhury
Anwesan Pal
Anwesha Sengupta
Archisman Adhikary
Arjan Kuijper
Arnab Kumar Das

Arnav Bhavsar
Arnav Varma
Arpita Dutta
Arshad Jamal
Artur Jordao
Arunkumar Chinnaswamy
Aryan Jadon
Aryaz Baradarani
Ashima Anand
Ashis Dhara
Ashish Phophalia
Ashok K. Bhateja
Ashutosh Vaish
Ashwani Kumar
Asifuzzaman Lasker
Atefeh Khoshkhahtinat
Athira Nambiar
Attilio Fiandrotti
Avandra S. Hemachandra
Avik Hati
Avinash Sharma
B. H. Shekar
B. Uma Shankar
Bala Krishna Thunakala
Balaji Tk
Balázs Pálffy
Banafsheh Adami
Bang-Dang Pham
Baochang Zhang
Baodi Liu
Bashirul Azam Biswas
Beiduo Chen
Benedikt Kottler
Beomseok Oh
Berkay Aydin
Berlin S. Shaheema
Bertrand Kerautret
Bettina Finzel
Bhavana Singh
Bibhas C. Dhara
Bilge Gunsel
Bin Chen
Bin Li
Bin Liu
Bin Yao
Bin-Bin Jia
Binbin Yong
Bindita Chaudhuri
Bindu Madhavi Tummala
Binh M. Le
Bi-Ru Dai
Bo Huang
Bo Jiang
Bob Zhang
Bowen Liu
Bowen Zhang
Boyang Zhang
Boyu Diao
Boyun Li
Brian M. Sadler
Bruce A. Maxwell
Bryan Bo Cao
Buddhika L. Semage
Bushra Jalil
Byeong-Seok Shin
Byung-Gyu Kim
Caihua Liu
Cairong Zhao
Camille Kurtz
Carlos A. Caetano
Carlos D. Martã-Nez-Hinarejos
Ce Wang
Cevahir Cigla
Chakravarthy Bhagvati
Chandrakanth Vipparla
Changchun Zhang
Changde Du
Changkun Ye
Changxu Cheng
Chao Fan
Chao Guo
Chao Qu
Chao Wen
Chayan Halder
Che-Jui Chang
Chen Feng
Chenan Wang
Cheng Yu
Chenghao Qian
Cheng-Lin Liu

Chengxu Liu
Chenru Jiang
Chensheng Peng
Chetan Ralekar
Chih-Wei Lin
Chih-Yi Chiu
Chinmay Sahu
Chintan Patel
Chintan Shah
Chiranjoy Chattopadhyay
Chong Wang
Choudhary Shyam Prakash
Christophe Charrier
Christos Smailis
Chuanwei Zhou
Chun-Ming Tsai
Chunpeng Wang
Ciro Russo
Claudio De Stefano
Claudio F. Santos
Claudio Marrocco
Connor Levenson
Constantine Dovrolis
Constantine Kotropoulos
Dai Shi
Dakshina Ranjan Kisku
Dan Anitei
Dandan Zhu
Daniela Pamplona
Danli Wang
Danqing Huang
Daoan Zhang
Daqing Hou
David A. Clausi
David Freire Obregon
David Münch
David Pujol Perich
Davide Marelli
De Zhang
Debalina Barik
Debapriya Roy (Kundu)
Debashis Das
Debashis Das Chakladar
Debi Prosad Dogra
Debraj D. Basu
Decheng Liu
Deen Dayal Mohan
Deep A. Patel
Deepak Kumar
Dengpan Liu
Denis Coquenet
Désiré Sidibé
Devesh Walawalkar
Dewan Md. Farid
Di Ming
Di Qiu
Di Yuan
Dian Jia
Dianmo Sheng
Diego Thomas
Diganta Saha
Dimitri Bulatov
Dimpy Varshni
Dingcheng Yang
Dipanjan Das
Dipanjyoti Paul
Divya Biligere Shivanna
Divya Saxena
Divya Sharma
Dmitrii Matveichev
Dmitry Minskiy
Dmitry V. Sorokin
Dong Zhang
Donghua Wang
Donglin Zhang
Dongming Wu
Dongqiangzi Ye
Dongqing Zou
Dongrui Liu
Dongyang Zhang
Dongzhan Zhou
Douglas Rodrigues
Duarte Folgado
Duc Minh Vo
Duoxuan Pei
Durai Arun Pannir Selvam
Durga Bhavani S.
Eckart Michaelsen
Elena Goyanes
Élodie Puybareau

Emanuele Vivoli
Emna Ghorbel
Enrique Naredo
Enyu Cai
Eric Patterson
Ernest Valveny
Eva Blanco-Mallo
Eva Breznik
Evangelos Sartinas
Fabio Solari
Fabiola De Marco
Fan Wang
Fangda Li
Fangyuan Lei
Fangzhou Lin
Fangzhou Luo
Fares Bougourzi
Farman Ali
Fatiha Mokdad
Fei Shen
Fei Teng
Fei Zhu
Feiyan Hu
Felipe Gomes Oliveira
Feng Li
Fengbei Liu
Fenghua Zhu
Fillipe D. M. De Souza
Flavio Piccoli
Flavio Prieto
Florian Kleber
Francesc Serratosa
Francesco Bianconi
Francesco Castro
Francesco Ponzio
Francisco Javier Hernández López
Frédéric Rayar
Furkan Osman Kar
Fushuo Huo
Fuxiao Liu
Fu-Zhao Ou
Gabriel Turinici
Gabrielle Flood
Gajjala Viswanatha Reddy
Gaku Nakano
Galal Binamakhashen
Ganesh Krishnasamy
Gang Pan
Gangyan Zeng
Gani Rahmon
Gaurav Harit
Gennaro Vessio
Genoveffa Tortora
George Azzopardi
Gerard Ortega
Gerardo E. Altamirano-Gomez
Gernot A. Fink
Gibran Benitez-Garcia
Gil Ben-Artzi
Gilbert Lim
Giorgia Minello
Giorgio Fumera
Giovanna Castellano
Giovanni Puglisi
Giulia Orrù
Giuliana Ramella
Gökçe Uludoğan
Gopi Ramena
Gorthi Rama Krishna Sai Subrahmanyam
Gourav Datta
Gowri Srinivasa
Gozde Sahin
Gregory Randall
Guanjie Huang
Guanjun Li
Guanwen Zhang
Guanyu Xu
Guanyu Yang
Guanzhou Ke
Guhnoo Yun
Guido Borghi
Guilherme Brandão Martins
Guillaume Caron
Guillaume Tochon
Guocai Du
Guohao Li
Guoqiang Zhong
Guorong Li
Guotao Li
Gurman Gill

Haechang Lee
Haichao Zhang
Haidong Xie
Haifeng Zhao
Haimei Zhao
Hainan Cui
Haixia Wang
Haiyan Guo
Hakime Ozturk
Hamid Kazemi
Han Gao
Hang Zou
Hanjia Lyu
Hanjoo Cho
Hanqing Zhao
Hanyuan Liu
Hanzhou Wu
Hao Li
Hao Meng
Hao Sun
Hao Wang
Hao Xing
Hao Zhao
Haoan Feng
Haodi Feng
Haofeng Li
Haoji Hu
Haojie Hao
Haojun Ai
Haopeng Zhang
Haoran Li
Haoran Wang
Haorui Ji
Haoxiang Ma
Haoyu Chen
Haoyue Shi
Harald Koestler
Harbinder Singh
Harris V. Georgiou
Hasan F. Ates
Hasan S. M. Al-Khaffaf
Hatef Otroshi Shahreza
Hebeizi Li
Heng Zhang
Hengli Wang
Hengyue Liu
Hertog Nugroho
Hieyong Jeong
Himadri Mukherjee
Hoai Ngo
Hoda Mohaghegh
Hong Liu
Hong Man
Hongcheng Wang
Hongjian Zhan
Hongxi Wei
Hongyu Hu
Hoseong Kim
Hossein Ebrahimnezhad
Hossein Malekmohamadi
Hrishav Bakul Barua
Hsueh-Yi Sean Lin
Hua Wei
Huafeng Li
Huali Xu
Huaming Chen
Huan Wang
Huang Chen
Huanran Chen
Hua-Wen Chang
Huawen Liu
Huayi Zhan
Hugo Jair Escalante
Hui Chen
Hui Li
Huichen Yang
Huiqiang Jiang
Huiyuan Yang
Huizi Yu
Hung T. Nguyen
Hyeongyu Kim
Hyeonjeong Park
Hyeonjun Lee
Hymalai Bello
Hyung-Gun Chi
Hyunsoo Kim
I-Chen Lin
Ik Hyun Lee
Ilan Shimshoni
Imad Eddine Toubal

Imran Sarker
Inderjot Singh Saggu
Indrani Mukherjee
Indranil Sur
Ines Rieger
Ioannis Pierros
Irina Rabaev
Ivan V. Medri
J. Rafid Siddiqui
Jacek Komorowski
Jacopo Bonato
Jacson Rodrigues Correia-Silva
Jaekoo Lee
Jaime Cardoso
Jakob Gawlikowski
Jakub Nalepa
James L. Wayman
Jan Čech
Jangho Lee
Jani Boutellier
Javier Gurrola-Ramos
Javier Lorenzo-Navarro
Jayasree Saha
Jean Lee
Jean Paul Barddal
Jean-Bernard Hayet
Jean-Philippe G. Tarel
Jean-Yves Ramel
Jenny Benois-Pineau
Jens Bayer
Jerin Geo James
Jesús Miguel García-Gorrostieta
Jia Qu
Jiahong Chen
Jiaji Wang
Jian Hou
Jian Liang
Jian Xu
Jian Zhu
Jianfeng Lu
Jianfeng Ren
Jiangfan Liu
Jianguo Wang
Jiangyan Yi
Jiangyong Duan
Jianhua Yang
Jianhua Zhang
Jianhui Chen
Jianjia Wang
Jianli Xiao
Jianqiang Xiao
Jianwu Wang
Jianxin Zhang
Jianxiong Gao
Jianxiong Zhou
Jianyu Wang
Jianzhong Wang
Jiaru Zhang
Jiashu Liao
Jiaxin Chen
Jiaxin Lu
Jiaxing Ye
Jiaxuan Chen
Jiaxuan Li
Jiayi He
Jiayin Lin
Jie Ou
Jiehua Zhang
Jiejie Zhao
Jignesh S. Bhatt
Jin Gao
Jin Hou
Jin Hu
Jin Shang
Jing Tian
Jing Yu Chen
Jingfeng Yao
Jinglun Feng
Jingtong Yue
Jingwei Guo
Jingwen Xu
Jingyuan Xia
Jingzhe Ma
Jinhong Wang
Jinjia Wang
Jinlai Zhang
Jinlong Fan
Jinming Su
Jinrong He
Jintao Huang

Jinwoo Ahn
Jinwoo Choi
Jinyang Liu
Jinyu Tian
Jionghao Lin
Jiuding Duan
Jiwei Shen
Jiyan Pan
Jiyoun Kim
João Papa
Johan Debayle
John Atanbori
John Wilson
John Zhang
Jónathan Heras
Joohi Chauhan
Jorge Calvo-Zaragoza
Jorge Figueroa
Jorma Laaksonen
José Joaquim De Moura Ramos
Jose Vicent
Joseph Damilola Akinyemi
Josiane Zerubia
Juan Wen
Judit Szücs
Juepeng Zheng
Juha Roning
Jumana H. Alsubhi
Jun Cheng
Jun Ni
Jun Wan
Junghyun Cho
Junjie Liang
Junjie Ye
Junlin Hu
Juntong Ni
Junxin Lu
Junxuan Li
Junyaup Kim
Junyeong Kim
Jürgen Seiler
Jushang Qiu
Juyang Weng
Jyostna Devi Bodapati
Jyoti Singh Kirar
Kai Jiang
Kaiqiang Song
Kalidas Yeturu
Kalle Åström
Kamalakar Vijay Thakare
Kang Gu
Kang Ma
Kanji Tanaka
Karthik Seemakurthy
Kaushik Roy
Kavisha Jayathunge
Kazuki Uehara
Ke Shi
Keigo Kimura
Keiji Yanai
Kelton A. P. Costa
Kenneth Camilleri
Kenny Davila
Ketan Atul Bapat
Ketan Kotwal
Kevin Desai
Keyu Long
Khadiga Mohamed Ali
Khakon Das
Khan Muhammad
Kilho Son
Kim-Ngan Nguyen
Kishan Kc
Kishor P. Upla
Klaas Dijkstra
Komal Bharti
Konstantinos Triaridis
Kostas Ioannidis
Koyel Ghosh
Kripabandhu Ghosh
Krishnendu Ghosh
Kshitij S. Jadhav
Kuan Yan
Kun Ding
Kun Xia
Kun Zeng
Kunal Banerjee
Kunal Biswas
Kunchi Li
Kurban Ubul

Lahiru N. Wijayasingha
Laines Schmalwasser
Lakshman Mahto
Lala Shakti Swarup Ray
Lale Akarun
Lan Yan
Lawrence Amadi
Lee Kang Il
Lei Fan
Lei Shi
Lei Wang
Leonardo Rossi
Lequan Lin
Levente Tamas
Li Bing
Li Li
Li Ma
Li Song
Lia Morra
Liang Xie
Liang Zhao
Lianwen Jin
Libing Zeng
Lidia Sánchez-González
Lidong Zeng
Lijun Li
Likang Wang
Lili Zhao
Lin Chen
Lin Huang
Linfei Wang
Ling Lo
Lingchen Meng
Lingheng Meng
Lingxiao Li
Lingzhong Fan
Liqi Yan
Liqiang Jing
Lisa Gutzeit
Liu Ziyi
Liushuai Shi
Liviu-Daniel Stefan
Liyuan Ma
Liyun Zhu
Lizuo Jin
Longteng Guo
Lorena Álvarez Rodríguez
Lorenzo Putzu
Lu Leng
Lu Pang
Lu Wang
Luan Pham
Luc Brun
Luca Guarnera
Luca Piano
Lucas Alexandre Ramos
Lucas Goncalves
Lucas M. Gago
Luigi Celona
Luis C. S. Afonso
Luis Gerardo De La Fraga
Luis S. Luevano
Luis Teixeira
Lunke Fei
M. Hassaballah
Maddimsetti Srinivas
Mahendran N.
Mahesh Mohan M. R.
Maiko Lie
Mainak Singha
Makoto Hirose
Malay Bhattacharyya
Mamadou Dian Bah
Man Yao
Manali J. Patel
Manav Prabhakar
Manikandan V. M.
Manish Bhatt
Manjunath Shantharamu
Manuel Curado
Manuel Günther
Manuel Marques
Marc A. Kastner
Marc Chaumont
Marc Cheong
Marc Lalonde
Marco Cotogni
Marcos C. Santana
Mario Molinara
Mariofanna Milanova

Markus Bauer
Marlon Becker
Mårten Wadenbäck
Martin G. Ljungqvist
Martin Kampel
Martina Pastorino
Marwan Torki
Masashi Nishiyama
Masayuki Tanaka
Massimo O. Spata
Matteo Ferrara
Matthew D. Dawkins
Matthew Gadd
Matthew S. Watson
Maura Pintor
Max Ehrlich
Maxim Popov
Mayukh Das
Md Baharul Islam
Md Sajid
Meghna Kapoor
Meghna P. Ayyar
Mei Wang
Meiqi Wu
Melissa L. Tijink
Meng Li
Meng Liu
Meng-Luen Wu
Mengnan Liu
Mengxi China Guo
Mengya Han
Michaël Clément
Michal Kawulok
Mickael Coustaty
Miguel Domingo
Milind G. Padalkar
Ming Liu
Ming Ma
Mingchen Feng
Mingde Yao
Minghao Li
Mingjie Sun
Ming-Kuang Daniel Wu
Mingle Xu
Mingyong Li
Mingyuan Jiu
Minh P. Nguyen
Minh Q. Tran
Minheng Ni
Minsu Kim
Minyi Zhao
Mirko Paolo Barbato
Mo Zhou
Modesto Castrillón-Santana
Mohamed Amine Mezghich
Mohamed Dahmane
Mohamed Elsharkawy
Mohamed Yousuf
Mohammad Hashemi
Mohammad Khalooei
Mohammad Khateri
Mohammad Mahdi Dehshibi
Mohammad Sadil Khan
Mohammed Mahmoud
Moises Diaz
Monalisha Mahapatra
Monidipa Das
Mostafa Kamali Tabrizi
Mridul Ghosh
Mrinal Kanti Bhowmik
Muchao Ye
Mugalodi Ramesha Rakesh
Muhammad Rameez Ur Rahman
Muhammad Suhaib Kanroo
Muming Zhao
Munender Varshney
Munsif Ali
Na Lv
Nader Karimi
Nagabhushan Somraj
Nakkwan Choi
Nakul Agarwal
Nan Pu
Nan Zhou
Nancy Mehta
Nand Kumar Yadav
Nandakishor Nandakishor
Nandyala Hemachandra
Nanfeng Jiang
Narayan Hegde

Narayan Ji Mishra
Narayan Vetrekar
Narendra D. Londhe
Nathalie Girard
Nati Ofir
Naval Kishore Mehta
Nazmul Shahadat
Neeti Narayan
Neha Bhargava
Nemanja Djuric
Newlin Shebiah R.
Ngo Ba Hung
Nhat-Tan Bui
Niaz Ahmad
Nick Theisen
Nicolas Passat
Nicolas Ragot
Nicolas Sidere
Nikolaos Mitianoudis
Nikolas Ebert
Nilah Ravi Nair
Nilesh A. Ahuja
Nilkanta Sahu
Nils Murrugarra-Llerena
Nina S. T. Hirata
Ninad Aithal
Ning Xu
Ningzhi Wang
Niraj Kumar
Nirmal S. Punjabi
Nisha Varghese
Norio Tagawa
Obaidullah Md Sk
Oguzhan Ulucan
Olfa Mechi
Oliver Tüselmann
Orazio Pontorno
Oriol Ramos Terrades
Osman Akin
Ouadi Beya
Ozge Mercanoglu Sincan
Pabitra Mitra
Padmanabha Reddy Y. C. A.
Palaash Agrawal
Palaiahnakote Shivakumara
Palash Ghosal
Pallav Dutta
Paolo Rota
Paramanand Chandramouli
Paria Mehrani
Parth Agrawal
Partha Basuchowdhuri
Patrick Horain
Pavan Kumar
Pavan Kumar Anasosalu Vasu
Pedro Castro
Peipei Li
Peipei Yang
Peisong Shen
Peiyu Li
Peng Li
Pengfei He
Pengrui Quan
Pengxin Zeng
Pengyu Yan
Peter Eisert
Petra Gomez-Krämer
Pierrick Bruneau
Ping Cao
Pingping Zhang
Pintu Kumar
Pooja Kumari
Pooja Sahani
Prabhu Prasad Dev
Pradeep Kumar
Pradeep Singh
Pranjal Sahu
Prasun Roy
Prateek Keserwani
Prateek Mittal
Praveen Kumar Chandaliya
Praveen Tirupattur
Pravin Nair
Preeti Gopal
Preety Singh
Prem Shanker Yadav
Prerana Mukherjee
Prerna A. Mishra
Prianka Dey
Priyanka Mudgal

Qc Kha Ng
Qi Li
Qi Ming
Qi Wang
Qi Zuo
Qian Li
Qiang Gan
Qiang He
Qiang Wu
Qiangqiang Zhou
Qianli Zhao
Qiansen Hong
Qiao Wang
Qidong Huang
Qihua Dong
Qin Yuke
Qing Guo
Qingbei Guo
Qingchao Zhang
Qingjie Liu
Qinhong Yang
Qiushi Shi
Qixiang Chen
Quan Gan
Quanlong Guan
Rachit Chhaya
Radu Tudor Ionescu
Rafal Zdunek
Raghavendra Ramachandra
Rahimul I. Mazumdar
Rahul Kumar Ray
Rajib Dutta
Rajib Ghosh
Rakesh Kumar
Rakesh Paul
Rama Chellappa
Rami O. Skaik
Ramon Aranda
Ran Wei
Ranga Raju Vatsavai
Ranganath Krishnan
Rasha Friji
Rashmi S.
Razaib Tariq
Rémi Giraud
René Schuster
Renlong Hang
Renrong Shao
Renu Sharma
Reza Sadeghian
Richard Zanibbi
Rimon Elias
Rishabh Shukla
Rita Delussu
Riya Verma
Robert J. Ravier
Robert Sablatnig
Robin Strand
Rocco Pietrini
Rocio Diaz Martin
Rocio Gonzalez-Diaz
Rohit Venkata Sai Dulam
Romain Giot
Romi Banerjee
Ru Wang
Ruben Machucho
Ruddy Théodose
Ruggero Pintus
Rui Deng
Rui P. Paiva
Rui Zhao
Ruifan Li
Ruigang Fu
Ruikun Li
Ruirui Li
Ruixiang Jiang
Ruowei Jiang
Rushi Lan
Rustam Zhumagambetov
S. Amutha
S. Divakar Bhat
Sagar Goyal
Sahar Siddiqui
Sahbi Bahroun
Sai Karthikeya Vemuri
Saibal Dutta
Saihui Hou
Sajad Ahmad Rather
Saksham Aggarwal
Sakthi U.

Salimeh Sekeh
Samar Bouazizi
Samia Boukir
Samir F. Harb
Samit Biswas
Samrat Mukhopadhyay
Samriddha Sanyal
Sandika Biswas
Sandip Purnapatra
Sanghyun Jo
Sangwoo Cho
Sanjay Kumar
Sankaran Iyer
Sanket Biswas
Santanu Roy
Santosh D. Pandure
Santosh Ku Behera
Santosh Nanabhau Palaskar
Santosh Prakash Chouhan
Sarah S. Alotaibi
Sasanka Katreddi
Sathyanarayanan N. Aakur
Saurabh Yadav
Sayan Rakshit
Scott McCloskey
Sebastian Bunda
Sejuti Rahman
Selim Aksoy
Sen Wang
Seraj A. Mostafa
Shanmuganathan Raman
Shao-Yuan Lo
Shaoyuan Xu
Sharia Arfin Tanim
Shehreen Azad
Sheng Wan
Shengdong Zhang
Shengwei Qin
Shenyuan Gao
Sherry X. Chen
Shibaprasad Sen
Shigeaki Namiki
Shiguang Liu
Shijie Ma
Shikun Li
Shinichiro Omachi
Shirley David
Shishir Shah
Shiv Ram Dubey
Shiva Baghel
Shivanand S. Gornale
Shogo Sato
Shotaro Miwa
Shreya Ghosh
Shreya Goyal
Shuai Su
Shuai Wang
Shuai Zheng
Shuaifeng Zhi
Shuang Qiu
Shuhei Tarashima
Shujing Lyu
Shuliang Wang
Shun Zhang
Shunming Li
Shunxin Wang
Shuping Zhao
Shuquan Ye
Shuwei Huo
Shuyue Lan
Shyi-Chyi Cheng
Si Chen
Siddarth Ravichandran
Sihan Chen
Siladittya Manna
Silambarasan Elkana Ebinazer
Simon Benaïchouche
Simon S. Woo
Simone Caldarella
Simone Milani
Simone Zini
Sina Lotfian
Sitao Luan
Sivaselvan B.
Siwei Li
Siwei Wang
Siwen Luo
Siyu Chen
Sk Aziz Ali
Sk Md Obaidullah

Sneha Shukla
Snehasis Banerjee
Snehasis Mukherjee
Snigdha Sen
Sofia Casarin
Soheila Farokhi
Soma Bandyopadhyay
Son Minh Nguyen
Son Xuan Ha
Sonal Kumar
Sonam Gupta
Sonam Nahar
Song Ouyang
Sotiris Kotsiantis
Souhaila Djaffal
Soumen Biswas
Soumen Sinha
Soumitri Chattopadhyay
Souvik Sengupta
Spiros Kostopoulos
Sreeraj Ramachandran
Sreya Banerjee
Srikanta Pal
Srinivas Arukonda
Stephane A. Guinard
Su O. Ruan
Subhadip Basu
Subhajit Paul
Subhankar Ghosh
Subhankar Mishra
Subhankar Roy
Subhash Chandra Pal
Subhayu Ghosh
Sudip Das
Sudipta Banerjee
Suhas Pillai
Sujit Das
Sukalpa Chanda
Sukhendu Das
Suklav Ghosh
Suman K. Ghosh
Suman Samui
Sumit Mishra
Sungho Suh
Sunny Gupta
Suraj Kumar Pandey
Surendrabikram Thapa
Suresh Sundaram
Sushil Bhattacharjee
Susmita Ghosh
Swakkhar Shatabda
Syed Ms Islam
Syed Tousiful Haque
Taegyeong Lee
Taihui Li
Takashi Shibata
Takeshi Oishi
Talha Ahmad Siddiqui
Tanguy Gernot
Tangwen Qian
Tanima Bhowmik
Tanpia Tasnim
Tao Dai
Tao Hu
Tao Sun
Taoran Yi
Tapan Shah
Taveena Lotey
Teng Huang
Tengqi Ye
Teresa Alarcon
Tetsuji Ogawa
Thanh Phuong Nguyen
Thanh Tuan Nguyen
Thattapon Surasak
Thibault Napolãon
Thierry Bouwmans
Thinh Truong Huynh Nguyen
Thomas De Min
Thomas E. K. Zielke
Thomas Swearingen
Tianatahina Jimmy Francky Randrianasoa
Tianheng Cheng
Tianjiao He
Tianyi Wei
Tianyuan Zhang
Tianyue Zheng
Tiecheng Song
Tilottama Goswami
Tim Büchner

Tim H. Langer
Tim Raven
Tingkai Liu
Tingting Yao
Tobias Meisen
Toby P. Breckon
Tong Chen
Tonghua Su
Tran Tuan Anh
Tri-Cong Pham
Trishna Saikia
Trung Quang Truong
Tuan T. Nguyen
Tuan Vo Van
Tushar Shinde
Ujjwal Karn
Ukrit Watchareeruetai
Uma Mudenagudi
Umarani Jayaraman
V. S. Malemath
Vallidevi Krishnamurthy
Ved Prakash
Venkata Krishna Kishore Kolli
Venkata R. Vavilthota
Venkatesh Thirugnana Sambandham
Verónica Maria Vasconcelos
Véronique Ve Eglin
Víctor E. Alonso-Pérez
Vinay Palakkode
Vinayak S. Nageli
Vincent J. Whannou De Dravo
Vincenzo Conti
Vincenzo Gattulli
Vineet Padmanabhan
Vishakha Pareek
Viswanath Gopalakrishnan
Vivek Singh Baghel
Vivekraj K.
Vladimir V. Arlazarov
Vu-Hoang Tran
W. Sylvia Lilly Jebarani
Wachirawit Ponghiran
Wafa Khlif
Wang An-Zhi
Wanli Xue
Wataru Ohyama
Wee Kheng Leow
Wei Chen
Wei Cheng
Wei Hua
Wei Lu
Wei Pan
Wei Tian
Wei Wang
Wei Wei
Wei Zhou
Weidi Liu
Weidong Yang
Weijun Tan
Weimin Lyu
Weinan Guan
Weining Wang
Weiqiang Wang
Weiwei Guo
Weixia Zhang
Wei-Xuan Bao
Weizhong Jiang
Wen Xie
Wenbin Qian
Wenbin Tian
Wenbin Wang
Wenbo Zheng
Wenhan Luo
Wenhao Wang
Wen-Hung Liao
Wenjie Li
Wenkui Yang
Wenwen Si
Wenwen Yu
Wenwen Zhang
Wenwu Yang
Wenxi Li
Wenxi Yue
Wenxue Cui
Wenzhuo Liu
Widhiyo Sudiyono
Willem Dijkstra
Wolfgang Fuhl
Xi Zhang
Xia Yuan

Xianda Zhang
Xiang Zhang
Xiangdong Su
Xiang-Ru Yu
Xiangtai Li
Xiangyu Xu
Xiao Guo
Xiao Hu
Xiao Wu
Xiao Yang
Xiaofeng Zhang
Xiaogang Du
Xiaoguang Zhao
Xiaoheng Jiang
Xiaohong Zhang
Xiaohua Huang
Xiaohua Li
Xiao-Hui Li
Xiaolong Sun
Xiaosong Li
Xiaotian Li
Xiaoting Wu
Xiaotong Luo
Xiaoyan Li
Xiaoyang Kang
Xiaoyi Dong
Xin Guo
Xin Lin
Xin Ma
Xinchi Zhou
Xingguang Zhang
Xingjian Leng
Xingpeng Zhang
Xingzheng Lyu
Xinjian Huang
Xinqi Fan
Xinqi Liu
Xinqiao Zhang
Xinrui Cui
Xizhan Gao
Xu Cao
Xu Ouyang
Xu Zhao
Xuan Shen
Xuan Zhou
Xuchen Li
Xuejing Lei
Xuelu Feng
Xueting Liu
Xuewei Li
Xueyi X. Wang
Xugong Qin
Xu-Qian Fan
Xuxu Liu
Xu-Yao Zhang
Yan Huang
Yan Li
Yan Wang
Yan Xia
Yan Zhuang
Yanan Li
Yanan Zhang
Yang Hou
Yang Jiao
Yang Liping
Yang Liu
Yang Qian
Yang Yang
Yang Zhao
Yangbin Chen
Yangfan Zhou
Yanhui Guo
Yanjia Huang
Yanjun Zhu
Yanming Zhang
Yanqing Shen
Yaoming Cai
Yaoxin Zhuo
Yaoyan Zheng
Yaping Zhang
Yaqian Liang
Yarong Feng
Yasmina Benmabrouk
Yasufumi Sakai
Yasutomo Kawanishi
Yazeed Alzahrani
Ye Du
Ye Duan
Yechao Zhang
Yeong-Jun Cho

Yi Huo
Yi Shi
Yi Yu
Yi Zhang
Yibo Liu
Yibo Wang
Yi-Chieh Wu
Yifan Chen
Yifei Huang
Yihao Ding
Yijie Tang
Yikun Bai
Yimin Wen
Yinan Yang
Yin-Dong Zheng
Yinfeng Yu
Ying Dai
Yingbo Li
Yiqiao Li
Yiqing Huang
Yisheng Lv
Yisong Xiao
Yite Wang
Yizhe Li
Yong Wang
Yonghao Dong
Yong-Hyuk Moon
Yongjie Li
Yongqian Li
Yongqiang Mao
Yongxu Liu
Yongyu Wang
Yongzhi Li
Youngha Hwang
Yousri Kessentini
Yu Wang
Yu Zhou
Yuan Tian
Yuan Zhang
Yuanbo Wen
Yuanxin Wang
Yubin Hu
Yubo Huang
Yuchen Ren
Yucheng Xing
Yuchong Yao
Yuecong Min
Yuewei Yang
Yufei Zhang
Yufeng Yin
Yugen Yi
Yuhang Ming
Yujia Zhang
Yujun Ma
Yukiko Kenmochi
Yun Hoyeoung
Yun Liu
Yunhe Feng
Yunxiao Shi
Yuru Wang
Yushun Tang
Yusuf Osmanlioglu
Yusuke Fujita
Yuta Nakashima
Yuwei Yang
Yuwu Lu
Yuxi Liu
Yuya Obinata
Yuyao Yan
Yuzhi Guo
Zaipeng Xie
Zander W. Blasingame
Zedong Wang
Zeliang Zhang
Zexin Ji
Zhanxiang Feng
Zhaofei Yu
Zhe Chen
Zhe Cui
Zhe Liu
Zhe Wang
Zhekun Luo
Zhen Yang
Zhenbo Li
Zhenchun Lei
Zhenfei Zhang
Zheng Liu
Zheng Wang
Zhengming Yu
Zhengyin Du

Zhengyun Cheng
Zhenshen Qu
Zhenwei Shi
Zhenzhong Kuang
Zhi Cai
Zhi Chen
Zhibo Chu
Zhicun Yin
Zhida Huang
Zhida Zhang
Zhifan Gao
Zhihang Ren
Zhihang Yuan
Zhihao Wang
Zhihua Xie
Zhihui Wang
Zhikang Zhang
Zhiming Zou
Zhiqi Shao
Zhiwei Dong
Zhiwei Qi
Zhixiang Wang
Zhixuan Li
Zhiyu Jiang
Zhiyuan Yan
Zhiyuan Yu
Zhiyuan Zhang
Zhong Chen
Zhongwei Teng
Zhongzhan Huang
Zhongzhi Yu
Zhuan Han
Zhuangzhuang Chen
Zhuo Liu
Zhuo Su
Zhuojun Zou
Zhuoyue Wang
Ziang Song
Zicheng Zhang
Zied Mnasri
Zifan Chen
Žiga Babnik
Zijing Chen
Zikai Zhang
Ziling Huang
Zilong Du
Ziqi Cai
Ziqi Zhou
Zi-Rui Wang
Zirui Zhou
Ziwen He
Ziyao Zeng
Ziyi Zhang
Ziyue Xiang
Zonglei Jing
Zongyi Xu

Contents – Part XI

SARCOVID: A Framework for Sarcasm Detection in Tweets Using Hybrid Transfer Learning Techniques

T. K. Balaji[1], Annushree Bablani[1(✉)], S. R. Sreeja[1], and Hemant Misra[2]

[1] Indian Institute of Information Technology Sri City, Sri City 517646, Andhra Pradesh, India
{balaji.tk,annushree.bablani,sreeja.sr}@iiits.in
[2] Simpl, Bengaluru, India

Abstract. The COVID-19 pandemic sparked a surge in online discussions, making sentiment analysis challenging due to the prevalence of sarcasm on social media. Identifying sarcastic expressions within the context of COVID-19 conversations poses a unique linguistic hurdle. To tackle this challenge, a novel framework called SARCOVID is proposed that leverages hierarchical transfer learning and ensemble techniques to detect sarcasm in the field. Through rigorous evaluation on a collected COVID-19 dataset, SARCOVID demonstrates superior performance in identifying sarcastic content with reduced bias compared to traditional methods. The findings reveal a significant presence of sarcasm in online COVID-19 discussions, underscoring the importance of robust sarcasm detection techniques. In a test, the framework outperforms other models with 0.61 accuracy on Sarcasm corpus V2. This approach not only advances sentiment analysis capabilities for evolving online conversations but also provides deeper insights into the nuanced expressions of sentiment on social media.

Keywords: Sentiment analysis · Transfer learning · Deep learning · Opinion mining · Neural networks

1 Introduction

Sentiment analysis is a crucial aspect of understanding public opinion and emotions, particularly during times of crisis such as the COVID-19 pandemic. The pandemic has led to a surge in discussions and opinions on various social media platforms, where people share their thoughts and feelings about the situation. As a result, sentiment analysis has become increasingly important for governments, businesses, and researchers to understand the public's concerns, reactions, and attitudes towards the pandemic.

However, sarcasm poses a significant challenge to sentiment analysis. Sarcasm, a form of irony, can completely change the meaning of a sentence, making it difficult for sentiment analysis models to classify the sentiment accurately. For

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15311, pp. 1–12, 2025.
https://doi.org/10.1007/978-3-031-78195-7_1

example, a sentence like "I'm so glad, I am not vaccinated" may seem positive at first glance, but it is actually sarcastic and expresses the opposite sentiment. Therefore, accurately detecting sarcasm is essential for improving the accuracy of sentiment analysis, especially in the context of COVID-19.

It is crucial to address sarcasm in sentiment analysis during the COVID-19 pandemic due to the potential impact on the accuracy of sentiment classification. Sarcasm, an irony often used in online communication, can alter the sentiment expressed in text, leading to misinterpretation by sentiment analysis models. In the context of the pandemic, where emotions and opinions are heightened, accurately capturing sentiments is essential for understanding public reactions, concerns, and attitudes towards COVID-19. Failure to detect sarcasm can result in misleading analyses and misrepresenting public sentiment, highlighting the significance of effectively dealing with sarcasm to ensure the reliability and precision of sentiment analysis during this critical period.

Numerous approaches are available for detecting sarcasm, transfer learning is one such effective method among them when a domain-specific dataset is limited. Transfer learning is a machine learning technique that allows models to learn from pre-trained models and adapt them to new tasks [1]. In the case of sarcasm detection, transfer learning can be used to train models on large datasets of sarcastic and non-sarcastic text, allowing the models to learn the patterns and nuances of sarcasm. This can lead to more accurate sarcasm detection, which in turn can improve the overall performance of sentiment analysis models. Addressing this challenge of detecting sarcasm in COVID-19 tweet sentiment analysis, a novel framework called SARCOVID is introduced. This framework leverages hierarchical transfer learning and ensemble methods to enhance the accuracy of sarcasm detection. The effectiveness of SARCOVID is evaluated using the SENSECOR dataset, which revealed the prevalence of sarcasm in COVID-19 discussions.

The subsequent sections unfold as Section 2, Section 3 explains the methodology of this study, section 4 discusses the results, and section 5 concludes the study.

2 Literature work

Several studies have explored sentiment analysis of COVID-19, revealing valuable insights into public opinion. Twitter data has become a popular resource for sentiment analysis [2]. In [3], authors explored public sentiment on Twitter in India during the early stages of the COVID-19 pandemic (December 2019 to May 2020). They used TextBlob to analyze the emotional tone (polarity) of tweets and NLTK to identify frequently used words. Their analysis, visualized by state and month, revealed some surprising findings. Despite the pandemic's challenges, the dominant sentiment among Indian Twitter users was positive. This positivity coincided with announcements of lockdowns, with higher tweet volumes coming from states hit hard by COVID-19. While there were negative tweets, the positive sentiment suggests a general trust in the government's response.

The study [4] evaluated various machine learning classifiers across different datasets for sentiment analysis of COVID-19-related Twitter data. Traditional methods like TF-IDF with SVM showed strong performance, with accuracy scores ranging from 0.829 to 0.845. Embedding-based models, particularly fastText, outperformed others due to their effective handling of out-of-vocabulary words. Deep learning approaches, such as using GloVe embeddings with deep convolutional neural networks (DCNN), demonstrated superiority over bidirectional long short-term memory (BiLSTM). Hybrid models like hybrid ranking outperformed IWV, emphasizing the importance of incorporating sentiment and context information. BERT stood out among transformer-based language models, surpassing all others with performance scores exceeding 0.85 across all datasets.

Authors in [5] investigated sentiment trends across eleven heavily affected countries during the pandemic. Analyzing over 50,000 tweets, the study revealed nuanced emotional responses, with some nations displaying predominantly positive sentiments while others showed a balance between positive and negative expressions. Emotion analysis highlighted shifts over time, from initial fear to growing trust as recovery rates improved. Utilizing the Syuzhet package in R, the research compared sentiment analysis algorithms to unveil the complex emotional dynamics amidst the global crisis.

Traditional sentiment analysis approaches primarily focus on surface-level sentiment in tweets, neglecting the crucial layer of sarcasm. This research introduces SARCOVID, a novel framework that handles sarcasm in COVID-19 tweets. SARCOVID leverages a hybrid approach, combining hierarchical transfer learning for improved knowledge transfer and an ensemble majority voting [6] technique to achieve more accurate sarcasm detection. By adjusting sentiment analysis based on identified sarcasm, SARCOVID aims to provide a more nuanced understanding of public opinion within COVID-19 discussions on social media.

3 Methodology

The SARCOVID framework development comprises the fusion of two methodologies. Initially, a hierarchical evaluation of tweets is conducted for sarcasm detection, employing transfer learning techniques in the first phase. Subsequently, in the second phase, the tweets undergo classification utilizing an ensemble majority voting technique. This dual approach ensures comprehensive analysis and robust sarcasm detection within COVID-19-related discourse on social media platforms.

3.1 Dataset

The datasets used for this study to train, evaluate and test the models are:

1. **News Headlines (26,709 samples)** [7]: This dataset is dedicated to identifying sarcasm in concise texts such as headlines, ensuring an equal distribution of sarcastic and non-sarcastic examples, denoted by 1 and 0 respectively.

2. **Reddit Sarcastic (1 million reviews)** [8]: With a vast collection of reviews, this dataset serves to analyze sarcasm within online discussions, with clearly labeled instances of sarcasm (1) and non-sarcasm (0).
3. **SemEval (5,735 samples)** [9]: Sourced from the iSarcasmEval GitHub repository, this dataset provides additional data with definitive yes/no labels for sarcasm.
4. **Twitter**[1] (2,000 samples): This dataset is instrumental in uncovering sarcasm within tweets, offering a balanced mix of both sarcastic and non-sarcastic instances.
5. **Sarcasm Corpus V2(9116 samples)** [10]:The Sarcasm Corpus comprises three balanced types of samples–Generic, Rhetorical Questions, and Hyperbole–each containing an equal number of sarcastic and non-sarcastic samples.

This study evaluates the proposed SARCOVID framework on the SENSECOR [11] dataset, which we previously collected for research on the COVID-19 Omicron variant [11]. The SENSECOR dataset comprises 160,000 tweets related to this specific variant.

3.2 Preprocessing

In the text preprocessing phase, a series of steps are implemented to enhance the data quality before the text given to the model to process. It involves

- Lowercasing all text
- Eliminating punctuation
- Tokenizing
- Removal of stopwords
- Stemming and lemmatization
- Removing special characters
- Handling Emoji and acronyms (if available)
- Removal of URLs, mentions and hashtags (if available)

3.3 Methodology implementation

The process involves four deep learning models (M1, M2, M3, and M4), each trained on individual sarcasm datasets (Reddit, News Headlines, SemEval, and Twitter). Each model is tested on the SENSECOR dataset for sarcasm identification to make the sentiment analysis free from biased opinions.

[1] https://github.com/surajr/SarcasmDetection/tree/master/Data

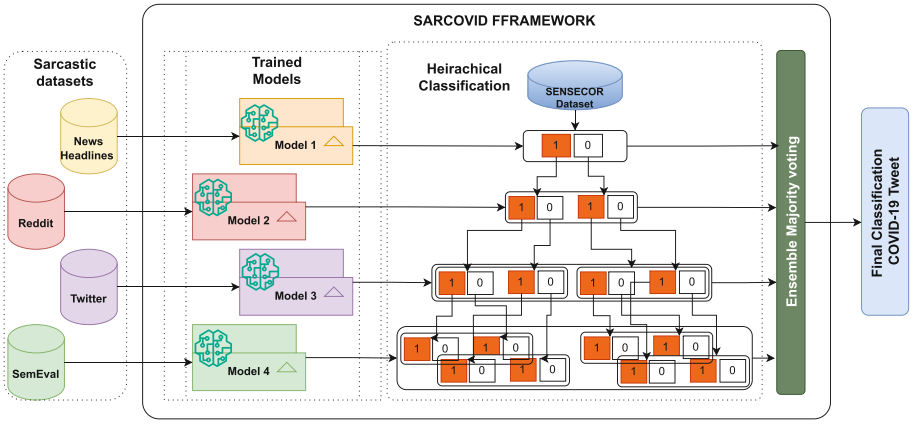


Fig. 1. Detailed architectural representation of SARCOVID framework.

hierarchical evaluation The hierarchical evaluation process outlined involves a multi-stage approach to sarcasm detection in COVID-19-related tweets using transfer learning techniques. The architectural representation of the proposed SARCOVID framework is presented in Fig. 1.

In the initial classification, model M1 is used to classify the tweets in the SENSECOR dataset, where each tweet is classified into two categories: those identified as sarcastic (Class 1) and non-sarcastic (Class 0).

In the subsequent evaluation phase, M2 conducts further analysis of the tweets based on M1's classifications. For tweets initially identified as sarcastic by M1, M2 evaluates them to either confirm or refute their sarcastic nature, resulting in two new categories: tweets confirmed as sarcastic and tweets reclassified as non-sarcastic. Similarly, for tweets initially labeled as non-sarcastic by M1, M2 evaluates them to detect any sarcastic content, leading to the creation of two additional categories: tweets flagged as sarcastic and tweets confirmed as non-sarcastic. Thus, the initial two results generated by M1 are now categorized into four.

This process continues with M3 and M4 analyzing the outputs from previous models, generating outputs at different levels for classifications based on M1, M2, M3, and M4. Consequently, M1 generates two outputs, M2 generates four outputs, and finally, M4 generates 16 outputs, totalling 30 outputs in a hierarchical fashion, comprising 15 for sarcastic and 15 for non-sarcastic tweets.

Ensemble majority Voting:

– In the final evaluation phase, all model outputs are considered to determine the final classifications for each tweet. An ensemble majority voting approach is then applied to assess the final classifications. Tweets that receive at least three votes(three model's votes) as sarcastic are categorized as sarcastic, while those receiving at least three votes(three model's votes) votes as

non-sarcastic are categorized as non-sarcastic. Overall, this multi-stage hierarchical evaluation process leverages transfer learning techniques to effectively identify sarcasm in COVID-19-related tweets, allowing for a comprehensive analysis of sarcasm prevalence within the SENSECOR dataset.

Algorithm 1. Hierarchical Sarcasm Detection

Input: SENSECOR dataset with n tweets T_i where $i > 0$
Output: Sarcastic and non-sarcastic tweet classifications
Run M_1 on all n tweets ▷ Classification using M1
$C_1 \leftarrow$ Tweets classified as sarcastic by M_2
$C_2 \leftarrow$ Tweets classified as non-sarcastic by M_2

Run M_2 on all C_1 tweets ▷ Classification using M2
$C_3 \leftarrow$ Tweets classified as sarcastic by M_2
$C_4 \leftarrow$ Tweets classified as non-sarcastic by M_2
Run M_2 on all C_2 tweets
$C_5 \leftarrow$ Tweets classified as sarcastic by M_2
$C_6 \leftarrow$ Tweets classified as non-sarcastic by M_2
Continue the classification using M3, and M4 models... ▷ assign tweets to c7,c8,...c30.
sarcastic_votes $\leftarrow \{0\}^n$
non_sarcastic_votes $\leftarrow \{0\}^n$
sarcastic_tweets $\leftarrow \{\}$
non_sarcastic_tweets $\leftarrow \{\}$
for $i = 1$ to n **do**
    for $j = 1$ to 30 **do**
        if T_i in C_j **then** ▷ T_i is tweet in a dataset
            if $j\%2 = 1$ **then**
                sarcastic_votes $\leftarrow$ sarcastic_votes $+ 1$
            else
                non_sarcastic_votes $\leftarrow$ non_sarcastic_votes $+ 1$
            end if
        end if
    end for
    if sarcastic_votes ≥ 3 **then** ▷ Threshold for classifying as sarcastic
        sarcastic_tweets.append(T_i)
    else if non_sarcastic_votes ≥ 3 **then** Threshold for classifying as non-sarcastic
        non_sarcastic_tweets.append(T_i)
    end if
end for
return sarcastic_tweets, non_sarcastic_tweets

The SARCOVID framework methodology is presented in algorithm 1. The algorithm takes the SENSECOR dataset with n tweets as input and initializes lists to store the final sarcastic and non-sarcastic tweet classifications. Model M1 is run on all tweets, and the initial sarcastic (C1) and non-sarcastic (C2)

classifications are obtained. Then, M2 runs on (C1) and classifies sarcastic tweets as (C3) and non-sarcastic tweets as (c4). Later, M2 runs on (C2) and classifies sarcastic tweets as (C5) and non-sarcastic tweets as (c6). The same procedure follows for M3 and M4 models, which classify tweets and assign them to C7, C8, ..., and C30. Here, Odd classes such as C1, C3, and C5... are holding Sarcastic tweets and even classes such as C2.C4, C6,... hold non-sarcastic tweets. After the hierarchical classification, the algorithm counts the number of sarcastic and non-sarcastic votes for each tweet based on the model outputs by iterating over the models and incrementing the respective vote counts for each tweet based on its classification. Finally, ensemble voting is applied: tweets with at least three sarcastic votes are added to the sarcastic_tweets list, and tweets with at least three non-sarcastic votes are added to the non_sarcastic_tweets list. The algorithm returns these two lists as the final sarcastic and non-sarcastic tweet classifications. The sample tweet classification of the SARCOVID framework is presented in Table 1.

Table 1. Sample COVID-19 Tweets detected in SENSECOR dataset

2 jabs taken, but tested positive for COVID-19. Meanwhile, my granny, who hasn't received the vaccine, tested negative. Thank you #FireFauci #COVIDsucks
I love lockdowns - no food, no job, no nothing! #pandemic
Can anyone play a song on COVID-19? My neighbours are disturbing me with their noise. #COVID19 #lockdown

4 Results and discussion

The performance of four distinct models BiLSTM [12], BERT [13], RoBERTa [14] and, DistilBERT [15] is evaluated across four diverse datasets: News Headlines, IsarcasmEval, Twitter, and Reddit. The performance evaluation of models is presented in table 2. On the News Headlines dataset, BiLSTM achieved a moderate accuracy of 0.81, while DistilBERT and BERT surpassed it with accuracies of 0.83 and 0.91, respectively. RoBERTa exhibited the highest accuracy of 0.93, indicating its superior performance in capturing nuanced linguistic cues associated with sarcasm in news articles. Transitioning to the IsarcasmEval dataset, BiLSTM performed moderately with an accuracy of 0.74, whereas DistilBERT, BERT, and RoBERTa showcased improved performance with accuracies of 0.758, 0.774 and 0.805, respectively, with RoBERTa achieving the highest accuracy. Moving to the Twitter dataset, BiLSTM achieved a reasonable accuracy of 0.87, while DistilBERT, BERT, and RoBERTa further improved upon this with accuracies of 0.884, 0.902, and 0.91, respectively. Lastly, on the Reddit dataset, BiLSTM demonstrated moderate performance with an accuracy of 0.7, whereas

DistilBERT, BERT, and RoBERTa exhibited superior performance with accuracies of 0.767, 0.782, and 0.79, respectively. These results underscored the varying capabilities of each model across different datasets. They highlighted the effectiveness of advanced transformer-based models like RoBERTa, BERT, and DistilBERT in sarcasm detection across diverse linguistic contexts.

Table 2. Models performances on different datasets

Model	Dataset			
	News Headlines	IsarcasmEval	Twitter	Reddit
BiLSTM	0.81	0.74	0.87	0.7
DistilBERT	0.83	0.758	0.884	0.767
BERT	0.91	0.774	0.902	0.782
RoBERTa	0.93	0.805	0.91	0.79

The SARCOVID framework incorporates the RoBERTa model as a key component in its construction. The RoBERTa model is selected for its superior performance compared to all other baseline methods examined in this study. Hence, Model M1 is RoBERTa trained on the News Headlines dataset, while Model M2 is RoBERTa trained on the Reddit Sarcastic dataset. Models M3 and M4 are RoBERTa trained on the SemEval and Twitter datasets, respectively. The proposed approach leverages transfer learning due to its adaptability to different domains. Therefore, the SARCOVID framework must be evaluated on a new labelled sarcastic benchmark dataset, which is not used for training these models. For this purpose, the evaluation of the proposed SARCOVID framework is conducted on the Sarcasm Corpus V2, utilizing 2000 randomly selected samples to assess the effectiveness of the transfer learning technique. The primary objective of the study is to detect sarcasm in COVID-19 tweets, which is accomplished using the SENSECOR dataset, a large unlabeled corpus. The analysis and findings are presented in the results section.

The performance evaluation of various models on the Sarcasm Corpus V2 dataset, as depicted in table 3, quantifies their effectiveness in sarcasm detection. BiLSTM, a traditional recurrent neural network model, achieved the lowest performance with a precision, recall, F1 score, and accuracy all at 0.50. DistilBERT exhibited a slight improvement, with a precision and recall of 0.54, an F1 score of 0.51, and an accuracy of 0.55. BERT further improved with a precision of 0.55, recall of 0.56, F1 score of 0.56, and accuracy of 0.561. RoBERTa outperformed the previous models with a precision of 0.57, recall of 0.55, F1 score of 0.59, and accuracy of 0.58. The SARCOVID model achieved the highest performance metrics, with a precision of 0.6, recall of 0.62, F1 score of 0.61, and accuracy of 0.61, highlighting its effectiveness in detecting sarcasm. Notably, the proposed SARCOVID framework outperformed all other models, boasting the highest accuracy of 0.61. This superior performance of SARCOVID underscores

its efficacy in sarcasm detection, indicating its potential to advance the field of natural language understanding and sentiment analysis. The ROC curve of the RoBERTa model using various datasets is presented in Fig. 2.

Table 3. Test analysis of models performed on Sarcasm Corpus V2 dataset.

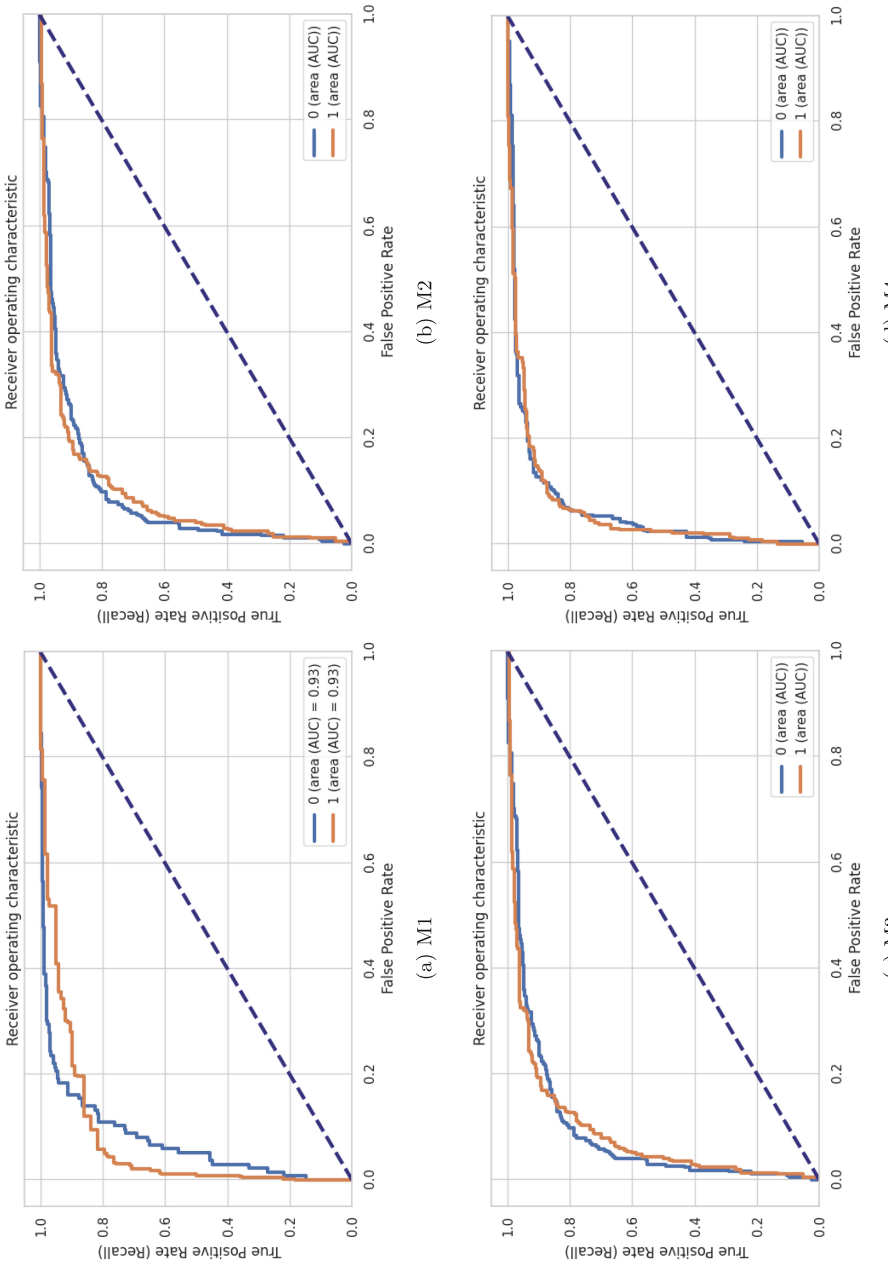
Model	Precision	Recall	F1	Accuracy
BiLSTM	0.49	0.50	0.50	0.5
DistilBERT	0.54	0.54	0.51	0.55
BERT	0.55	0.56	0.56	0.561
RoBERTa	0.57	0.55	0.59	0.58
SARCOVID	**0.6**	**0.62**	**0.61**	**0.61**

The performance of sarcasm detection models M1 to M4 on the SENSECOR dataset of 160,000 tweets is presented in Table 4. The effectiveness of each model is expressed as a percentage, indicating the proportion of tweets it classified as containing sarcasm. A lower percentage implies a more conservative approach, where only tweets exhibiting clear sarcastic cues specific to the SENSECOR dataset are identified as sarcastic. This selective identification helps reduce false positives, thereby improving the precision of sarcasm detection. Conversely, a higher percentage suggests a more liberal approach, increasing the risk of misclassifying non-sarcastic tweets as sarcastic, which could lead to lower accuracy in identifying truly sarcastic content.

Table 4. Sarcastic tweets detected in the SENSECOR dataset by models used in this study.

Models	**sarcasm detected**(in %)
M1	47.2
M2	46
M3	39
M4	48
SARCOVID	**24**

Among the evaluated models, Model M1 detected sarcasm in 47.2% of tweets, closely followed by Model M4, which identified sarcasm in 48% of tweets. Meanwhile, Models M2 and M3 exhibited lower detection rates, flagging sarcasm in 46% and 39% of tweets, respectively. Interestingly, the SARCOVID framework showcased the most discerning performance, with a detection rate of only 24%. This suggests that SARCOVID employs a rigorous approach to sarcasm detection, prioritizing accuracy by minimizing false positives and ensuring precise identification of sarcastic content within the SENSECOR dataset.



(a) M1 (b) M2 (c) M3 (d) M4

Fig. 2. ROC curves for RoBERTa model on various datasets.

The SARCOVID framework offers several advantages in its methodology for sarcasm detection in COVID-19 tweets. One notable advantage is its hierarchical evaluation approach, which allows for a comprehensive analysis by leveraging multiple deep-learning models trained on diverse sarcasm datasets. Most of the existing sarcasm detection techniques work on domain-specific labelled datasets to overcome false positive classification. This technique has a better approach to reducing the false positives without having a specific dataset. However, this methodology may present challenges regarding computational resources required for training and evaluating multiple models iteratively, as well as potential complexities in interpreting conflicting model outputs. Additionally, the reliance on pre-existing sarcasm datasets for transfer learning may introduce biases or limitations in detecting sarcasm in COVID-19 tweets.

5 Conclusion

Detecting sarcasm in COVID-19 tweets presents a significant challenge due to the scarcity of domain-specific datasets for model training. To address this issue, a novel framework called SARCOVID is proposed. SARCOVID employs a hybrid approach, combining hierarchical transfer learning and ensemble majority voting. The SARCOVID achieves high confidence in identifying sarcasm within the SENSECOR dataset, a collection of COVID-19 tweets. This framework exhibits a lower tendency for false positives in sarcasm detection, making it a robust solution for overcoming limitations in dataset availability for COVID-19 tweets. The model can be adapted to any application with limited dataset availability. A potential future direction could involve exploring the integration of additional modalities like visual data and employing continual learning techniques that could further enhance SARCOVID's ability to adapt and evolve with emerging linguistic patterns in this ever-changing domain.

References

1. Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., Liu, C.: A Survey on Deep Transfer Learning. In: Kůrková, V., Manolopoulos, Y., Hammer, B., Iliadis, L., Maglogiannis, I. (eds.) ICANN 2018. LNCS, vol. 11141, pp. 270–279. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01424-7_27
2. Ahmad, T., Ramsay, A., Ahmed, H.: Detecting emotions in english and arabic tweets. Information **10**(3), 98 (2019)
3. T. Vijay, A. Chawla, B. Dhanka, and P. Karmakar, "Sentiment analysis on covid-19 twitter data," in *2020 5th IEEE international conference on recent advances and innovations in engineering (ICRAIE)*, pp. 1–7, IEEE, 2020
4. Naseem, U., Razzak, I., Khushi, M., Eklund, P.W., Kim, J.: Covidsenti: A large-scale benchmark twitter data set for covid-19 sentiment analysis. IEEE transactions on computational social systems **8**(4), 1003–1015 (2021)
5. M. A. Kausar, A. Soosaimanickam, and M. Nasar, "Public sentiment analysis on twitter data during covid-19 outbreak," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 2, 2021

6. Kuncheva, L.I., Rodríguez, J.J.: A weighted voting framework for classifiers ensembles. Knowl. Inf. Syst. **38**, 259–275 (2014)
7. Misra, R., Arora, P.: Sarcasm detection using news headlines dataset. AI Open **4**, 13–18 (2023)
8. "A large self-annotated corpus for sarcasm." 2017
9. I. Abu Farha, S. V. Oprea, S. Wilson, and W. Magdy, "SemEval-2022 task 6: iSarcasmEval, intended sarcasm detection in English and Arabic," in *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, (Seattle, United States), pp. 802–814, Association for Computational Linguistics, July 2022
10. S. Oraby, V. Harrison, L. Reed, E. Hernandez, E. Riloff, and M. Walker, "Creating and characterizing a diverse corpus of sarcasm in dialogue,"arXiv preprint arXiv:1709.05404, 2017
11. Balaji, T., Bablani, A., Sreeja, S., Misra, H.: Sensecor: A framework for covid-19 variants severity classification and symptoms detection. Evol. Syst. **15**(1), 65–82 (2024)
12. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional lstm and other neural network architectures. Neural Netw. **18**(5–6), 602–610 (2005)
13. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding,"arXiv preprint arXiv:1810.04805, 2018
14. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," arXiv preprint arXiv:1907.11692, 2019
15. V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," arXiv preprint arXiv:1910.01108, 2019

Contrastive Conditional Alignment Based on Label Shift Calibration for Imbalanced Domain Adaptation

Xiaona Sun, Zhenyu Wu(✉), Zhiqiang Zhan, and Yang Ji

School of Communication and Information Engineering, Beijing University of Posts and Telecommunications, Beijing, China
{subxiaona,shower0512,zqzhan,jiyang}@bupt.edu.cn

Abstract. Many existing unsupervised domain adaptation (UDA) methods primarily focus on covariate shift, limiting their effectiveness in imbalanced domain adaptation (IDA) where both covariate shift and label shift coexist. Recent IDA methods have achieved promising results based on self-training using target pseudo labels. However, under the IDA scenarios, the classifier learned in the source domain will exhibit different decision bias from the target domain. It will potentially make target pseudo labels unreliable, and will further lead to error accumulation with incorrect class alignment.Thus, we propose contrastive conditional alignment based on label shift calibration (CCA-LSC) for IDA, to address both covariate shift and label shift. Initially, our contrastive conditional alignment resolve covariate shift to learn representations with domain invariance and class discriminability, which include domain adversarial learning, sample-weighted moving average centroid alignment and discriminative feature alignment. Subsequently, we estimate the probability distribution of the target domain, and calibrate target sample classification predictions based on label shift metrics to encourage labeling pseudo-labels more consistently with the distribution of real target data. Extensive experiments are conducted and demonstrate that our method outperforms existing UDA and IDA methods on benchmarks with both label shift and covariate shift. Our code is available at https://github.com/ysxcj-hub/CCA-LSC.

Keywords: Unsupervised domain adaptation · Label shift · Covariate shift · Long-tailed distribution

1 Introduction

Unsupervised Domain Adaptation (UDA) [1–4] aims to transfer knowledge from labeled source domain to unlabeled target domain. A common scenario in UDA is covariate shift, where the conditional distributions of the labels given the features are the same across domains, i.e., $P_S(y|x) = P_T(y|x)$, but the marginal distributions of the features are different, i.e., $P_S(x) \neq P_T(x)$. Many UDA

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15311, pp. 13–28, 2025.
https://doi.org/10.1007/978-3-031-78195-7_2

methods have been proposed to deal with covariate shift, such as distribution matching-based methods [3–5], which aim to align the feature distributions of the source and target domains by minimizing some distance measure. However, when there exists label distribution shift, i.e. $P_S(y|x) \neq P_T(y|x)$, distribution matching-based methods may suffer from negative transfer. In real-world scenarios, domain adaptation often faces the challenge of both data distribution shift (covariate shift) and label distribution shift (label shift). Moreover, real-world data is usually imbalanced, where some classes are more frequent than others. For example, in the domainnet [6] dataset, the head classes that are abundant in the source domain may be scarce in the target domain. This scenario is referred to as imbalanced domain adaptation (IDA). To enable domain adaptation to cope with such realistic situations, effective IDA algorithms are essential.

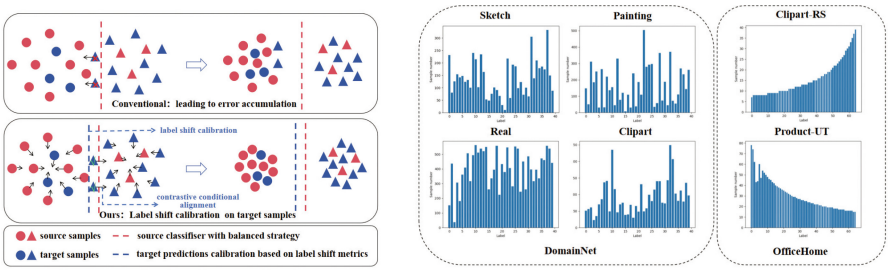


Fig. 1. (a)Top: In cases of substantial label shift, classifier learned from source domain may mislabel target samples due to the unknown target label distribution. This can result in error accumulation and misalignment in IDA methods that use self-training with pseudo-labels. **(a)Bottom:** Our approach rectifies the classification boundary to predict target samples based on the label shift metric M_{ls}, effectively reducing the error rates in estimating target pseudo-labels. We employ calibrated pseudo-labels in CCA to learn feature representations that are both domain-invariant and class-discriminative. **(b):** Label distributions on DomainNet and OfficeHome

Recent studies attempt to address the IDA problem through self-training with target pseudo-labels. However, these methods prove unstable as the classifier's output tends to align more closely with the source than the target label distribution under label shift. This discrepancy results in noisier pseudo-labels for target samples. The issue is particularly pronounced for classes with a large label shift, leading to error accumulation, as depicted in Figure 1(a), top.

To tackle this issue, we introduce a novel method termed contrastive conditional alignment based on label shift calibration (CCA-LSC). This method adjusts the classification of target samples in accordance with the degree of label shift. First, we propose to align the conditional distributions of two domains inspired by contrastive learning by using domain adversarial learning, sample-weighted moving average centroid alignment, and discriminative feature alignment. We then estimate the label distribution of the target domain ($\widehat{P}_T$) after a

simple pre-training. Second, we utilize $\widehat{P}_T$ and the label distribution of the source domain P_S to calculate the degree of label shift. Finally, we adjust the classification prediction of target samples according to the degree of label shift during the training process. Our experiments reveal that the pseudo-labels procured by CCA-LSC consistently outperform the pseudo-labels obtained directly from the classifier's output. This observation, confirmed across all tasks on the Office-Home and DomainNet datasets, suggests that this strategy effectively enhances the reliability of pseudo-labels, thereby promoting a more accurate alignment across the two domains. See Figure 1(a), bottom.

The contributions of this article are as follows:

- Contrastive conditional alignment (CCA) leverages the principles of contrastive learning for extracting domain-invariant and class-discriminative features to resist covariate shift. And it weights samples to reduce misalignment from unreliable target pseudo labels.
- Label shift calibration (LSC) introduce a novel metric to quantify label shift and leverage this metric to rectify the classification predictions of target samples, which reduce target false pseudo-rate and resist label shift. CCA and LSC jointly resolve the IDA problem.
- Experiments were conducted on the OfficeHome and DomainNet datasets, which have both label shift and covariate shift, and it was shown that CCA-LSC achieved state-of-the-art performance.

2 Related Work

Unsupervised Domain Adaptation With Covariate Shift Covariate shift in UDA is primarily addressed by three kind of methods: statistic divergence alignment, adversarial training, and self-training. Statistic divergence alignment learns invariant features by minimizing domain discrepancy, with the divergence measure selection being key. Measures such as maximum mean discrepancy (MMD) [2,4,7,8], correlation alignment [3], wasserstein distance [9–11], marginal discrepancy measures [12], and other distance-based methods [13,14] are commonly employed. Adversarial training, taking inspiration from generative adversarial networks (GANs) [15], aims to extract domain invariant features [5,16–18] through an adversarial process. These UDA methods align the marginal distribution during training, assuming invariant label distributions. However, label shifts could lead to bad performance or even negative transfer. Self-training [19–21] employs pseudo-labels generated from the target domain for training on target domain data. However, these pseudo-labels may suffer from miscalibrated probabilities [22], potentially leading to the errors accumulation.

Unsupervised Domain Adaptation With Label Shift These techniques strive to tackle the challenge of varying label distributions across domains. Predominant strategies include class-weighting methods [23–25] and those that address cross-domain label shift by predicting and estimating the distribution

of the target label [23,26]. However, these methods presume the feature distribution is invariant across domains, only concentrating on label shift. Additional methods have investigated DA scenarios where the label spaces across domains do not entirely overlap, such as open set domain adaptation [27,28] and partial domain adaptation [29–31]. These methods pertain to specific label shift problems, which are not the focus of this paper.

Imbalanced Domain Adaptation IDA is designed to tackle the coexistence of covariate shift and label shift. Typical methods include conditional distribution alignment based on pseudo-labels [32,33], class-weighting strategies [34,35], implicit alignment methods based on sampling [36], asymmetric relaxed distribution alignment [37], and cluster-level discrepancy minimization [38]. These methods typically utilize pseudo-labels for self-training. However, under strong label shift, pseudo-labels are often unreliable, leading to error accumulation and erroneous class alignment. To address this, SENTRY [39] proposed that minimizes the entropy of reliable instances and maximizes the entropy of unreliable instances. ISFDA [40] proposed a method using secondary label correction. However, as label shift varies for different classes, unreliable instances are class-biased. These methods overlook the varying label shift across classes and do not essentially address the label shift issue. In this work, we introduce CCA-LSC. It adjusts the classification prediction of target samples based on each class's label shift degree, M_{ls}, enhancing the precision of pseudo-labels.

3 Method

3.1 Problem Setup

In this work, we investigated C-way image classification. In imbalanced domain adaptation (IDA), we are given a source domain $\mathcal{S} = \{\mathcal{X}_i^s, \mathcal{Y}_i^s\}$ with N_s labeled samples $\{(x_i^s, y_i^s)_{i=1}^{N_s}\}$ and a target domain $\mathcal{T} = \{\mathcal{X}_i^t\}$ with N_t unlabeled samples $\{(x_i^t)_{i=1}^{N_t}\}$, where the input x are images and label $y \in \{1, 2, \ldots, C\}$ are categorical variables. For the joint case of label shift and covariate shift, we adopt the same assumption in [32], i.e., $p(y|x) = q(y|x)$, $p(x) \neq q(x)$, $p(y) \neq q(y)$ and $p(x|y) \neq (x|y)$. Our goal is to learning a CNN mapping function f_t: $\mathcal{X}_t \rightarrow \mathcal{Y}_t$.

3.2 Contrastive Conditional Alignment (CCA)

Domain Adversarial Learning In domain adversarial learning, an auxiliary domain classifier **D** is employed to determine whether the features extracted by **G** are derived from the source or target domain. Simultaneously, **G** is trained to deceive **D**. When this adversarial game reaches a state of equilibrium, the features produced by G demonstrate domain invariance. Formally,

$$L_{DC}(x_s, x_t) = E_{x\sim D_S}[\log(1 - D \circ G(x))] + E_{x\sim D_T}[\log(D \circ G(x))]. \quad (1)$$

Sample-weighted Moving Average Centroid Alignment However, domain-invariance does not mean cross domain class-invariance. In [41], they propose to use moving average centroid alignment strategy. This strategy explicitly constrains the distance between centroids with identical class but different domains, ensuring close mapping of same-class features. The transfer objective is:

$$L_{SM}(x_s, y_s, x_t) = \sum_{k=1}^{K} \Phi(C_S^k, C_T^k), \tag{2}$$

where C_S^k and C_T^k represent the centroid of class K of the source and target domains respectively, and $\Phi(\cdot)$ represents the Euclidean distance between the two. This strategy is designed to mitigate the adverse effects of incorrect pseudo-labels. However, in situations with severe label shift, an excess of unreliable pseudo-labels can misalign centroids. We suggest that each sample's contribution to the centroid calculation varies based on its reliability. For instance, in a binary classification problem, if samples x_1 and x_2 have probability outputs $[0.9, 0.1]$ and $[0.6, 0.4]$ respectively, x_1 is more reliable. Hence, we use confidence as a sample weight. For a sample x, the final probability output through a deep model parameterized by θ is represented as $p_\theta(y|x)$, with a weight of $w = \max p_\theta(y|x)$.

Inspired by contrastive learning [43], the centroids with same class label but different domains should be closer, while the centroids with different class labels and domains should be futher away. We rewrite Eq.(2) as follows:

$$L_{DSM}(x_s, y_s, x_t, \widehat{y}_t) = \frac{\sum_{k=1}^{K} \Phi(C_{wS}^k, C_{wT}^k)}{\sum_{i \neq k} \Phi(C_{wS}^i, C_{wT}^k)}, \tag{3}$$

where C_{wS}^k and C_{wT}^k represent the centroids weighted by w.

Discriminative Feature Alignment In scenarios with two domains exhibiting significant distribution disparities, our goal is to ensure domain-invariant and class-discriminative features. Features with identical class labels across domains align closely, while those with different labels are distinctly separated. We propose discriminative feature alignment, a contrastive learning-based method, to facilitate this. It computes the difference between each feature pair from the source and target domains, using actual labels for the source and classifier-produced pseudo-labels for the target. Identical class labels draw features closer, while differing labels push them apart, effectively enabling cluster learning for robust classification boundaries. To avoid over-attracting unreliable samples, we persist in using w as a sample weight. Formally,

$$L_{DFA} = \frac{1/N_{same} \sum_i \sum_j \sqrt{w_i^s w_j^t} \Phi(x_i^s, x_j^t)|_{y_i^s = \widehat{y}_j^t}}{1/N_{diff} \sum_i \sum_j \sqrt{w_i^s w_j^t} \Phi(x_i^s, x_j^t)|_{y_i^s \neq \widehat{y}_j^t}} \tag{4}$$

The above strategies address covariate shift by aligning conditional distributions. However, when class imbalance is present, label shift becomes more

pronounced, leading to a biased classifier and impacting the reliability of pseudo-labels. Given the unknown target domain, for generality and simplicity, we employ class-balanced sampling on the source domain. Specifically, when selecting samples for a mini-batch, each class has an equal probability of being selected.

3.3 Label Shift Calibration (LSC)

Label Shift Metrics M_{ls} Label shift, quantifies the disparity in label distributions between source and target domains. It varies per class due to differing quantity distributions across domains. M_{ls} is defined with respect to the probability distributions P_S and P_T of the source and target domains respectively.

$$M_{ls}^i = P_T^i / P_S^i. \tag{5}$$

M_{ls}, a $1 \times C$ tensor, measures label shift, where C is the number of classes and M_{ls}^i represents the label shift degree for class i. If $M_{ls}^i = 1$, there's no label shift for class i. If $M_{ls}^i > 1$, class i is more prevalent in the target domain, and if $0 < M_{ls}^i < 1$, it's more prevalent in the source. Both cases indicate label shift, affecting pseudo-label reliability and potentially leading to error accumulation and performance degradation. We derive P_S from source labels. The unlabeled target domain's P_T is approximated using pseudo-labels, denoted as $\widehat{P}_T$.

Label Shift Calibration (LSC) In deep learning classification models, we decompose them into a feature extractor G and a classifier F. The goal of domain adaptation is to align the features extracted by G from two domains. When dealing with long-tailed source data, F tends to favor head classes due to their larger quantity, which can lead to suboptimal learning for tail classes with fewer instances. However, even if we adopt class-balanced sampling for the source domain, ensuring an unbiased F, the reliability of target sample labeling remains uncertain when the target domain follows a class-imbalanced long-tail distribution. To address this, we propose LSC based on the degree of label shift M_{ls}. LSC calibrates the classification predictions for target samples during training, making the pseudo labels more consistent with the real target data's probability distribution, thus improving the reliability of target pseudo-labels.

For a target sample through a model with parameters θ, we use $p_\theta(y|x_T)$ to represent its final probability output. The idea of LSC is to reweight $p_\theta(y|x_T)$ based on the degree of label shift M_{ls}, in order to re-estimate the target pseudo-labels. The class weighting matrix W_m is designed as:

$$W_m = \frac{1}{h_m + exp(-\sqrt{M_{ls}})}. \tag{6}$$

Then we obtain target pseudo labels after calibration and its confidence weight:

$$\widehat{y}_t^m = argmax\{p_\theta(y|x_T) \cdot W_m\}. \tag{7}$$

$$w^m = p_\theta(\widehat{y}_t^m|x_T) \tag{8}$$

A larger $M_{ls}[i]$ suggests that class i is less frequent in the source but more so in the target domain, and vice versa for a smaller $M_{ls}[i]$. As per Eq.6 and Eq.7, when a sample's feature is on the boundary of two classes and $M_{ls}[i] > M_{ls}[j]$, we prefer to label the sample as i, as shown in Figure 1. W_m bounds the class weighting values, with h_m set to 1.5, indicating that only unreliable samples at the classification boundary are calibrated to prevent over-calibration. The sample's confidence weight w^m is determined by the classifier's output, mitigating the negative effects of incorrect classification calibration.

3.4 Overall Optimization and Analysis

Overall Optimization In summary, our training process comprises two stages. The first stage involves pre-training for three epochs, utilizing high-confidence target samples from the training results to estimate the target domain's label distribution. The optimization objective of the first stage is:

$$L_{total} = L_C(x_s, y_s) + \lambda L_{DSM}(x_s, y_s, x_t, \widehat{y}_t) + \mu L_{DFA}(x_s, y_s, x_t, \widehat{y}_t) + \gamma L_{DC}(x_s, y_s) \quad (9)$$

In the second stage, we employ LSC to rectify target pseudo-labels $\widehat{y}_t^m$, and utilize $\widehat{y}_t^m$ for the training of CCA. Then our optimization objective is:

$$L_{total}^m = L_C(x_s, y_s) + \lambda L_{DSM}(x_s, y_s, x_t, \widehat{y}_t^m) + \mu L_{DFA}(x_s, y_s, x_t, \widehat{y}_t^m) + \gamma L_{DC}(x_s, y_s) \quad (10)$$

where λ and μ and γ are hyperparameters no less than zero.

Analysis Next, we demonstrate how our approach reduces the expected error on the target samples from domain adaptation theory.

Theorem 1. ([1]). *Denote $h \in \mathcal{H}$ as the hypothesis. Given two domains $\mathcal{S}$ and $\mathcal{T}$, the target error $\varepsilon_{\mathcal{T}}$ is bounded by three terms: (i) $\varepsilon_{\mathcal{S}}$: source error, (ii) $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T})$: the discrepancy distance between two distributions S and T, (iii) C_0:shared expected loss. We have:*

$$\forall h \in \mathcal{H}, \varepsilon_{\mathcal{T}}(h) \leq \varepsilon_{\mathcal{S}}(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T}) + C_0. \quad (11)$$

It is defined as $C_0 = \min_{h \in \mathcal{H}} \varepsilon_{\mathcal{S}}(h, f_{\mathcal{S}}) + \varepsilon_{\mathcal{T}}(h, f_{\mathcal{T}})$ where $f_{\mathcal{S}}$ and $f_{\mathcal{T}}$ are labeling functions for source and target domain respectively. Previous methods often assume that C_0 is negligible. However, when C_0 is large, ignoring C_0 can prevent the learning of an effective target classifier.

Theorem 2. ([41]). *According to the triangle inequality for classification error [1,42], an upper bound for C_0 is:*

$$\begin{aligned} C_0 &= \min_{h \in \mathcal{H}} \varepsilon_{\mathcal{S}}(h, f_{\mathcal{S}}) + \varepsilon_{\mathcal{T}}(h, f_{\mathcal{T}}) \\ &\leq \min_{h \in \mathcal{H}} \varepsilon_{\mathcal{S}}(h, f_{\mathcal{S}}) + \varepsilon_{\mathcal{T}}(h, f_{\mathcal{S}}) + \varepsilon_{\mathcal{T}}(f_{\mathcal{S}}, f_{\mathcal{T}}) \\ &\leq \min_{h \in \mathcal{H}} \varepsilon_{\mathcal{S}}(h, f_{\mathcal{S}}) + \varepsilon_{\mathcal{T}}(h, f_{\mathcal{S}}) + \varepsilon_{\mathcal{T}}(f_{\mathcal{S}}, f_{\widehat{\mathcal{T}}}) + \varepsilon_{\mathcal{T}}(f_{\mathcal{T}}, f_{\widehat{\mathcal{T}}}) \end{aligned} \quad (12)$$

In the given formula, the first two terms quantify the discrepancy between the hypothesis h and the source labeling function $f_{\mathcal{S}}$. Given the availability of source labels, these terms are typically minimal, facilitating the learning of a hypothesis space h that closely approximates $f_{\mathcal{S}}$. The third term measures the inconsistency between the source and pseudo-target labeling functions on target samples, while the final term indicates the divergence between the pseudo-target labeling function and the true target label, serving as a reliability measure for the pseudo-labels. Our method seeks to minimize the last two terms to optimize the upper bound of C_0. The moving average centroid alignment strategy, discussed in [41], optimizes the third term by aligning the centroids of target and source features in class C_0, ensuring prediction consistency. Our approach employs both sample-weighted moving average centroid alignment and discriminative feature alignment to foster feature alignment across different domains but within the same class, thereby minimizing the third term.

However, [41] presumes the fourth term will minimize over time and disregards it. This assumption falls short in the presence of data imbalance and label shift, where optimizing the third term could induce class bias in the pseudo-target labeling function, amplifying the fourth term. Our proposed LSC rectifies this by adjusting the classification prediction of the pseudo-target labeling function based on the label shift index M_{ls}, reducing the false pseudo-rate, and aligning the prediction with the true target data's label distribution, thereby also minimizing the fourth term. Our experiments demonstrate that LSC consistently curtails the false pseudo-rate on target samples (refer to section 4.4).

In essence, the efficacy of domain adaptation methods hinges on managing each term that could escalate the target classification error, thus broadening the applicability of domain adaptation methods.

4 Experiments

4.1 Set up

Datasets We utilized three datasets . First, we employed **Office-Home (RS-UT)**, an imbalanced version of Office-Home created by [32], where the source and target domains follow two reverse Paredo distributions. This benchmark includes three domains: Clipart (Cl), Product images (Pr), and Real-world images (Rw). The Art images (Ar) domain in Office-Home, being too small for sampling an imbalanced subset, is not considered here. Second, we used a subset of **DomainNet** created by [32], which includes 40 classes from four domains (Real (R), Clipart (C), Painting (P), Sketch (S)). As a noticeable label shift already exists, we made no additional modifications. The label distributions can be seen in Figure 1(b). **Office-31** [44] contains 4,110 images of 31 categories. The domains are Amazon (A), Webcam (W), and DSLR (D).

Baselines We benchmarked our method against eight state-of-the-art techniques that tackle both covariate shift and label shift. (i) **COAL** [32] aligns

feature and label distributions using prototype-based conditional alignment and self-training on confident pseudo-labels. (ii) **MDD+Implicit Alignment (I.A)** [36] removes explicit model parameter optimization from pseudo-labels via sampled implicit alignment. (iii) **InstaPBM** [45] employs instance-based prediction behavior matching. (iv) **F-DANN** [37] introduces a DANN based on asymmetric relaxed distribution matching. (v) **SENTRY** [39] minimizes the entropy of reliable instances and maximizes that of unreliable ones. (vi)**TIToK** [46] and (vii)**BIWAA-I** [47] and (viii)**RHWD** [48] also solve both label and feature shifting problems. All methods, except F-DANN, use target pseudo-labels. We also compared with conventional UDA methods like BBSE [23], which only addresses label shift, and MCD [17], DAN [4], DANN [5], JAN [7], BSP [49], which solely focus on covariate shift.

Implementation details All experiments are conducted using the Pytorch framework with resnet50. The model's hyper-parameters are $\lambda = 3$, $\mu = 0.6$, and $\gamma = 1$. The bottleneck layer dimension is 256, and the batch size is 50. We use the SGD optimizer with a momentum of 0.9. The initial learning rate for the classifier is 0.005 for OfficeHome and 0.01 for DomainNet, adjusted as [5]. The model trains for 20 epochs, with the first 3 forming the initial stage. After this, the model evaluates target samples and uses pseudo-labels with a confidence level of $w > 0.5$ to estimate the target domain's label distribution. The model then enters the second stage. The random seed is set to 100 for reproducibility. For imbalanced data, we use per-class mean accuracy, as suggested by [32], for a fair performance assessment.

4.2 Results

DomainNet and OfficeHome. The experimental results on DomainNet and OfficeHome are presented in Tables 1 and 2, respectively. Our method outperforms the second best method SENTRY, by improving the average accuracy by 1.90% on OfficeHome (RS-UT) and by 0.88% on DomainNet. Table 1 reveals that our method significantly surpasses SENTRY in scenarios with higher label shifts, such as R→S, P→S, and S→P, registering increases of 4.57%, 3.97%, and 3.50%, respectively. Table 2 shows a better promotion since there are severe label shift. These results highlight our method's efficacy in simultaneously tackling label shift and covariate shift.

Office-31. The experimental results are shown in Tables 3. There are few label shifts but feature shifts in this dataset. It can be seen that our method also has good performance for solving the problem of feature shifting.

Different Degrees of Label Shift. We measure imbalance using the imbalance factor IF [50], defined as the ratio of maximum to minimum class sizes. A larger IF indicates more imbalance. We created four splits on Cl→Pr with IF$\in \{1, 5, 10, 20\}$. For IF=1, we used the original Cl and Pr data from OfficeHome. For other splits, we maintained the maximum class size and adjusted

Table 1. Per-class average accuracies on DomainNet

Methods	R→C	R→P	R→S	C→R	C→P	C→S	P→R	P→C	P→S	S→R	S→C	S→P	AVG
source	65.75	68.84	59.15	77.71	60.60	57.87	84.45	62.35	65.07	77.10	63.00	59.72	66.80
MCD	61.97	69.33	56.26	79.78	56.61	53.66	83.38	58.31	60.98	81.74	56.27	66.78	65.42
DANN	63.37	73.56	72.63	86.47	65.73	70.58	86.94	73.19	70.15	85.73	75.16	70.04	74.46
F-DANN	66.15	71.80	61.53	81.85	60.06	61.22	84.46	66.81	62.84	81.38	69.62	66.50	69.52
JAN	65.57	73.58	67.61	85.02	64.96	67.17	87.06	67.92	66.10	84.54	72.77	67.51	72.48
BSP	67.29	73.47	69.31	86.50	67.52	70.90	86.83	70.33	68.75	84.34	72.40	71.47	74.09
COAL	73.85	75.37	70.50	89.63	69.98	71.29	89.81	68.01	70.49	87.97	73.21	70.53	75.89
MDD+I.A	78.54	75.09	69.43	88.50	70.59	70.44	88.37	75.71	71.65	89.35	77.97	72.41	77.33
InstaPBM	80.10	75.87	70.84	89.67	70.21	72.76	89.60	74.41	72.19	87.00	79.66	71.75	77.84
SENTRY	83.89	76.72	74.43	90.61	76.02	79.47	90.27	82.91	75.60	90.41	82.40	73.98	81.39
BIWAA-I	79.93	75.24	75.35	87.93	72.07	75.71	88.87	77.81	76.66	88.78	80.49	74.49	79.44
RHWD [48]	84.80	76.90	75.20	91.80	75.60	81.20	91.90	84.60	76.10	91.30	83.20	74.60	82.00
Ours	83.74	77.10	79.00	90.21	76.54	78.55	89.62	81.86	79.57	90.49	83.06	77.48	82.27

Table 2. Per-class average accuracies on OfficeHome (RS-UT)

Methods	Rw→Pr	Rw→Cl	Pr→Rw	Pr→Cl	Cl→Rw	Cl→Pr	AVG
source	70.74	44.24	67.33	38.68	53.51	51.85	54.39
BBSE	61.10	33.27	62.66	31.15	39.70	38.08	44.33
MCD	66.03	33.17	62.95	29.99	44.47	39.01	45.94
DAN	69.35	40.84	66.93	34.66	53.55	52.09	52.90
DANN	71.62	46.51	68.40	38.07	58.83	58.05	56.91
F-DANN	68.56	40.57	67.32	37.33	55.84	53.67	53.88
JAN	67.20	43.60	68.87	39.21	57.98	48.57	54.24
COAL	73.65	42.58	73.26	40.61	59.22	57.33	58.40
MDD+I.A	76.08	50.04	74.21	45.38	61.15	63.15	61.67
InstaPBM	75.56	42.93	70.30	39.32	61.87	63.40	58.90
SENTRY	76.12	56.80	73.60	54.75	65.94	64.29	65.25
TIToK	77.09	52.84	72.15	44.32	60.06	59.95	61.07
Ours	79.18	60.53	78.26	50.13	65.79	68.99	67.15

the Pareto distribution parameters based on OfficeHome (RS-UT). All methods used class-balanced sampling in the source domain for fairness. As shown in Figure 2(a), accuracy decreases for all methods with increasing imbalance due to label shift, but our method consistently outperforms the others.

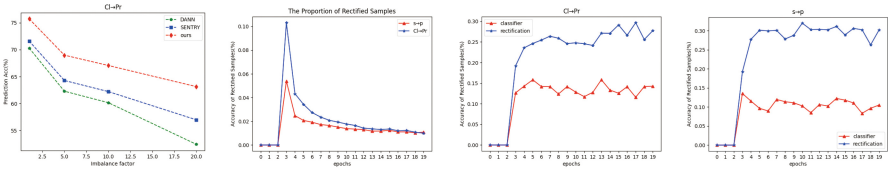
4.3 Ablation Study

To mitigate the influence of source data imbalance, we evaluated each domain adaptation component using class-balanced sampling on the source domain. Table 4 presents the results. Model performance is bad with only source cross-entropy loss. Performance improves with the addition of adversarial learning and sample-weighted moving average centroid alignment loss (L_{DC}+L_{DSM}). Significant improvement is observed with the inclusion of discriminative feature alignment loss (L_{DFA}), which ensures both domain invariance and class discriminability of the learned representation. Label shift calibration on target samples

Table 3. Accuracy results on Office-31 dataset.

Methods	A→W	D→W	W→D	A→D	D→A	W→A	AVG
DAN	68.5	96.0	99.0	67.0	54.0	53.1	72.9
DANN	82.0	96.9	99.1	79.7	68.2	67.4	82.2
MCD	88.6	98.5	100.	92.2	69.5	69.7	86.5
MDD	94.5	98.4	100.	93.5	74.6	72.2	88.9
BIWAA-I	95.6	99.0	100.	95.4	75.9	77.3	90.5
Ours	96.0	99.1	100.	94.6	77.1	77.3	90.7

further enhances performance by reducing the target false pseudo rate during training, ensuring correct execution of the two pseudo label-based strategies.



(a) (b) (c) (d)

Fig. 2. Analysis of label shift calibration. (a) Accuracy under different degrees of imbalance on Cl→Pr. (b) The proportion of target samples with calibrated pseudo labels($\widehat{y} \neq \widehat{y}^m$) via CCA-LSC. (c) and (d) :The accuracy of the target pseudo labels $\widehat{y}$ (obtained by the classifier) and $\widehat{y}^m$ (calibrated based on label shift metric M_{ls}) in all calibrated target samples ($\widehat{y} \neq \widehat{y}^m$) on Cl→Pr and S→P respectively.

4.4 Analysis of Label Shift Calibration

The label shift calibration strategy calibrates only some target pseudo labels at the classification boundary, leading to two scenarios: consistency ($\widehat{y} = \widehat{y}^m$) and inconsistency ($\widehat{y} \neq \widehat{y}^m$) between the classifier's output pseudo labels and the calibrated ones. Figure 2(b) illustrates the proportion of samples with calibrated pseudo labels ($\widehat{y} \neq \widehat{y}^m$) during training, which decreases over time, indicating an increasing number of samples moving away from the classification boundary. Figures 2(c) and 2(d) show the right proportion of $\widehat{y}$ and $\widehat{y}^m$ in these calibrated samples. During the initial 3 epochs of pre-training, label shift calibration is not applied. Throughout the training, the accuracy of $\widehat{y}^m$ consistently surpasses that of $\widehat{y}$, demonstrating the strategy's effectiveness in reducing the false pseudo rate of the classifier's target output, supporting the analysis in Section 3.4. In fact, higher accuracy of $\widehat{y}^m$ over $\widehat{y}$, was observed in all 18 transfer tasks on OfficeHome and DomainNet during training.

Table 4. Ablation Study: effectiveness of adaptation components. We adopt class-balanced sampling on the source domain to counteract the adverse effect caused by imbalance to examine the effectiveness of each component.

Methods	Cl→Pr	S→P	P→S
source	51.85	63.00	65.07
L_C+L_{DC}	62.23	71.22	73.32
L_C+L_{DC}+L_{DSM}	63.37	72.80	75.08
L_C+L_{DC}+L_{DSM}+L_{DFA}	66.57	76.74	77.35
L_C+L_{DC}+L^m_{DSM}+L^m_{DFA}	68.99	77.48	79.57

Table 5. The influence of h_m

h_m	1	1.5	2
Cl→Pr	68.32	68.99	68.44

Table 6. Hyper-parameter sensitivity on Cl→Pr

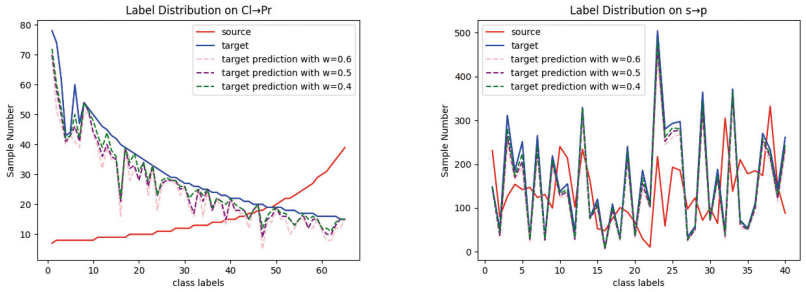
λ \ μ	0.4	0.6	0.8
1	67.18	67.73	68.06
3	68.22	68.99	68.34
5	68.56	68.11	67.16

A question naturally arises: given the label shift between source and target domains, could we diminish this shift by implementing pseudo-label balanced sampling on the target domain and class-balance sampling on the source domain? Initially, the balanced sampling strategy curbs imbalance by regulating the utilization of input data, inevitably leading to an over-sampling of certain classes. This is more likely to negatively impact the quality of the learned representation for unlabeled target domain data. Furthermore, our application of pseudo-label balanced sampling on the OfficeHome dataset resulted in a reduction of per-class accuracy by about 1%. Consequently, we have decided not to use pseudo-label balanced sampling strategy on target data in our method.

4.5 Hyper-parameter Discussion

The Influence of the Parameter h_m. The h_m dictates the proportion of calibrated samples. A smaller h_m value leads to a larger proportion of $\widehat{y} \neq \widehat{y}^m$. Although our calibration strategy is effective, more calibrations aren't always better. Over-calibration can lead to over-representation of the dominant class in target samples, while under-calibration can lessen its effectiveness. Table 5 illustrates the impact of the h_m. To counteract the effects of incorrect calibrations, we derive the confidence of all target pseudo labels from the classifier's probability output. For instance, if a sample's probability output is [0.6, 0.4] and the calibrated output is [0.45, 0.55], its confidence is 0.4 and its weight $w = 0.4$. This can effectively reduce the adverse impact of incorrect calibrations.

Hyper-parameter Analysis. We fixed γ to 1 and discussed the impact of λ and μ. The experimental results are shown in the Table 6. It can be seen that our experimental results are not sensitive to each hyperparameter.



(a) (b)

Fig. 3. The impact of selecting target samples with different confidence levels on the estimation of target label distribution

4.6 Analysis of Two Stage Learning

Our LSC strategy relies on the distribution estimation of the target domain in the first stage. When this estimation is highly unreliable, the LSC strategy may fail. Therefore, we discuss the impact of the pre-training of CCA in the first stage on the LSC strategy in the second stage. Figure 3 shows the results of estimating the target domain distribution by selecting pseudo-labels of target samples with different confidence levels. It can be seen that when the confidence level $w > 0.4$, $w > 0.5$, $w > 0.6$, our estimated distribution of the target domain is generally close to its true distribution, indicating that $\widehat{P}_T$ is reliable. In fact, our estimation of the target domain distribution does not need to be very accurate, as long as it can generally reflect the target label distribution. In our experiments, we use pseudo-labels of target samples with a confidence level of $w > 0.5$ to estimate the target domain distribution.

5 Conclusion

We introduce CCA-LSC to tackle label shift and covariate shift in imbalanced domain adaptation. Our approach employs domain adversarial learning, sample-weighted moving average centroid alignment, and discriminative feature alignment for contrastive conditional alignment, facilitating the learning of feature representations that are both domain-invariant and class-discriminative. To counter label shift, we introduce the label shift measure M_{ls}, using it to calibrate the classification prediction of target samples. Experimental evidence demonstrates that CCA-LSC delivers state-of-the-art results on benchmark datasets.

Acknowledgements. The paper is supported by the National Natural Foundation Science of China (62101061).

References

1. Ben-David, S., Blitzer, J., Crammer, K., et al.: A theory of learning from different domains[J]. Mach. Learn. **79**, 151–175 (2010)
2. Long M, Zhu H, Wang J, et al. Unsupervised domain adaptation with residual transfer networks[J]. Advances in neural information processing systems, 2016, 29
3. Sun, B., Saenko, K.: Deep CORAL: Correlation Alignment for Deep Domain Adaptation. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9915, pp. 443–450. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-49409-8_35
4. Long M, Cao Y, Wang J, et al. Learning transferable features with deep adaptation networks[C]//International conference on machine learning. PMLR, 2015: 97-105
5. Ganin, Y., Ustinova, E., Ajakan, H., et al.: Domain-adversarial training of neural networks[J]. J. Mach. Learn. Res. **17**(59), 1–35 (2016)
6. Peng X, Bai Q, Xia X, et al. Moment matching for multi-source domain adaptation[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 1406-1415
7. Long M, Zhu H, Wang J, et al. Deep transfer learning with joint adaptation networks[C]//International conference on machine learning. PMLR, 2017: 2208-2217
8. Ge, P., Ren, C.X., Xu, X.L., et al.: Unsupervised domain adaptation via deep conditional adaptation network[J]. Pattern Recogn. **134**, 109088 (2023)
9. Balaji Y, Chellappa R, Feizi S. Normalized wasserstein for mixture distributions with applications in adversarial learning and domain adaptation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 6500-6508
10. Lee J, Raginsky M. Minimax statistical learning with wasserstein distances[J]. Advances in Neural Information Processing Systems, 2018, 31
11. Shen J, Qu Y, Zhang W, et al. Wasserstein distance guided representation learning for domain adaptation[C]//Proceedings of the AAAI conference on artificial intelligence. 2018, 32(1)
12. Zhang Y, Liu T, Long M, et al. Bridging theory and algorithm for domain adaptation[C]//International conference on machine learning. PMLR, 2019: 7404-7413
13. Li, J., Jing, M., Lu, K., et al.: Locality preserving joint transfer for domain adaptation[J]. IEEE Trans. Image Process. **28**(12), 6103–6115 (2019)
14. Li, J., Jing, M., Su, H., et al.: Faster domain adaptation networks[J]. IEEE Trans. Knowl. Data Eng. **34**(12), 5770–5783 (2021)
15. Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[J]. Advances in neural information processing systems, 2014, 27
16. Long M, Cao Z, Wang J, et al. Conditional adversarial domain adaptation[J]. Advances in neural information processing systems, 2018, 31
17. Saito K, Watanabe K, Ushiku Y, et al. Maximum classifier discrepancy for unsupervised domain adaptation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 3723-3732
18. Rangwani H, Aithal S K, Mishra M, et al. A closer look at smoothness in domain adversarial training[C]//International conference on machine learning. PMLR, 2022: 18378-18399
19. Mei, K., Zhu, C., Zou, J., Zhang, S.: Instance Adaptive Self-training for Unsupervised Domain Adaptation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12371, pp. 415–430. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58574-7_25
20. Wei C, Shen K, Chen Y, et al. Theoretical analysis of self-training with deep networks on unlabeled data[J]. arXiv preprint arXiv:2010.03622, 2020

21. Zou Y, Yu Z, Liu X, et al. Confidence regularized self-training[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 5982-5991
22. Guo C, Pleiss G, Sun Y, et al. On calibration of modern neural networks[C]//International conference on machine learning. PMLR, 2017: 1321-1330
23. Lipton Z, Wang Y X, Smola A. Detecting and correcting for label shift with black box predictors[C]//International conference on machine learning. PMLR, 2018: 3122-3130
24. Azizzadenesheli K, Liu A, Yang F, et al. Regularized learning for domain adaptation under label shifts[J]. arXiv preprint arXiv:1903.09734, 2019
25. Azizzadenesheli, K.: Importance weight estimation and generalization in domain adaptation under label shift[J]. IEEE Trans. Pattern Anal. Mach. Intell. **44**(10), 6578–6584 (2021)
26. Alexandari A, Kundaje A, Shrikumar A. Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation[C]//International Conference on Machine Learning. PMLR, 2020: 222-232
27. Panareda Busto P, Gall J. Open set domain adaptation[C]//Proceedings of the IEEE international conference on computer vision. 2017: 754-763
28. Yang, X., Deng, C., Liu, T., et al.: Heterogeneous graph attention network for unsupervised multiple-target domain adaptation[J]. IEEE Trans. Pattern Anal. Mach. Intell. **44**(4), 1992–2003 (2020)
29. Cao Z, Ma L, Long M, et al. Partial adversarial domain adaptation[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 135-150
30. Zhang J, Ding Z, Li W, et al. Importance weighted adversarial nets for partial domain adaptation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 8156-8164
31. Zhang, Y., Ji, J.C., Ren, Z., et al.: Digital twin-driven partial domain adaptation network for intelligent fault diagnosis of rolling bearing[J]. Reliability Engineering & System Safety **234**, 109186 (2023)
32. Tan, S., Peng, X., Saenko, K.: Class-Imbalanced Domain Adaptation: An Empirical Odyssey. In: Bartoli, A., Fusiello, A. (eds.) ECCV 2020. LNCS, vol. 12535, pp. 585–602. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-66415-2_38
33. Tachet des Combes R, Zhao H, Wang Y X, et al. Domain adaptation with conditional distribution matching and generalized label shift[J]. Advances in Neural Information Processing Systems, 2020, 33: 19276-19289
34. Wang J, Chen Y, Hao S, et al. Balanced distribution adaptation for transfer learning[C]//2017 IEEE international conference on data mining (ICDM). IEEE, 2017: 1129-1134
35. Yan H, Ding Y, Li P, et al. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2272-2281
36. Jiang X, Lao Q, Matwin S, et al. Implicit class-conditioned domain alignment for unsupervised domain adaptation[C]//International conference on machine learning. PMLR, 2020: 4816-4827
37. Wu Y, Winston E, Kaushik D, et al. Domain adaptation with asymmetrically-relaxed distribution alignment[C]//International conference on machine learning. PMLR, 2019: 6872-6881
38. Yang J, Yang J, Wang S, et al. Advancing imbalanced domain adaptation: Cluster-level discrepancy minimization with a comprehensive benchmark[J]. IEEE Transactions on Cybernetics, 2021

39. Prabhu V, Khare S, Kartik D, et al. Sentry: Selective entropy optimization via committee consistency for unsupervised domain adaptation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 8558-8567
40. Li X, Li J, Zhu L, et al. Imbalanced source-free domain adaptation[C]//Proceedings of the 29th ACM international conference on multimedia. 2021: 3330-3339
41. Xie S, Zheng Z, Chen L, et al. Learning semantic representations for unsupervised domain adaptation[C]//International conference on machine learning. PMLR, 2018: 5423-5432
42. Crammer K, Kearns M, Wortman J. Learning from Multiple Sources[J]. Journal of Machine Learning Research, 2008, 9(8)
43. Hjelm R D, Fedorov A, Lavoie-Marchildon S, et al. Learning deep representations by mutual information estimation and maximization[J]. arXiv preprint arXiv:1808.06670, 2018
44. Saenko, K., Kulis, B., Fritz, M., Darrell, T.: Adapting Visual Category Models to New Domains. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6314, pp. 213–226. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15561-1_16
45. Li B, Wang Y, Che T, et al. Rethinking distributional matching based domain adaptation[J]. arXiv preprint arXiv:2006.13352, 2020
46. Wang, Y., Chen, Q., Liu, Y., et al.: TIToK: A solution for bi-imbalanced unsupervised domain adaptation[J]. Neural Netw. **164**, 81–90 (2023)
47. Westfechtel T, Yeh H W, Meng Q, et al. Backprop induced feature weighting for adversarial domain adaptation with iterative label distribution alignment[C]//Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2023: 392-401
48. Si, L., Dong, H., Qiang, W., et al.: Regularized hypothesis-induced wasserstein divergence for unsupervised domain adaptation[J]. Knowl.-Based Syst. **283**, 111162 (2024)
49. Chen X, Wang S, Long M, et al. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation[C]//International conference on machine learning. PMLR, 2019: 1081-1090
50. Cui Y, Jia M, Lin T Y, et al. Class-balanced loss based on effective number of samples[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 9268-9277

A New StyleGAN Latent Space Based Model for Image Style Transfer

Rakesh Dey[1], Shivakumara Palaiahnakote[2(✉)], Saumik Bhattacharya[3], Sukalpa Chanda[4], and Umapada Pal[1]

[1] Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata, Baranagar, India
umapada@isical.ac.in
[2] School of Science, Engineering and Environment, University of Salford, Salford, UK
s.palaiahnakote@salford.ac.uk
[3] Department of Electrical and Electronics Communication, IIT-Kharagpur, Kharagpur, India
saumik@ece.iitkgp.ac.in
[4] Østfold University College, Halden, Norway
sukalpa@ieee.org

Abstract. Cross-domain image style transfer task is an attractive topic for several applications, such as image-to-image style transfer, text-to-image style transfer, artistic image generation, etc. In cross-domain image style transfer tasks (e.g., image-to-image style transfer, artistic image-to-image style transfer, text-to-image style transfer, etc.), training becomes cumbersome due to differences in data distribution across domains and complex model architectures. Unlike existing domain adaptation and domain-independent methods that focus on robust and sufficient feature extraction, this work focuses on disentangling the latent space through latent optimization. For this purpose here we propose a new idea of styled image generation from the latent space of StyleGAN which works well for image-to-image and text-to-image style transfer. We critically analyzed the low-dimensional latent structure and its effect on cross-domain image style transfer tasks and finally proposed a method along with a latent optimizing procedure to overcome the problem of style transfer. The experimental results on different standard datasets show that the proposed model is robust, effective, and generic compared to the state-of-the-art models.

Keywords: Generative adversarial network · Optimization · StyleGAN · Style transfer · Morphing

1 Introduction

With the advancement in artificial neural networks and their variants, artistic image generation, image style transfer, and synthesis have undergone a transformative journey. The recent sophisticated generative models can efficiently generate high-fidelity visual content that is indistinguishable from reality. These advancements have elevated image synthesis from an intriguing theoretical concept to a practical tool with real-world

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15311, pp. 29–45, 2025.
https://doi.org/10.1007/978-3-031-78195-7_3

applications. Traditional artistic image generation relies on conventional methodologies such as feature modulation, thin plate spline interpolation [1, 2], etc. In the same way, recently, to alleviate the dependencies of large amounts of training data and humongous computation power, GAN inversion methods have been proposed [3–8]. The core idea of these approaches is to find a properly optimized latent vector in the latent space that is the most suitable for performing a given particular style transfer task. These approaches may work well for the data of the same domains but not for cross-domains, which include image and text domains for image style transfer.

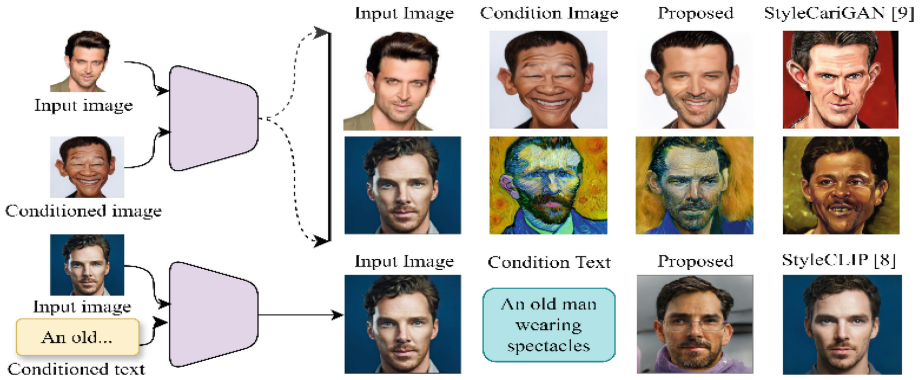


Fig. 1. Block diagram of proposed method. It can be observed that the proposed method successfully generates samples of superior quality compared to the existing methods.

The performance of the state-of-the-art model [8, 9] is illustrated in Fig. 1, where for the candidate image and condition image including the text domain, the existing method does not output the correct images which are close to the input candidate image. When we look at the results presented in the last row in Fig. 2, the existing method [8] does not fix the reading glass over the output while the proposed method fixes the reading glass for the candidate of "An old man wearing spectacles". The key reasons for the poor results of the existing method are the cross-domain gap and lack of generalization ability. Therefore, the proposed model works well, especially for textual guidance. This observation motivated us to propose a new method to explore low-level latent space for improving the image style transfer performance. Inspired by the success of StyleGAN, which can generate different variant images of input depending on candidate images, we explore StyleGAN for generating correct output through developing latent space and a new approach to the loss function.

As noted from the related work, there are models that work well for image to image and there are other methods which work well for text to image. But there are a few methods which perform both the tasks, image to image and text to image. However, these methods are not robust to adverse situations and the method lacks generalization. Therefore, developing a new model for addressing the challenges of both image to image and text to image irrespective of adverse situations is an open issue. This is the motivation to propose a new idea that exploits styled image generation from the latent

space of StyleGAN for performing image to image and text to image successfully. This is the key contribution of the proposed work compared to the state-of-the-art methods.

Therefore, the key contributions of the proposed work are as follows. (i) For cross-domain image manipulation, we propose a simple and effective architecture with a novel optimization procedure that involves creating the latent-space disentanglement and taking advantage of it to find the optimal point in cross-domain latent space. (ii) Exploring CLIP for generating proper output through textual guidance.

The paper is organized as follows: Sect. 2 provides a comprehensive review of the foundational works on image synthesis, tracing its evolution from early experiments to modern deep-learning approaches. In Sect. 3, we present a new architecture and propose a novel optimization scheme to solve the problem. Experimental results on different datasets and tasks are conducted in Sect. 4. Finally, we conclude the effectiveness of the method by highlighting the potential future directions in Sect. 5.

2 Related Work

2.1 Image Style Transfer

Image style editing and manipulation have achieved unprecedented progress in recent years. This objective includes image morphing [10], image style transfer [11, 12], text-to-image generation and manipulation, etc. Modern sophisticated deep learning algorithms such as GANs [2], transformers [12], diffusion models [13], etc. can successfully transfer different styles. With these sophisticated architectures, many critical tasks such as image restoration [14], semantic editing [15], style transfer, and photo-realistic rendering have achieved an unforeseen advancement. The models work well for particular type of images or dataset but not different datasets of applications.

2.2 Generative Models

In conventional GANs the generator takes a noise vector as an input to produce new samples and the discriminator tries to differentiate the real and the generated samples. The progressive architectural and training methodological refinement [16, 17] has significantly contributed to the advancement of numerous researches such as image super-resolution [18], out-of-domain image synthesis [19], conditional image synthesis [20, 21], text-to-image synthesis etc. However, as the existing generative models learn the underlying data distribution of a particular domain, they often lack the ability of fine-grained editing control in the case of cross-domain setups. Instead of a random noise vector zz, StyleGAN, and its variants [22] focus on style injection in progressive fine-grained image generation. With the improvement in StyleGAN, many pieces of research have been conducted such as conditioned attributes sampling editing [23], domain-gap analysis, disentangling variation in video and text-driven image synthesis and editing [24], etc. are names of a few. Additionally, residual feature-based transformation [25] to encode finer details, and CLIP adapter [26] based approaches have extended StyleGAN's efficiency in text to image editing. In summary, the models are focused on image to image but not both image to image and text to image.

2.3 Diffusion Models

Recently diffusion models [13, 27] have shown unprecedented success in text to image T2I and image to image I2I generation. For T2I, Gal et al. [28] encoded the subject in the latent space of a pretrained CLIP and then the noise scheduling and denoising were performed based on a given text prompt. Although this method can generate overall structure, in case of particular structural identity e.g. face, hairstyle, etc., the model fails to generate faithful samples. For I2I, diffusion-based generation method [29] mainly encodes a particular art style as a textual description as textual embedding and guides the generation process in the denoising step. But the limitation of this model is that the perceptual guidance of a particular style may not be optimal as our proposed method as shown in the figure. On the other hand, encoding in StyleGAN latent space in our proposed method not only generates samples with superior perceptual quality, but also preserves the identity of the subject. Although the methods perform both image to image and text to image, the results are not consistent and stable for both tasks.

In summary, although the existing models perform various forms of image style transfer in cross-domain, they mainly focus on learning from huge data distribution and explicit architectural design for domain adaptation and hence lack generalization.

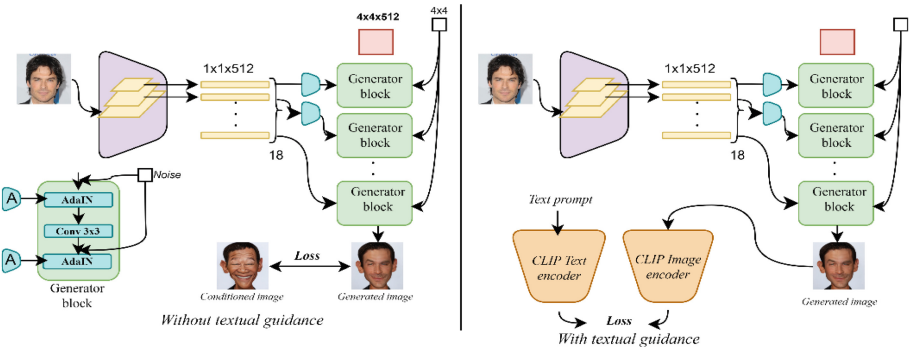


Fig. 2. Architecture diagram of the proposed method. Input image tensor $\mathcal{I}^{b\times C\times H\times W}$ is fed into the *pSp* encoder which provides latent of shape $b \times 18 \times 512$. This tensor is then sent to the generator blocks of StyleGAN and loss is computed between the conditioned and generated image. In the case of textual guidance, the loss is computed by incorporating the CLIP text encoder and CLIP image encoder.

3 Proposed Methodology

Given a set of input candidate images of real domain and condition images of different domains such as paint, cartoon, etc., our goal is to perform different image style transfer tasks such as transferring the style of the target image to that of the input candidate image, guiding the shape of the input images according to the condition images, etc. While transferring the shape of the target image, our main objective is to preserve the

identity of the input image. In addition, for textual style transfer, we aim to find a specific latent direction in a disentangled $\mathcal{W}^+$ space where the textual guidance can change a specific attribute of the image.

As shown in the complete architecture of the proposed work in Fig. 3, the input and candidate images are sent to Pixel2Style2Pixel (*pSp*) [30] encoder to obtain initial latent in the $\mathcal{W}^+$ latent space. Then that latent is fed to the generator blocks of StyleGAN and finally the loss is estimated with the help of image and textual encoder of the CLIP. Overall, the combination of *pSp* encoder, StyleGAN and CLIP enables the model to achieve the best accuracy with robustness and generic properties.

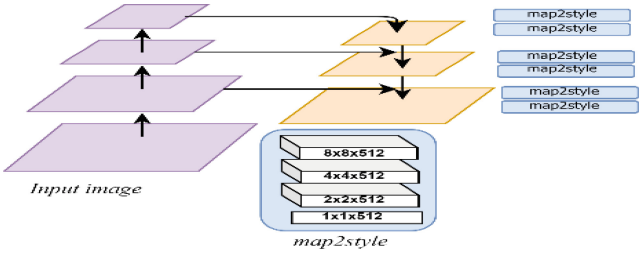


Fig. 3. Hierarchical architecture of the encoder (*pSp*) for latent initialization.

3.1 Creating Latent Space Using Pixel2Style2Pixel (pSp) Encoder

Image encoder: Conventional GANs are capable of generating photo-realistic images of a given data distribution by taking a noise vector $z \in \mathcal{Z}$ from a Gaussian distribution. But, in the case of facial portrait image generation, it has been observed that, although the random noise sample can generate a photo-realistic portrait, it often fails to faithfully reconstruct the identity of the main input image. Moreover, this unstructured random latent fails to encode attribute-specific information in its latent space making it highly entangled. To alleviate these preliminary problems, we aim to sample a latent z.

in such a way that while generating the image, this latent will infuse important information about the person-specific identity and hopefully contain attribute-specific information in its latent space in a disentangled manner.

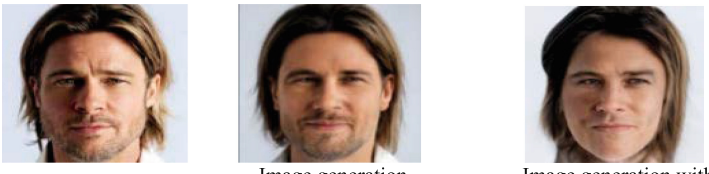


Fig. 4. The effect of the optimization process is illustrated. It is noted that optimization using the latent obtained from *pSp* encoder gives better generation result

Hence, for encoding the input images, the *pSp* encoder is used and it is illustrated in Fig. 4, which is trained on FFHQ faces [31]. As an input, the *pSp* encoder takes a tensor $\mathcal{I}^{b\times C\times H\times W}$ where b, C, H, W are the batch size, channel, height, and width respectively. Each element of the batch b is then mapped by the *pSp* encoder into the $\mathcal{W}^+$ latent manifold. Hence, for an input tensor $\mathcal{I}^{b\times C\times H\times W}$ the *pSp* encoder gives $w \in \mathbb{R}^{b\times n\times d}$ tensor as the output. This tensor $w \in \mathbb{R}^{b\times n\times d}$ represents the initial latent representation of the images. The effectiveness of hierarchical encoding through the *pSp* encoder can be observed in Fig. 5 where it clearly shows that the image generated from *pSp* encoded latent can reconstruct a high-fidelity sample compared to the image generated from randomly initialized latent.

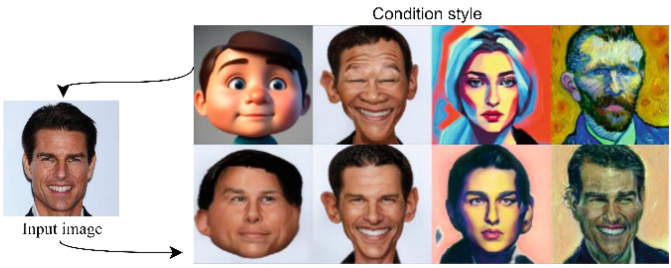


Fig. 5. The effect of style loss is illustrated. Here, the leftmost image is the candidate image and top row images are the conditional images. Second row images show the results of the respective image style transfer of the proposed method using style loss.

Generator: For synthesizing the images from their latent representations and editing them in a different style domain such as caricature, Pixar, artistic, etc. a StyleGAN generator [31] is used. To guide the style and the shape of the target domain to the input images, this generator is pre-trained with the WebCari-A dataset. As discussed earlier this generator is composed of 18 sequences. The generator takes input $w^{b\times n\times d}$ and gives $\mathcal{I}_g^{b\times C\times H\times W}$. For each progressive up-sampling layer of the StyleGAN, it takes w^{512} latent vector which controls the style-related characteristics of the generated images. For this task, the w^{512} latent vector is passed through an affine transformation layer to generate a style code s. This style code infuses the style information by implementing *adaptive instance normalization.* Moreover, for statistically varying features such as moles, etc., a noise vector n is added in each progressive layer. Next, optimization is done by comparing $\mathcal{I}_{real}^{b\times C\times H\times W}$ and $\mathcal{I}_{gen}^{b\times C\times H\times W}$. To fulfill the objective function, we propose a novel and non-trivial optimization scheme. The optimization procedure will be discussed in the subsequent section.

3.2 Image and Textual Encoder for Deriving Optimization Procedure

Optimization Procedure: In this work, we follow the optimization-based GAN inversion procedure. To find out the optimized latent w, we use the perceptual loss as defined

in Eq. (1).

$$\mathcal{L}^{\mathcal{G}}_{p\_\mu} = \frac{1}{h_\mu w_\mu c_\mu} \sum_{x=1}^{h_\mu} \sum_{y=1}^{w_\mu} \sum_{z=1}^{c_\mu} \left|\left| \phi_\mu\left(\mathcal{I}^j_{real}\right) - \phi_\mu\left(\mathcal{I}^j_{gen}\right) \right|\right| \tag{1}$$

where $\mathcal{L}^{\mathcal{G}}_{p\_\mu}$ represents the perceptual loss computed from the μ^{th} layer of a pre-trained VGG-16 network. ϕ_μ denotes the output of the μ^{th} layer having a feature space dimension $h_\mu \times w_\mu \times c_\mu$. For transferring the style information of the target image $\mathcal{I}_{style}$ to the content image $\mathcal{I}_{real}$, for each generation iteration style loss is calculated as defined in Eq. (2).

$$\mathcal{L}_{style} = \frac{1}{N_l} \sum_{i=0}^{N_l} \left|\left| \mu\left(\phi_i\left(\mathcal{I}_{gen}\right) - \mu\left(\phi_i\left(\mathcal{I}_{style}\right)\right)\right)\right|\right|_2 + ||\sigma(\phi(\mathcal{I}_0) - \sigma(\phi(\mathcal{I}_s)))||_2 \tag{2}$$

where $\phi_i(\cdot)$ represents features extracted from a i^{th} layer of a pre-trained VGG-19 model and N_l represents the number of layers. $\mu(\cdot)$ and $\sigma(\cdot)$ are the mean and standard deviation respectively.

Fig. 6. Example of the identity generation with ID loss. (a) Input image, the result of (b) existing StycarGAN [9] method and (c) proposed method. From the images, the identity generation of the proposed method is better than the existing method.

Although the perceptual loss guarantees visual fidelity and the style loss successfully transfers the style of the target image to the content image, to capture identity-specific information, ID loss is also incorporated. The ID loss can be written mathematically as defined in Eq. (3). Figure 6 shows that style loss can guide condition style successfully while generating through the proposed optimization process.

$$\mathcal{L}_{id} = 1 - CosineSim\left(ArcFace\left(\mathcal{I}_{gen}\right), ArcFace(\mathcal{I}_{real})\right) \tag{3}$$

where, $ArcFace(\cdot)$ denotes the features extracted from the *ArcFace* [32] model which is a popular face recognition model. Here, the facial features from the *ArcFace* are extracted and the cosine similarity of them is calculated and subtracted from 1. For controlling the textual guidance, while generating, the CLIP loss is used as shown in Eq. (4).

$$\mathcal{L}_{CLIP} = 1 - CosineSim(\mathcal{E}_{TEXT}(text), \mathcal{E}_{IMAGE}(image)) \tag{4}$$

Loss: Finally, the loss is derived as defined in Eq. (5).

$$\mathcal{L} = \lambda_{percept}\mathcal{L}_{percept} + \lambda_{style}\mathcal{L}_{style} + \lambda_{id}\mathcal{L}_{id} + \lambda_{CLIP}\mathcal{L}_{CLIP}||\sigma(w)|| \tag{5}$$

where, $||\sigma(\cdot)||$ is the standard error for 18×512 vectors which is used as a regularizer. $\lambda_{percept}$, λ_{style}, λ_{id}, and λ_{CLIP} are weights of the corresponding perceptual loss, style loss, ID loss and CLIP loss respectively. Finally, for style transfer, the preservation of the identity of the candidate image is very crucial. Figure 6 shows that the proposed methods can preserve identity while transferring conditioned style.

4 Experimental Results

4.1 Datasets and Evaluation

For the experiment, the FFHQ dataset is divided into 9:1 ratio for training and testing purposes. The WebCariA dataset has 252 different identity classes containing a variable number of styled images and real images. For our task, we used all styled images of every class i.e. a total of 6042 styled images were used for learning the image style. For inference, we took 1000 images chosen randomly from the WebCariA dataset and used each of them as the conditional image for style guidance.

Implementation Details: We used Nvidia GeForce GTX 1080 GPU. For optimization in stage 3.2, we use $\lambda_{percept} = 1$, $\lambda_{id} = 0.1$. For image-to-image style transfer tasks-related experiments, we use $\lambda_{style} = 0.4$. While text-to-image style transfer task, we use $\lambda_{CLIP} = 0.5$. We set the iteration number to 360 to optimize the objective function. In an input batch of 16, to optimize each image in the batch, it took roughly 2 min. In this experiment, the *pSp* encoder which is pre-trained on the FFHQ dataset is used. This encoder takes a batch input $\mathcal{I}^{b\times C\times H\times W}$ where $b = 16, C = 3, H = 256, W = 256$. This encoder maps this input tensor to the $\mathcal{W}^+$ latent space. The input is mapped to the w latent where $w^{b\times n\times d}$ with $b = 16, n = 18, d = 512$. This $w^{b\times n\times d}$ acts as the initial starting point in the latent space. Next this latent is fed into the StyleGAN generator which has been trained on the WebCari-A dataset. This generator takes this input and sequentially constructs the output tensor $\mathcal{O}^{b\times c\times H\times W}$. To find the proper latent which faithfully reconstructs the image, an optimization process is completed. We use perceptual loss, style loss, ID loss, CLIP loss, and noise regularization loss for the optimization procedure. Adam [33] optimizer with $lr = 0.2$ is used.

4.2 Ablation Study

4.2.1 Loss Functions

For FFHQ dataset training, the model is trained with perceptual loss, and ID loss. Downstream tasks like image style transfer, text to image editing, image morphing, etc. The results reported in Table 1 shows that each loss listed in Table 1 contributes equally for achieving the best results. It is also inferred that single loss is not sufficient to achieve the best results compared to the combined loss of the proposed method.

4.2.2 GAN Inversion

For an effective cross-domain style transfer task, a specific setup for GAN inversion, identity regeneration, and disentangled latent analysis is shown in Table 2 which presents

Table 1. Effectiveness of the different loss function on FFHQ

$\mathcal{L}$				Metrics			
$\mathcal{L}_{percept}$	$\mathcal{L}_{id}$	$\mathcal{L}_{style}$	$\mathcal{L}_{CLIP}$	Id score↑	SSIM ↑	PSNR ↑	CLIP score ↑
✓	×	×	×	0.788	0.640	15.72	0.603
×	✓	×	×	0.826	0.668	14.04	0.487
×	×	✓	×	0.581	0.410	10.31	0.313
×	×	×	✓	0.653	0.580	13.44	0.792
✓	✓	✓	✓	**0.961**	**0.729**	**18.07**	**0.918**

the quantitative validation for the specific design choice of the proposed model. We first verify the proposed optimization-based inversion technique to validate the faithfulness of the proposed pipeline toward the end goal of image style editing such as cross-domain style transfer, shape guidance, image morphing, etc. Two types of initializations are explored:

Table 2. Id score comparison for different setups

	Inversion		Identity regeneration	Disentanglement analysis					
	Random	pSp		1–4	4–8	8–12	12–18	10–14	14–18
ID_metric	0.766	0.924	0.961	0.880	0.761	0.749	0.827	0.918	0.921

Random vs pSp Initialization: It is observed that latent obtained from *pSp* provides better initialization than the random setup. To get the initial latent w_i of an input image $\mathcal{I}$, we first pass it through the *pSp* encoder $\mathcal{E}(\cdot)$. Then w_i is optimized in an iterative process to find a latent that can accurately reconstruct the image $\mathcal{I}$.

Input image StyleCariGAN [9] Proposed model

Fig. 7. Examples of identity regeneration in style transfer tasks. It can be observed that the proposed model outperforms the existing model in terms of identity generation for style transfers task.

4.2.3 Identity Regeneration

We are particularly interested in the question – '*can the conventional ID loss $\mathcal{L}_{id}$ preserve identity?*' The conventional ID loss $\mathcal{L}_{id}$ uses the face recognition model *ArcFace*. It can be observed that during the optimization process, for the $\mathcal{L}_{id}$ loss, we can expect minima in the loss surface where the identity is maximized. This point on which the minima is attained in latent vector space, captures the identity. Figure 7 shows that in cross-domain image generation such as real-image-to-caricature, our method can preserve the identity of the real input image and transfer it to the caricature output image. It is worth noting that our method finds a better solution than the current state-of-the-art method *StyleCariGAN*.

4.2.4 Disentanglement Analysis

According to the definition of disentanglement of $\mathcal{W}^+$ the change in some direction in $\mathcal{W}^+$ results in particular characteristics in the image domain. In Fig. 8, the first column represents the input images and the second column represents the generated caricature by optimizing all 18 latents obtained from the *pSp* encoder. It is observed that optimizing all of the 18 latents obtain the highest visual fidelity after generation. For the last column, 1^{st} image represents generated output by only 1–4 latents and the rest with random Gaussian initialization. Similarly, 2^{nd}, 3^{rd}, 4^{th} images optimize $4-8$, $8-12$, $12-18$ latents. It can be seen that initial layers generally store person-specific information such as facial structures, etc. The last 2 images of this column represent the effect of optimizing the last 8 latents divided into 1:1 ratio and keeping 10 previous latents as similar as that of column 2. These results suggest that the last latent set only contains texture-related information. This observation supports that optimizing all 18 latents provide better results.

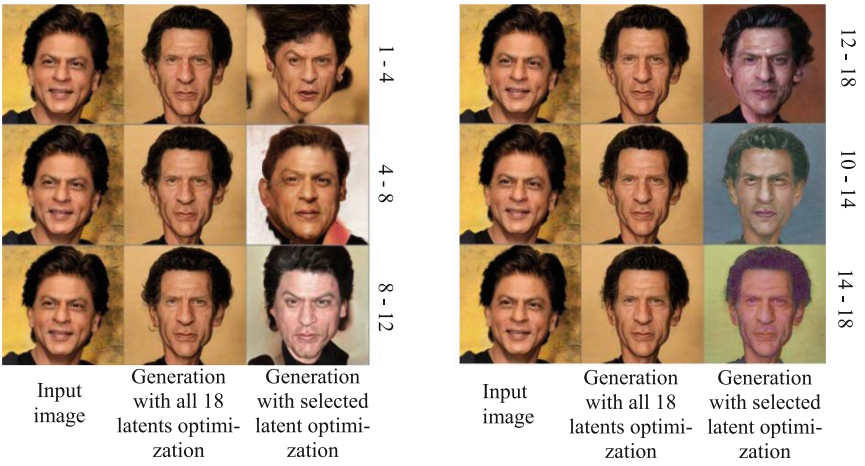


Fig. 8. The effect of different latent optimization is illustrated. The initial latents are mainly responsible for structural consistency while the final latents are responsible for texture generation.

4.3 Qualitative Analysis on Different Image Style Transfer Tasks

In this section, we compare the performance of the proposed method with the existing method qualitatively and quantitatively. This section is divided into three parts:

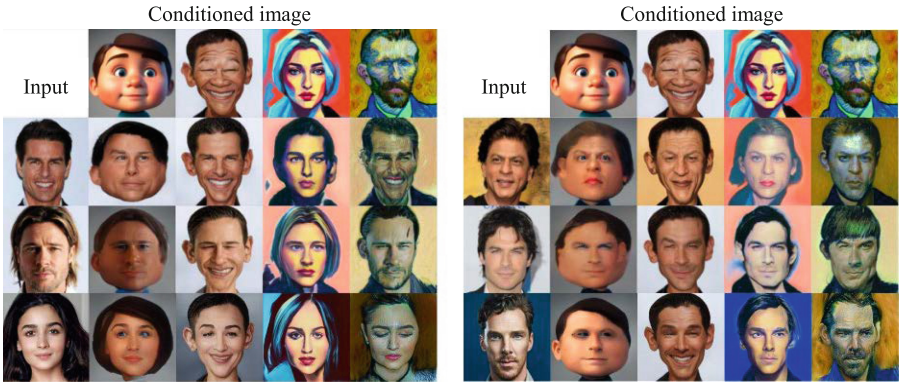


Fig. 9. Samples of style transfer. 1^{st} column and 1^{st} row for each block represents ground truth image and style image. Rest of the images represents output of the proposed model for style transfer.

4.3.1 Experiments on Style Transfer

Style transfer refers to image style editing where the candidate image is manipulated by the condition image's texture, color, features, etc. The proposed method uses novel losses such as *perceptual-loss* and *style-loss* for transferring the style of the condition image to the candidate image while improving visual fidelity. In addition to that, *ID_loss* ensures the identity preservation of the candidate image after generation. The weighted combination of these three losses has been shown to perform wonderfully and produce satisfactory results for image style transfer. Figure 9 shows the robustness of the proposed model for transferring different styles (Pixar, caricature, artistic painting, Gogh) on multiple candidate images. This validates the robustness of the model for style transfer tasks.

Qualitative Comparison with SOTA: Transferring a specific art style (e.g. caricature, cartoon, artistic painting, Gogh, etc.) while generating portraits is a well-known problem. The existing methods often suffer from unwanted identity mixing from the conditioned-style image domain to the generated output. Figure 10 presents the superiority of our proposed model. It can be observed that for style transfer, existing state-of-the art models either fail to learn the style [21] of the conditional art image or inefficiently transfer the shape exaggeration [9] from the condition image domain which results in poor identity regeneration. Although recent advancements in conditional generation have succeeded in copying the conditional art style [11], they often copy the shape of the conditional image and this influences the generated image. As a result, it fails to preserve the identity of the subject. It can be observed that the proposed latent space optimization in

$\mathcal{W}^+$ space for style transfer successfully captures the style and shape of the conditioned image, curates them and faithfully generates images where the identity of the subject is also preserved. At this end, while optimizing, the model finds the optimal latent directions in $\mathcal{W}^+$ that are responsible for shape and style transfer and edits them according to the conditioned style image while preserving the latents that are responsible for identity generation.

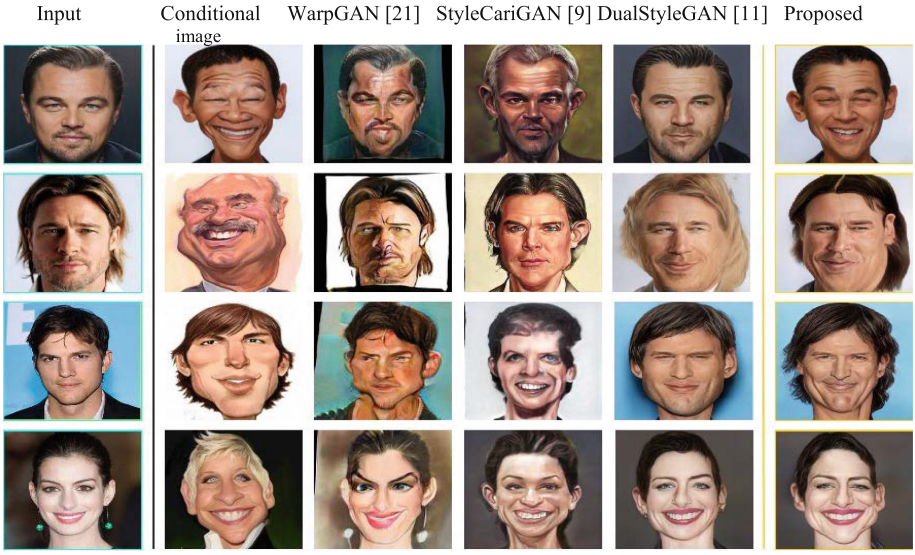


Fig. 10. Qualitative results of the proposed and the existing methods. The proposed method transfers the style and shape of the conditioned images more effectively compared to the existing methods.

4.3.2 Experiments on Text Guided Image Style Transfer

Transferring style of images with textual guidance is particularly challenging due to (i) text-to-image cross-modality distribution difference between text and image and (ii) feature localization for style editing. To solve this non-trivial problem, conditioned text and the generated image are passed through the CLIP text encoder and image encoder respectively and optimization is done with $\mathcal{L}_{CLIP}$ loss. Figure 11 represents the faithfulness of the generation guided by text. It can be observed that the proposed model can localize text-specific latent in $\mathcal{W}^+$ space and perform style transfer tasks in accordance with the condition text. It is worth noting that, while transferring the style, the identity of the subject after generation does not get altered, and ensures reliable generation.

Qualitative Comparison with SOTA: We compare the text guided generation with the existing state-of-the art model. It can be observed from Fig. 12 that although the proposed method has similar or better performance than the existing model [8] for in-domain text-to-image style transfer i.e. real image to real image, for out-of-domain image

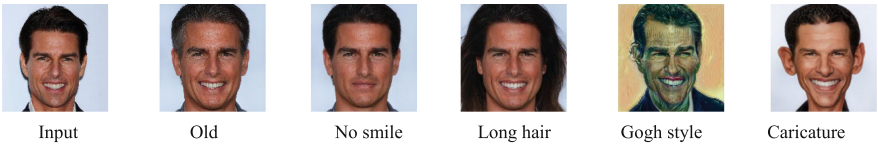


Fig. 11. The effect of textual guidance in image generation of the proposed model. It can be observed that the proposed model can faithfully generate output according to the textual guidance for local style editing (old, no-smile, long hair) as well as global style editing (Gogh style, caricature).

styling, e.g. real image to art style image with text it performs significantly better than the existing model. The reason being the proposed architecture and optimization method can successfully find the latent direction that are responsible for cross-domain image manipulation while keeping the identity of the input image intact.

Fig. 12. Sample results of image generation with textual guidance. It can be observed that for text guided style editing, even though the proposed method performs similar or better than the existing method for local textual guidance, it outperforms the existing methods for global textual editing.

4.4 Quantitative Analysis and Comparison with Diffusion Model

To evaluate the performance of the model quantitatively we use the FFHQ dataset as the source of candidate images. The styled images of the WebCariA dataset are used as conditioned images. Here, at first, we measure the ID metric based on the *ArcFace* to validate the faithful identity regeneration. It essentially compares the identity-specific features of the intermediate layers of the *ArcFace* encoder. A high Id score represents the success of the method for reconstructing the identity of the generated image. Next,

SSIM [34], and PSNR are also calculated to validate the structural consistency of the proposed model. Finally to check the textual generation control CLIP score is measured. Table 3 shows the comparative results. It can be observed that in terms of ID metric, our method achieves the best results signifying the preservation of identity for cross-domain style transfer. Additionally, the proposed model performs better in terms of SSIM and PSNR indicating the high structural similarity and high peak signal-to-noise ratio while generation. However, for CLIP score, Patashnik et al. [8] perform slightly better than the proposed model. We believe that the explicit latent mapper used in [8] is responsible for this slight improvement in CLIP score. However, the exclusion of this mapper in our model makes it faster for generation.

Table 3. Quantitative evaluation of the proposed and existing methods on the FFHQ dataset

		Image to Image			Text to Image
Method	Type	Id score ↑	SSIM ↑	PSNR ↑	CLIP score ↑
WarpGAN [21]	GAN	0.784	0.587	12.44	-
StyleCariGAN [9]	GAN	0.770	0.664	15.40	-
StyleCLIP [8]	GAN	0.959	0.721	17.91	**0.942**
DualStyleGAN [11]	GAN	0.948	0.704	18.03	0.881
Styleres [25]	GAN	0.927	0.710	16.91	0.916
CLIPInverter [26]	GAN	0.901	0.718	16.70	0.910
Textual inversion [28]	Diffusion	0.844	0.680	16.02	0.90
Key locked [29]	Diffusion	0.760	0.651	15.47	0.910
InST [27]	Diffusion	0.820	0.700	17.73	0.860
Proposed	GAN	**0.961**	**0.729**	**18.07**	0.918

Diffusion Models: To compare with state-of-the-art diffusion methods [27–29] we randomly selected 50k images from the FFHQ dataset and trained the model. As the methods [28, 29] take images in-the-wild in consideration for inference, for faithfulness of the comparison, we choose internet-available images (also shown in their main paper) and show the qualitative comparison. Gal et al. [28] encoded in the latent space of a pretrained CLIP. While this method can generate the overall structure, it struggles with specific structural elements such as the face, hairstyle, etc. In contrast, encoding in the StyleGAN latent space in our proposed method not only produces samples of higher perceptual quality but also preserves the subject's identity effectively. For image-to-image style transfer (I2I) tasks, diffusion-based generation methods [27, 29] primarily encode specific artistic styles using textual descriptions as embeddings, guiding the generation process during denoising. However, a limitation of this approach is that the perceptual guidance for a particular style may not be as effective as demonstrated in our proposed method, as illustrated in Fig. 13.

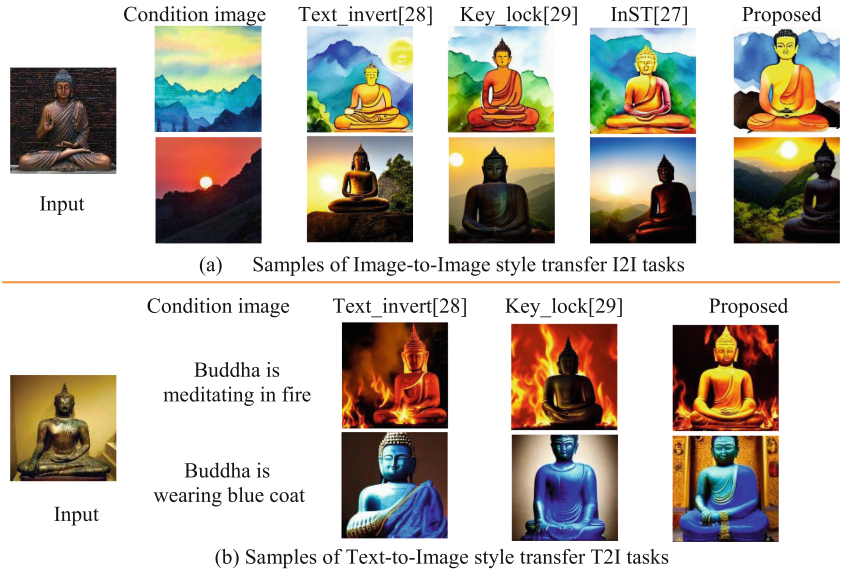


Fig. 13. For I2I (a), existing models do not reconstruct facial attributes faithfully while proposed method can transfer both style and facial information. For T2I (b), the proposed method successfully edited the text specific attribute while existing methods failed to do so.

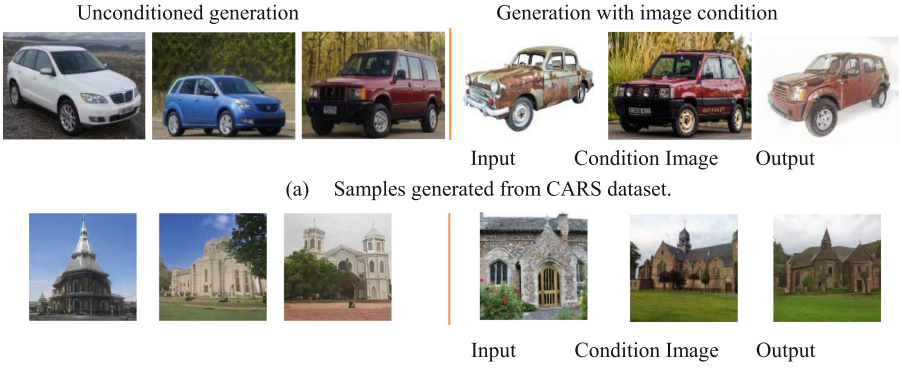


Fig. 14. Unconditional and text conditioned sample results of image generation for CARS (a) and LSUN Church (b) dataset.

4.5 Qualitative Results on Non-Face Datasets

As face data has minute features for person identification, training with the FFHQ dataset proves the robustness of the model. Other than that, for ablation, other dataset like CARS [35] which has 197 classes and a total of 16185 images and CHURCH [36] which has 126227 images of size 64 × 64 of church are also taken in consideration, whose generated sample images are also provided. It can be seen from the results in Fig. 14

that the proposed method can work on non-face images. Thus, one can conclude that the proposed method is application independent and dataset independent.

5 Conclusion and Future Work

In this work, we have proposed a new idea of generating from low dimensional latent space and optimizing in that latent space for different image style transfer tasks with task specific loss functions. Unlike the existing models which are not effective in handling cross-domain distribution, the proposed method is effective in handling cross-domain distribution. To achieve an efficient, effective and generic system for image style transfer, the proposed work derives optimal perceptual loss. Experimental results on several image style transfer, morphing and reconstruction using two standard datasets show that the proposed model is effective, efficient and robust. Since the scope of the work is limited to proposing optimal perceptual loss to handle the low dimensional latent space effectively for image style transfer, our future work is to propose a new generic model that can address other challenges of image style transfer.

References

1. Zhao, J., Zhang, H (2022). Thin-plate spline motion model for image animation. In: Proc. CVPR. pp. 3657–3666
2. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y (2014). Generative adversarial nets. Advances in neural information processing systems. 27
3. Hinton, G.E., Salakhutdinov, R.R (2006). Reducing the dimensionality of data with neural networks. Science. 313, 504–507
4. Kingma, D.P., Welling, M (2013). Auto-encoding variational bayes. arXiv preprint arXiv: 1312.6114
5. Huang, X., Belongie, S (2017). Arbitrary style transfer in real-time with adaptive instance normalization. In Proc. ICCV. pp. 1501–1510
6. Johnson, J., Alahi, A., Fei-Fei, L (2016). Perceptual losses for real-time style transfer and super-resolution. In Proc. ECCV. pp. 694–711
7. Radford, A. et al. (2021). Learning transferable visual models from natural language supervision. In Proc. PMLR, pp. 8748–8763
8. Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., Lischinski, D (2021). Styleclip: Text-driven manipulation of stylegan imagery. In Proc. ICCV. pp. 2085–2094
9. Jang, W., Ju, G., Jung, Y., Yang, J., Tong, X., Lee, S (2021). StyleCariGAN: caricature generation via StyleGAN feature map modulation. ACM Transactions on Graphics (TOG). 40, 1–16
10. Yang, T et al. (2022). Beyond a Video Frame Interpolator: A Space Decoupled Learning Approach to Continuous Image Transition. In Proc. ECCV. pp. 738–755
11. Yang, S., Jiang, L., Liu, Z., Loy, C.C (2022). Pastiche master: Exemplar-based high-resolution portrait style transfer. In Proc. CVPR. pp. 7693–7702
12. Deng, Y., Tang, F., Dong, W., Ma, C., Pan, X., Wang, L., Xu, C (2022). Stytr2: Image style transfer with transformers. In Proc. CVPR. pp. 11326–11336
13. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B (2022). High-resolution image synthesis with latent diffusion models. In Proc. CVPR. pp. 10684–10695

14. Ulyanov, D., Vedaldi, A., Lempitsky, V (2018). Deep image prior. In Proc. CVPR. pp. 9446–9454
15. Luo, W., Yang, S., Zhang, X., Zhang, W (2023). SIEDOB: Semantic Image Editing by Disentangling Object and Background. In Proc. CVPR. pp. 1868–1878
16. Karras, T., Aila, T., Laine, S., Lehtinen, J (2017). Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196
17. Karras, T et al. (2021), Alias-free generative adversarial networks. Advances in neural information processing systems. 34, 852–863
18. Wang, X., Xie, L., Dong, C., Shan, Y (2021). Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In Proc. ICCV. pp. 1905–1914
19. Choi, Y., Uh, Y., Yoo, J., Ha, J.-W (2020). Stargan v2: Diverse image synthesis for multiple domains. In Proc. CVPR. pp. 8188–8197
20. Chen, Y., Liu, S., Wang, X (2021). Learning continuous image representation with local implicit image function. In Proc. CVPR. pp. 8628–8638
21. Shi, Y., Deb, D., Jain, A.K (2019). Warpgan: Automatic caricature generation. In Proc. CVPR. pp. 10762–10771
22. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T (2020). Analyzing and improving the image quality of stylegan. In Proc. CVPR. pp. 8110–8119
23. Abdal, R., Zhu, P., Mitra, N.J., Wonka, P (2021). Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. ACM Transactions on Graphics (ToG). 40, 1–21
24. Abdal, R., Zhu, P., Mitra, N.J., Wonka, P (2022). Video2stylegan: Disentangling local and global variations in a video. arXiv preprint arXiv:2205.13996
25. Pehlivan, H., Dalva, Y., Dundar, A (2022). StyleRes: Transforming the Residuals for Real Image Editing with StyleGAN. abs/2212.14359
26. Baykal, A et al. (2023). CLIP-Guided StyleGAN Inversion for Text-Driven Real Image Editing. https://doi.org/10.1145/3610287
27. Gal, R et al. (2022). An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. Presented at https://doi.org/10.48550/arXiv.2208.01618.
28. Zhang, Y et al. (2023). Inversion-Based Creativity Transfer with Diffusion Models, In Proc. CVPR, 10146–10156
29. Richardson, E et al. (2021). Encoding in style: a stylegan encoder for image-to-image translation. In Proc. CVPR. pp. 2287–2296
30. Karras, T., Laine, S., Aila, T (2019). A style-based generator architecture for generative adversarial networks. In Proc. CVPR. pp. 4401–4410
31. Deng, J., Guo, J., Xue, N., Zafeiriou, S (2019). Arcface: Additive angular margin loss for deep face recognition. In Proc. CVPR. pp. 4690–4699
32. Kingma, D.P., Ba, J (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
33. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P (2004). Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing. 600–612.
34. Krause, J., Deng, J., Stark, M., & Fei-Fei, L (2023). Collecting a large-scale dataset of fine-grained cars. https://ai.stanford.edu/~jkrause/papers/fgvc13.pdf
35. Yu, F et al. (2015). Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint arXiv:1506.03365

SSMT: Few-Shot Traffic Forecasting with Single Source Meta-transfer

Kishor Kumar Bhaumik[1,3], Minha Kim[1], Fahim Faisal Niloy[2], Amin Ahsan Ali[3], and Simon S. Woo[1(✉)]

[1] Sungkyunkwan University, Seoul, South Korea
{kishor25,sunshine01,swoo}@g.skku.edu
[2] University of California, Riverside, Riverside, USA
fnilo001@ucr.edu
[3] Center for Computational and Data Sciences, Independent University, Dhaka, Bangladesh
aminali@iub.edu.bd

Abstract. Traffic forecasting in Intelligent Transportation Systems (ITS) is vital for intelligent traffic prediction. Yet, ITS often relies on data from traffic sensors or vehicle devices, where certain cities might not have all those smart devices or enabling infrastructures. Also, recent studies have employed meta-learning to generalize spatial-temporal traffic networks, utilizing data from multiple cities for effective traffic forecasting for data-scarce target cities. However, collecting data from multiple cities can be costly and time-consuming. To tackle this challenge, we introduce **S**ingle **S**ource **M**eta-**T**ransfer Learning (*SSMT*) which relies only on a single source city for traffic prediction. Our method harnesses this transferred knowledge to enable few-shot traffic forecasting, particularly when the target city possesses limited data. Specifically, we use memory-augmented attention to store the heterogeneous spatial knowledge from the source city and selectively recall them for the data-scarce target city. We extend the idea of sinusoidal positional encoding to establish meta-learning tasks by leveraging diverse temporal traffic patterns from the source city. Moreover, to capture a more generalized representation of the positions we introduced a meta-positional encoding that learns the most optimal representation of the temporal pattern across all the tasks. We experiment on five real-world benchmark datasets to demonstrate that our method outperforms several existing methods in time series traffic prediction. Our code is available at https://github.com/Kishor-Bhaumik/SSMT.

Keywords: Traffic Forecasting · Time Series · Meta Learning · GNN

1 Introduction

Accurate traffic forecasting is crucial for Intelligent Transportation Systems (ITS) to enable a wide range of AI services that rely on real time traffic information, such as food delivery, taxi services, etc. Traditional methods such as

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15311, pp. 46–61, 2025.
https://doi.org/10.1007/978-3-031-78195-7_4

ARIMA [1] and Kalman filter often rely on historical data for univariate time series forecasting. Lately, spatiotemporal traffic forecasting methods [9,27,32] are proposed to integrate temporal and topological sensor relationships for multivariate traffic forecasting using Graph Neural Networks (GNNs) [27]. However, such methods generally rely on abundant training data and fail to generalize for the data-scarce cities.

Recently, transfer learning-based methods such as RegionTrans [24] and Cross TRes [12] focused on grid-based traffic prediction by transferring source knowledge from multiple source cities that have abundant data to the target cities that have limited data. They include large-scale auxiliary data to better match regions that are similar to one another. However, these methods are not directly applicable to graph based traffic prediction because their grid structure differs significantly from the graph structure of the traffic network in target cities. To tackle this problem, few transferable time series forecasting models have been proposed. MetaST [26] uses a global learnable memory, while ST-MetaNet [18] and ST-GFSL [17] employ meta-knowledge from multiple cities for graph-based traffic forecasting. However, the aforementioned approaches tend to overlook the challenges and disadvantages of collecting data from diverse cities. Indeed, gathering sensor data from multiple cities can be prohibitively expensive. Also, possibly, negative transfer [30], when an underdeveloped city or the city significantly differs from the target city is included in the source data, can further undesirably decrease the forecasting performance for the target city. Addressing the aforementioned prior challenges, we introduce Few-Shot Traffic Forecasting with Single Source Meta-Transfer Learning (*SSMT*), a novel framework for transferable spatiotemporal traffic prediction. Our method utilizes data from a single source city learning important spatial footprints and effectively adapting to new temporal patterns. Overall, the main contributions of our work are summarized as follows:

- We present *SSMT*, a new meta-learning based time series traffic forecasting approach that allows single source knowledge transfer from source city to the target city. We carefully construct three meta-learning tasks to help the model quickly adapt to different temporal resolution, each based on a particular periodic pattern (daily, weekly, or monthly).
- Our *SSMT* employs an external memory module to facilitate the effective transfer of topological information from source to target cities in the presence of sensor node discrepancy. This allows the model to gain a deeper understanding of the spatial context, resulting in more accurate predictions for the target city.
- We empirically demonstrate the effectiveness of our proposed *SSMT* framework on five real-world spatiotemporal datasets and show that our method achieves superior performances compared to the existing baselines.

2 Related Work

With the emergence of deep learning and graph neural networks, graph is applied to tackle a variety of urban challenges to explore spatial structural interactions.

Andrea et al. [3] proposed a framework to explain empirical results associated with the use of trainable node embeddings and discuss different architectures and regularization techniques to account for local effects. Bai et al. [2] present STG2Seq, a graph-based model for multi-step citywide passenger demand. Yuan et al. [29] proposed recasting spatio-temporal few-shot learning as pre-training a generative diffusion model, which creates tailored neural networks guided by prompts. This approach allows for adaptability to diverse data distributions and city-specific characteristics. To capture the dynamic aspects of urban traffic flow, Lu et al. [16] propose spatial and semantic neighbors of road segments. Do et al. [5] use IoT sensors on automobiles to assess city air quality and use variational graph autoencoders to predict unknown air pollutants. Nevertheless, these approaches are non-transferable and primarily concentrate on single-city traffic forecasting. Few-shot learning (FSL) has shown promising performance in various domains such as computer vision, natural language processing, and reinforcement learning when dealing with data scarcity. In our context, when data-rich source cities are used to transfer knowledge to data-scarce target cities, this problem is referred to as few-shot traffic forecasting. And, recently cross-city transfer learning models [12,24] have gained significant popularity in this area. MetaST [26] employs a global memory queried by the target region. Moreover, STrans-GAN [31] generates future traffic speed using GANs, and TPB method [15] proposes a traffic pattern bank to store similar patterns from multiple source cities for the downstream fine-tuning task. However, these methods heavily depend on data-rich multiple source cities, and can be cost-prohibitive in practice.

Recently, memory-augmented attention (MAA) models have gained much attention for capturing long-term dependency, particularly anomaly detection task [8]. Park et al.[19] used MAA for anomaly detection in video sequences. These studies explicitly utilize memory augmented attention to enhance model performance by storing data patterns. Inspired by these, in our work, we propose a separate memory module to address non-transferability due to node count mismatches between source and target cities.

3 Methodology

3.1 Preliminaries and Problem Formulation

Traffic networks are represented as $\mathcal{G}_s$ and $\mathcal{G}_t$, with $\mathcal{V}$ as vertices (e.g., traffic sensors) and $\mathcal{E}$ as edges (connectivity). V is the set of vertices where $V \in (v_1, v_2, v_3, ...v_n)$. In our context, we use subscripts s and t to denote the data of the source and target city, respectively. Next, we define the source data as $X_s \in \mathbb{R}^{N_s \times T_s \times C_s}$ and target data as $X_t \in \mathbb{R}^{N_t \times T_t \times C_t}$, where N, T, and C represents the total number of nodes, time window length, and the number of traffic features, respectively. The adjacency matrix $\mathcal{A} \in \mathbb{R}^{N \times N}$ is the spatiotemporal graph of $\mathcal{G}$. $A_{ij} = 1$ indicates that there is an edge between node v_i and v_j, otherwise, $v_{ij} = 0$. In our work, we mainly focus on investigating the transferability of a single feature, namely traffic speed, thereby reducing C to 1. Hence,

we finally consider $X_s \in \mathbb{R}^{N_s \times T_s}$ as the source, and $X_t \in \mathbb{R}^{N_t \times T_t}$ as the target data. Also, as the number of nodes differs between the source and the target city, we denote $N_s \neq N_t$ to indicate this discrepancy in the number of nodes.

In general, the transferable traffic forecasting problem can be divided into two stages: 1) pre-training and 2) fine-tuning. First, in the pre-training stage, we can formulate the forecasting problem by training a mapping function f_{θ_s} on the source data, which predicts future timestamps (T') based on past timestamps (T). We denote the historical spatiotemporal input data and the predicted observations as $\langle X_{\mathcal{G}s}^{(t-T+1)}, X_{\mathcal{G}s}^{(t-T+2)}, \ldots, X_{\mathcal{G}s}^{(t)} \rangle$ and $\langle X_{\mathcal{G}s}^{(t+1)}, X_{\mathcal{G}s}^{(t+2)}, \ldots, X_{\mathcal{G}_s}^{(t+T')} \rangle$, respectively. Thus, the time series forecasting with the input graph $\mathcal{G}_s$ for the source city can be defined as follows:

$$\langle X_{\mathcal{G}s}^{(t-T+1)}, X_{\mathcal{G}s}^{(t-T+2)}, \ldots, X_{\mathcal{G}s}^{(t)} \rangle \xrightarrow{f_{\theta_s}} \langle X_{\mathcal{G}s}^{(t+1)}, X_{\mathcal{G}s}^{(t+2)}, \ldots, X_{\mathcal{G}s}^{(t+T')} \rangle$$

Next, in the fine-tuning stage, the forecasting task is performed by fine-tuning the same mapping function with parameters θ_s, which are initially shared from the pre-trained function. The fine-tuning process aims to improve the model's ability to predict graph signals specifically for the target road network. And, the fine-tuning process can be defined as follows:

$$\langle X_{\mathcal{G}t}^{(t-T+1)}, X_{\mathcal{G}t}^{(t-T+2)}, \ldots, X_{\mathcal{G}t}^{(t)} \rangle \xrightarrow{f_{\theta_t^*;\theta_s}} \langle X_{\mathcal{G}t}^{(t+1)}, X_{\mathcal{G}t}^{(t+2)}, \ldots, X_{\mathcal{G}t}^{(t+T')} \rangle$$

In particular, the notation $f_{\theta_t^*;\theta_s}$ represents the adjusted function parameters from θ_s to fit the target domain. It basically signifies the updated or adapted set of parameters that have been modified from the source model to better align with the characteristics and requirements of the target city.

3.2 Meta-Learning Framework

In our model, θ represents the encoded spatiotemporal model parameter for the traffic prediction. To adapt the source model parameters effectively to a target city, we adopt the approach proposed in model-agnostic meta-learning (MAML) [7] In MAML framework, the dataset is divided into multiple tasks and each task is divided further into support and query sets. Then, the support set is used for model adaptation during meta-training, while the query set evaluates the model's performance after this adaptation. We use MAML to initialize θ with multiple tasks from the source city as θ_s, such that it can minimize the average generalization loss across all source tasks. When adapting to a new task $\mathcal{T}^i$, the model's parameters θ_s are updated to θ_i'. In MAML, the inner loop is for training the model's parameters for a specific task, while the outer loop is to adjust the model's initial parameters to enhance the learning performance across a range of tasks. The inner loop optimization of MAML can be expressed as follows:

$$\theta_i' = \theta_s - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}^i} \left(f_{\theta_s} \right) \tag{1}$$

And, practically, we can perform multiple steps of gradient descent to update the initialization θ_s to θ_i'. For each task $\mathcal{T}^i$, the training process is iterated on batches of tasks sampled from source tasks $S(\mathcal{T})$. More intuitively, the meta-learning objective is defined as follows:

$$\min_{\theta_s} \sum_{\mathcal{T}^i \sim S(\mathcal{T})} \mathcal{L}_{\mathcal{T}^i}\left(f_{\theta_i'}\right) = \sum_{\mathcal{T}^i \sim S(\mathcal{T})} \mathcal{L}_{\mathcal{T}^i}\left(f_{\theta_s - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}^i}(f_{\theta_s})}\right) \tag{2}$$

Finally, the outer loop optimization across all the tasks can be defined as:

$$\theta = \theta - \beta \nabla_{\theta_i'} \sum_{\mathcal{T}^i \sim S(\mathcal{T})} \mathcal{L}_{\mathcal{T}^i}\left(f_{\theta_i'}\right) \tag{3}$$

Defining β as the meta-step size, we assume that θ will yield superior generalization performance to a target city, as it provides an initialization that performs well across multiple tasks.

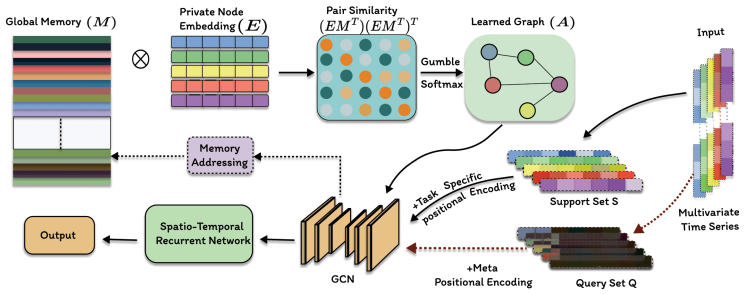


Fig. 1. The overall architecture of our proposed *SSMT* consists of several components. Firstly, the discrete graph A is learned by computing the pair similarity between M and E. The memory component then addresses the GCN module's output, which is subsequently delivered to the spatiotemporal recurrent block. Finally, the output of the spatiotemporal block is utilized to compute the loss.

3.3 Positional-Encoding Driven Task Partitioning

Let us define each batch of input data as D_j where $j \in [1, 2, 3, .., BatchSize]$. We first represent this split of a single batch in three distinct sets to present daily, weekly and monthly periodic patterns as follows:

$$\begin{aligned} D_j &= \mathcal{B}_j^{(1)} \cup \mathcal{B}_j^{(2)} \cup \mathcal{B}_j^{(3)} \\ \mathcal{B}_j^i &= \{(x_1^{(i)}, y_1^{(i)}), (x_2^{(i)}, y_2^{(i)}), \ldots, (x_{n^{(i)}}^{(i)}, y_{n^{(i)}}^{(i)})\} \end{aligned} \tag{4}$$

where $\mathcal{B}^i$ represents the i-th set of the batch and $i \in [1, 2, 3]$. $\mathcal{B}^i$ contains multiple data points and each data point $(x_j^{(i)}, y_j^{(i)})$ consists of input features $x_j^{(i)}$ and corresponding labels $y_j^{(i)}$ for the i-th set. $n^{(i)}$ is the number of data points in the i-th set of the batch. Inspired by the widely adopted concept of relative positional encoding in transformer-based attention mechanisms [13,22], we derive positional encoding as follows:

$$\begin{aligned} PE_{\text{pos},2k} &= \sin\left(\frac{2\pi \cdot \text{pos}}{24 \cdot \text{samples\_per\_hour} \cdot V}\right) \\ PE_{\text{pos},2k+1} &= \cos\left(\frac{2\pi \cdot \text{pos}}{24 \cdot \text{samples\_per\_hour} \cdot V}\right), \end{aligned} \tag{5}$$

where the value of V is influenced by the periodic patterns observed in the source city with k representing specific positions (either odd or even). For instance, to capture the daily pattern, we set $V = 1$. Similarly, for the weekly pattern, V is set to 7, and for the monthly pattern, V is set to 30. We next add these positional encodings to each of $\mathcal{B}^i$ resulting in three unique sets of batches becoming three distinct tasks. We then equally divide each task into support and query sets for the inner loop optimization.

However, traditional sinusoidal positional encodings might not be optimal to use for all tasks. To address this challenge, we further introduce meta-positional encoding that is learned throughout the outer loop optimization process, denoted as $\eta \in \mathbb{R}^{N \times T}$. We define the meta-positional encoding as follows:

$$\eta = \alpha \otimes E; \text{ where, } \alpha \in \mathbb{R}^{1 \times T} \text{ and } E \in \mathbb{R}^{N \times T}, \tag{6}$$

where $\otimes$ represents the matrix multiplication, α is the scaling parameter and E is the embedding vector. And, the scaling parameter, α, acts as a weight that determines the importance of the positional information. The meta-positional encoding η remains static in the inner loop and is only updated during the outer loop optimization. By learning a positional encoding in the outer loop, our model can potentially capture a more generalized representation of position that works across different time scales (daily, weekly, and monthly). Furthermore, it can help the meta-model start with a more suitable positional representation for a variety of tasks before inner loop adaptation.

3.4 Memory-based Spatial Knowledge Transfer

While traditional methods [11,28] generate graphs from a similarity matrix, our transfer-learning context for inter-city traffic forecasting presents the following unique challenges: When transferring knowledge from a source city to a target city, node embeddings become a problem due to discrepancies in the number of sensors between the two cities. For example, let us consider a scenario where the model is pre-trained using the source data, resulting in the learned node embedding $E_s \in \mathbb{R}^{N_s \times d}$. However, creating a target city specific graph requires the node embedding $E_t \in \mathbb{R}^{N_t \times d}$ which disables to use of the learnable node embedding from the source city directly.

In order to overcome this non-transferability issue due to the mismatch of node number between source and target cities, we adopt a memory module inspired by [21,23]. And, our proposed module supplements node embeddings that cannot be directly transferred. Let $E_s \in \mathbb{R}^{N_s \times d}$ and $E_t \in \mathbb{R}^{N_t \times d}$ denote the private node embeddings of the source and the target, respectively, which are not shared but trained independently. In particular, we leverage the idea of memory network [8] and learn a global memory that is shared in both the source and target city. Specifically, the memory is defined as $M \in \mathbb{R}^{b \times d}$, where b and d denote the number of memory items and the dimension of each item, respectively. During the pre-training step for the source data, the inter-node similarity matrix $\xi_s \in \mathbb{R}^{N_s \times N_s}$ is defined as follows:

$$\xi_s = (E_s M^T)(E_s M^T)^T \tag{7}$$

Upon completion of the pre-training using the source data, we use only the learned global memory M in the fine-tuning stage for the target data. Likewise, the inter-node similarity matrix $\xi_t \in \mathbb{R}^{N_t \times N_t}$ for the target city can be formulated as follows:

$$\xi_t = (E_t M^T)(E_t M^T)^T, \tag{8}$$

where the global memory M enables us to transfer the extracted knowledge on the road topological structure of the source data, which could have been potentially lost due to the disparity in the number of nodes between the cities.

To effectively ensure a decent level of sparsity of the graph structure, we apply the Gumbel softmax trick to retrieve the final sparse adjacency matrix $A \in \mathbb{R}^{N \times N}$ for both source and target, where σ and τ are the activation function and the temperature variable, respectively. This process can be expressed as follows:

$$\begin{aligned} A &= \sigma((log(\xi_{ij}/(1-\xi_{ij}) + (n^1_{ij} - n^2_{ij}))/\tau) \\ &s.t.\ n^1_{ij}, n^2_{ij} \sim Gumbel(0,1) \end{aligned} \tag{9}$$

Equation 9 implements the Gumbel Softmax algorithm [10] for our task, where $A_{i,j} = 1$ with the probability $\xi_{i,j}$ and 0 with the remaining probability. Gumbel Softmax maintains the same probability distribution as the normal Softmax, ensuring statistical consistency in generating the trainable probability matrix for the graph forecasting network. And, let I_n denote an identity matrix and D represent a diagonal degree matrix satisfying $D_{ii} = \Sigma_j A_{ij}$, then the specific operation of graph convolutional network (GCN) can be expressed as follows:

$$O = GCN_{\star A}(X) = W(I_n + D^{-\frac{1}{2}} A D^{-,2})X, \tag{10}$$

where GCN is parameterized by $W \in \mathbb{R}^{T \times d}$, and the output of GCN is denoted as $O \in \mathbb{R}^{N \times d}$.

3.5 Memory Addressing

The memory M is designed to explicitly record the topological pattern during the training. We define the memory that computes attention weights w based on the similarity of the memory items and the query from the GCN output O. We compute each w_j via a softmax operation as a memory addressing scheme by following:

$$w_j = \frac{exp(sim(O_t^{(a)}, M_j))}{\Sigma_{j=1}^{M} exp(sim(O_t^{(a)}, M_j))} \tag{11}$$

$$sim\left(O_t^{(a)}, M_j\right) = \frac{O_t^{(a)} M_j^{\top}}{\|O_t^{(a)}\| \, \|M_j\|} \tag{12}$$

where we denote a as a row index, and derive the memory reading operation by matching $O^{(a)}$ with each memory M_j. We then calculate a scaler w_j that represents the cosine similarity between vector $O^{(a)}$ and memory M_j as shown in Eq. 12. After that, We recover a node embedding vector $P_t^{(a)} \in \mathbb{R}^d$ combining all the memory item as follows:

$$P_t^{(a)} = \Sigma_{j=1}^{M} w_j m_j \tag{13}$$

3.6 Spatial Temporal Recurrent Network

To capture the spatiotemporal pattern in the traffic data, we use the Spatiotemporal Recurrent Graph Convolution Module (STRGC) proposed in [14]. This module integrates node embedding P_t with a Gated Recurrent Unit (GRU) network and processes the input series $X_{t-T+1:t}$ and matrix P_t to ultimately produce the predicted future traffic data, $X_{t+1:t+T'}$. This prediction process is precisely defined as follows:

$$\begin{aligned} z_t &= \sigma(P_t([X_{0:t} \,||\, h_{t-1}])) \\ r_t &= \sigma(P_t([X_{0:t} \,||\, h_{t-1}])) \\ c_t &= tanh(P_t([X_{0:t} \,||\, (r_t \odot h_{t-1})])) \\ h_t &= z_t \odot h_{t-1} + (1 - z_t) \odot c_t. \end{aligned} \tag{14}$$

In particular, Figure 1 presents the overall architecture of our method, where we use mean absolute error (MAE) as our objective criterion which is formulated as follows:

$$\mathcal{L}_{MAE} = \frac{1}{T_y} \sum_{i=1}^{T_y} \|Y_{[:,\tau]} - \widehat{Y}_{[:,\tau]}\|, \tag{15}$$

To further enhance our model's discriminating power for diverse scenarios on different roads over time, we regulate the memory parameters with two constraints [20] including a contrastive loss $\mathcal{L}_{\text{separate}}$ and a consistency loss $\mathcal{L}_{\text{compact}}$,

as follows:

$$\mathcal{L}_{\text{separate}} = \sum_{t}^{T} \sum_{a}^{N} \left[\left\| O_t^{(a)} - M_p \right\|_2 - \left\| O_t^{(a)} - M_n \right\|_2 + \lambda \right]_+$$

$$\mathcal{L}_{\text{compact}} = \sum_{t}^{T} \sum_{a}^{N} \left\| O_t^{(a)} - M_p \right\|_2$$

where T indicates the total number of sequences (i.e., samples) in the training set, and p, n signify the top two memory item indices determined by ranking w_j in Eq. 11 given localized query $O_t^{(a)}$. And, we regard $O_t^{(a)}$ as the anchor, its most comparable prototype M_p as the positive sample, and the second similar prototype M_n as the negative sample by applying these two constraints, where λ signifies the margin between the positive and negative pairings. Here, the idea is to keep memory items as compact as possible, at the same time, as dissimilar as possible through contrastive loss. These two competing objectives limits the memory's ability to directly discriminate between diverse spatiotemporal patterns at the node level. In practice, we found that including them within the objective criteria (i.e., MAE) promotes training convergence (with balancing factors C_1, C_2 and C_3):

$$\mathcal{L}_{total} = C_1 \mathcal{L}_{MAE} + C_2 \mathcal{L}_{\text{separate}} + C_3 \mathcal{L}_{\text{compact}} \tag{16}$$

It should be noted that during source pre-training, we use only the MAE loss Eq. 15 to update the model parameters. And, for fine-tuning in the target dataset, we employ Eq. 16 as the loss function. This is because we expect the memory to include a variety of topological patterns from the source city. Since data is limited in the target city, the memory should only retain certain patterns related to its own topological structure in order to prevent negative transfer.

3.7 Pre-training and Fine-tuning Process

As mentioned, the training process for *SSMT* consists of two primary stages: 1) pre-training with the data from the source city and 2) subsequent fine-tuning using the data from the target city. In the pre-training stage, data from each task is divided into support and query sets to facilitate MAML training. During the inner loop optimization, the model parameters are updated using the support set for each task. Importantly, the learnable meta-positional encoding used for the query set remains static during the inner loop optimization. In the subsequent outer loop optimization phase, the meta-positional encoding is updated based on parameters learned from all the tasks. This enables our model to learn task-universal representations across all temporal patterns. When the model undergoes the fine-tuning stage with the target data, we use the pre-trained weights as the initial parameters, except for the node embedding for target data which is initialized separately. The entire learning process is presented in Algorithm 1

Algorithm 1 :Pre-training and fine-tuning process of *SSMT*

Input: Source input data$(\mathcal{X}^s, \mathcal{Y}^s)$ and target input data $\left(\mathcal{X}^t, \mathcal{Y}^t\right)$
Output: Spatial temporal prediction in the target city

/* ——— Source pre-training ——— */
randomly initialize θ
$\theta_s \longleftarrow \theta$
while not done **do**
  // sample batch from the source dataset
  $D_j^s \longleftarrow SampleBatch(\mathcal{X}^s, \mathcal{Y}^s)$
  for all D_j^s **do**
    // sample task from a single batch
    $\mathcal{T}^i \longleftarrow D_j^s$ by Eq. (4)
    for all $\mathcal{T}^i$ **do**
      // sample support set
      $\left(\mathcal{X}_{sup}^s, \mathcal{Y}_{sup}^s\right) \longleftarrow \mathcal{T}^i$
      calculate $\alpha \nabla_\theta \mathcal{L}_{\mathcal{T}^i}\left(\mathcal{Y}_{supp}^s, \hat{\mathcal{Y}}_{supp}^s\right)$ by Eq. 15
      $\theta_s \longleftarrow \theta_s - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}^i}\left(\mathcal{Y}_{supp}^s, \hat{\mathcal{Y}}_{supp}^s\right)$
    end for
    sample query set $\left(\mathcal{X}_{qu}^s, \mathcal{Y}_{qu}^s\right) \longleftarrow \mathcal{T}^i$
    compute $\mathcal{L}_{\mathcal{T}^i}\left(\mathcal{Y}_{qu}^s, \hat{\mathcal{Y}}_{qu}^s\right)$ by Eq. 15
    $\theta \longleftarrow \beta \nabla_{\theta_s} \sum_{\mathcal{T}^i \sim D_j^s} \mathcal{L}_{\mathcal{T}_i}\left(\mathcal{Y}_{qu}^s, \hat{\mathcal{Y}}_{qu}^s\right)$
  end for
end while
/* ——— Target fine-tuning ——— */
while not done **do**
  // sample batch from the target dataset
  $D_j^t \longleftarrow SampleBatch(\mathcal{X}^t, \mathcal{Y}^t)$
  for all batches **do**
    $\left(\mathcal{X}_i^t, \mathcal{Y}_i^t\right) \longleftarrow D_j^t$
    $\hat{\mathcal{Y}}_i^t \longleftarrow f_\theta\left(\mathcal{X}_i^t\right)$
    calculate $\mathcal{L}_{total}\left(\mathcal{Y}_i^t, \hat{\mathcal{Y}}_i^t\right)$ by Eq (16)
    calculate $\nabla_\theta \mathcal{L}_{total}\left(\mathcal{Y}_i^t, \hat{\mathcal{Y}}_i^t\right)$ by Eq (16)
    $\theta \longleftarrow AdamOptimizer\left(\nabla_\theta \mathcal{L}_{total}\left(\mathcal{Y}_i^t, \hat{\mathcal{Y}}_i^t\right)\right)$
  end for
end while
return θ

4 Experiments

4.1 Datasets

We perform our experiments on five publicly available benchmark traffic datasets: METR-LA, PEMS-BAY, PEMSD4, Didi-Chengdu and Didi-Shenzhen [4,14]. These datasets contain months of traffic speed data. METR-LA and PEMS-BAY are collected every five minutes, while Didi-Chengdu and Didi-Shenzhen are collected every ten minutes. In our experiment, we use the PEMS-BAY dataset as the source dataset for the target METR-LA and PEMSD4 datasets, and we use the Didi-Shenzhen dataset as the source dataset for the target Didi-Chengdu dataset since PEMS-BAY and Didi-Shenzhen have significantly more traffic data compared to their respective target cities. Our primary motivation in this paper is to explore single-source transfer learning. Therefore, we strategically selected the largest datasets from source cities to ensure robustness. We assume that target cities will have smaller datasets, as they may have recently deployed sensors. Additionally, we restricted our experiments to cities within the same country due to the high security and privacy concerns associated with traffic data.

Table 1. Dataset description.

Dataset	Number of sensors (Nodes)	Time Steps	Mean
PEMS-BAY	325	52,116	61.77
METRA-LA	207	34,272	58.27
PEMSD4	170	17,856	35.38
Didi-Shenzhen	627	17,280	31.01
Didi-Chengdu	524	17,280	29.02

4.2 Few-Shot Setting

We use the similar few-shot traffic forecasting setting proposed in [17]. We divide the data of the cities into source, target, and test sets, where the source data consists of data from a single city, while the target and test data consist of data from the target city. For example, if PEMSD4 is the target city, we use the full PEMS-BAY dataset as the source data, 1 week of PEMSD4 data as the target data, and the remaining PEMSD4 data as the test data. We pre-train and learn multiple tasks on the source data and fine-tune the model on the target data. Finally, we evaluate our framework on the test data. We implemented our model using Pytorch training with a single NVIDIA RTX A5000 GPU with 24 GB memory. The training setup involved a batch size of 64 and a maximum of 100 epochs. We set the learning rates for the inner and outer loops at 0.01 and 0.001, respectively. And the memory module is comprised of 20 units, each with

a dimension of 64. We empirically chose the values of C_1, C_2, and C_3 as 0.5, 0.2, and 0.3, respectively for Eq. 16.

Table 2. The overall performance of *SSMT* vs. baseline methods, where our method achieves the best performance.

Methods	**PEMS-BAY to METRA-LA**						**PEMS-BAY to PEMSD4**					
	MAE			RMSE			MAE			RMSE		
	5 min	15 min	30 min	5 min	15 min	30 min	5 min	15 min	30 min	5 min	15 min	30 min
DCRNN	3.05	3.45	4.38	4.78	5.97	7.55	19.55	20.56	21.15	29.19	30.24	32.33
GWN	3.12	3.59	4.27	4.87	5.94	7.66	19.37	20.58	21.12	29.26	30.15	32.24
AdaRNN	2.95	3.39	4.13	4.66	5.86	7.25	19.45	20.44	20.92	29.12	30.05	32.02
ST-GFSL	2.81	3.21	4.12	4.30	5.82	7.38	19.02	20.32	20.88	28.85	29.93	31.73
TPB	2.75	3.11	3.88	4.25	5.75	6.97	19.08	20.17	20.98	28.81	29.95	31.69
SSMT (ours)	**2.66**	**3.01**	**3.80**	**4.14**	**5.60**	**6.82**	**18.58**	**19.45**	**20.34**	**28.31**	**29.35**	**31.02**

Methods	**Didi-Shenzhen to Didi-Chengdu**					
	MAE			RMSE		
	10 min	30 min	60 min	10 min	30 min	60 min
DCRNN	2.68	3.19	3.41	3.55	4.05	7.62
GWN	2.79	3.05	3.49	3.41	4.12	4.77
AdaRNN	2.64	2.91	3.35	3.33	3.97	4.56
ST-GFSL	2.48	2.81	3.28	3.24	3.88	4.42
TPB	2.35	2.88	3.21	3.19	3.87	4.45
SSMT (ours)	**2.23**	**2.71**	**3.16**	**3.10**	**3.83**	**4.39**

4.3 Baselines

To demonstrate the superiority and effectiveness of our proposed method in terms of transferability, we use the following baselines: 1) Diffusion Convolutional Recurrent Neural Network (DCRNN) [14] 2) Graph wavenet for deep spatial-temporal graph modeling (GWN) [25] 3) Adaptive learning and forecasting of time series (Ada-RNN) [6], 4) Spatio-Temporal Graph Few-Shot Learning with Cross-City Knowledge Transfer (ST-GFSL) [17], and 5) Cross-city Few-Shot Traffic Forecasting via Traffic Pattern Bank (TPB) [15]. In particular, we compare our model with both transferable and non-transferable methods. The first two models are traditional non-transferable forecasting baselines, whereas the others are transferable baselines including SOTA methods. In this paper, we use the mean absolute error (MAE) and the root mean squared error (RMSE) to evaluate the prediction performance.

4.4 Results and Analysis

In Table 2, we present a performance comparison of *SSMT* with other SOTA baselines. Our *SSMT* outperforms all other methods following the same evaluation settings in [17]. We can observe that our proposed *SSMT* outperforms

the baselines in both short-term and long-term forecasting. For the METR-LA dataset, our model surpassed the second-best baseline scores, showing enhancements of 2.84% in MAE and 2.44% in RMSE. Similarly, for the PEMSD4 dataset, we observed improvements of 2.90% in MAE and 1.92% in RMSE. Lastly, with the Didi-Chengdu dataset, our gains were 3.14% and 1.50% in MAE and RMSE, respectively. The improved performance demonstrates that the added meta-knowledge from the memory bank and meta-positional encoding indeed enhance the model's forecasting accuracy.

4.5 Ablation Study

We delve deeper into analyzing the effectiveness of different modules and the sensitivity of hyperparameters through ablation study. Table 3 demonstrates the effectiveness of the two modules employed in our method. Specifically, we show the experiment on Didi-Chengdu dataset for 1 hour prediction. It clearly illustrates that both strategies are individually effective for our task. Furthermore, combining the results of both methods leads to even greater improvements in performance.

Table 3. Ablation study for the effectiveness of the memory bank and meta-positional encoding (MPE) used in our method.

Memory	MPE	Error Score (MAE)
×	×	4.12%
×	✓	3.74%
✓	×	3.62%
✓	✓	**3.16%**

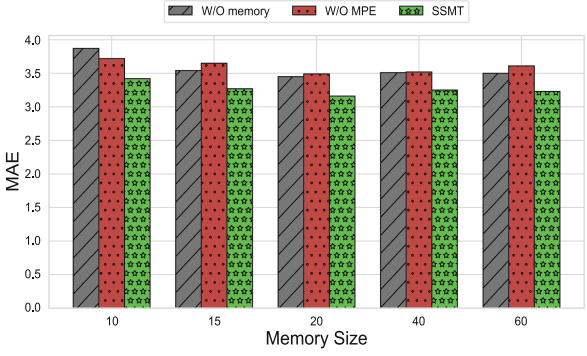


Fig. 2. MAE performance with different memory sizes on PeMSD4 dataset.

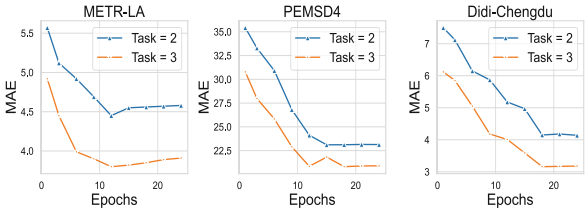


Fig. 3. The performance of *SSMT* vs. the number of tasks.

To further investigate the effect of the memory module of *SSMT*, we performed experiments with different memory size settings and presented the corresponding MAE scores for the Didi-Chengdu dataset. The results summarized in Figure 2 demonstrate that when a sufficiently large memory size is used, our *SSMT* consistently generates reliable and plausible outcomes, indicating that the memory module succeeds in transferring informative spatial knowledge from source to the target.

In Figure 3, we examine the impact of the number of tasks on traffic prediction. Our findings indicate that increasing the number of tasks can lead to significant improvements of our model. And, this suggests that diverse temporal tasks have the potential to capture essential temporal patterns specific to the target city.

5 Conclusion

In this paper, we present *SSMT*, a transferable model tailored for time series traffic forecasting. During the initial pre-training phase, our proposed model is trained using abundant data from a single source city, while the fine-tuning phase involves refining the pre-trained model using limited data from the target city. We present a carefully designed memory mechanism that stores the diverse patterns from the source city and then retrieve only the target-specific patterns from the memory to predict traffic speed accurately. We further propose a meta-positional encoding that consolidates universal patterns from daily, weekly, and monthly positional encodings. Our experimental results demonstrate the superior transferability, outperforming the SOTA traffic forecasting methods.

Acknowledgements. This work was partly supported by Institute for Information & communication Technology Planning & evaluation (IITP) grants funded by the Korean government MSIT: (RS-2022-II221199, RS-2024-00337703, RS-2022-II220688, RS-2019-II190421, RS-2023-00230337, RS-2024-00356293, RS-2022-II221045, and RS-2021-II212068).

References

1. M. S. Ahmed and A. R. Cook. *Analysis of freeway traffic time-series data by using Box-Jenkins techniques*. Number 722. 1979

2. L. Bai, L. Yao, S. Kanhere, X. Wang, Q. Sheng, et al. Stg2seq: Spatial-temporal graph to sequence model for multi-step passenger demand forecasting. arXiv preprint arXiv:1905.10069, 2019
3. A. Cini, I. Marisca, D. Zambon, and C. Alippi. Taming local effects in graph-based spatiotemporal forecasting. *Advances in Neural Information Processing Systems*, 36, 2024
4. Didi-Chuxing. Didi chuxing gaia initiative., 2020. Accessed: 2020-02-14
5. Do, T.H., Tsiligianni, E., Qin, X., Hofman, J., La Manna, V.P., Philips, W., Deligiannis, N.: Graph-deep-learning-based inference of fine-grained air quality from mobile iot sensors. IEEE Internet Things J. **7**(9), 8943–8955 (2020)
6. Y. Du, J. Wang, W. Feng, S. Pan, T. Qin, R. Xu, and C. Wang. Adarnn: Adaptive learning and forecasting of time series. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 402–411, 2021
7. C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017
8. D. Gong, L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh, and A. v. d. Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1705–1714, 2019
9. Guo, S., Lin, Y., Wan, H., Li, X., Cong, G.: Learning dynamics and heterogeneity of spatial-temporal graph data for traffic forecasting. IEEE Trans. Knowl. Data Eng. **34**(11), 5415–5428 (2021)
10. E. Jang, S. Gu, and B. Poole. Categorical reparameterization with gumbel-softmax. arXiv preprint arXiv:1611.01144, 2016
11. Jiang, R., Wang, Z., Yong, J., Jeph, P., Chen, Q., Kobayashi, Y., Song, X., Fukushima, S., Suzumura, T.: Spatio-temporal meta-graph learning for traffic forecasting. In Proceedings of the AAAI Conference on Artificial Intelligence **37**, 8078–8086 (2023)
12. Y. Jin, K. Chen, and Q. Yang. Selective cross-city transfer learning for traffic prediction via source city region re-weighting. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 731–741, 2022
13. A. J. R. Kumar and B. Bhanu. Micro-expression classification based on landmark relations with graph attention convolutional network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1511–1520, 2021
14. Y. Li, R. Yu, C. Shahabi, and Y. Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. arXiv preprint arXiv:1707.01926, 2017
15. Z. Liu, G. Zheng, and Y. Yu. Cross-city few-shot traffic forecasting via traffic pattern bank. arXiv preprint arXiv:2308.09727, 2023
16. B. Lu, X. Gan, H. Jin, L. Fu, and H. Zhang. Spatiotemporal adaptive gated graph convolution network for urban traffic flow forecasting. In *Proceedings of the 29th ACM International conference on information & knowledge management*, pages 1025–1034, 2020
17. B. Lu, X. Gan, W. Zhang, H. Yao, L. Fu, and X. Wang. Spatio-temporal graph few-shot learning with cross-city knowledge transfer. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1162–1172, 2022

18. Z. Pan, Y. Liang, W. Wang, Y. Yu, Y. Zheng, and J. Zhang. Urban traffic prediction from spatio-temporal data using deep meta learning. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1720–1730, 2019
19. H. Park, J. Noh, and B. Ham. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14372–14381, 2020
20. H. Park, J. Noh, and B. Ham. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14372–14381, 2020
21. A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, pages 1842–1850. PMLR, 2016
22. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017
23. O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016
24. L. Wang, X. Geng, X. Ma, F. Liu, and Q. Yang. Cross-city transfer learning for deep spatio-temporal prediction. arXiv preprint arXiv:1802.00386, 2018
25. Z. Wu, S. Pan, G. Long, J. Jiang, and C. Zhang. Graph wavenet for deep spatial-temporal graph modeling. arXiv preprint arXiv:1906.00121, 2019
26. H. Yao, Y. Liu, Y. Wei, X. Tang, and Z. Li. Learning from multiple cities: A meta-learning approach for spatial-temporal prediction. In *The World Wide Web Conference*, pages 2181–2191, 2019
27. B. Yu, H. Yin, and Z. Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. arXiv preprint arXiv:1709.04875, 2017
28. H. Yu, T. Li, W. Yu, J. Li, Y. Huang, L. Wang, and A. Liu. Regularized graph structure learning with semantic knowledge for multi-variates time-series forecasting. arXiv preprint arXiv:2210.06126, 2022
29. Y. Yuan, C. Shao, J. Ding, D. Jin, and Y. Li. A generative pre-training framework for spatio-temporal graph transfer learning. arXiv preprint arXiv:2402.11922, 2024
30. Zhang, W., Deng, L., Zhang, L., Wu, D.: A survey on negative transfer. IEEE/CAA Journal of Automatica Sinica **10**(2), 305–329 (2022)
31. Y. Zhang, Y. Li, X. Zhou, X. Kong, and J. Luo. Strans-gan: Spatially-transferable generative adversarial networks for urban traffic estimation. In *2022 IEEE International Conference on Data Mining (ICDM)*, pages 743–752. IEEE, 2022
32. Zhao, L., Song, Y., Zhang, C., Liu, Y., Wang, P., Lin, T., Deng, M., Li, H.: T-gcn: A temporal graph convolutional network for traffic prediction. IEEE Trans. Intell. Transp. Syst. **21**(9), 3848–3858 (2019)

Advanced Endoscopy Imaging with Automatic Feedback

D. M. Bappy[1], Donghwa Kang[1], Jinkyu Lee[2], Youngmoon Lee[3], Minsuk Koo[1], and Hyeongboo Baek[4(✉)]

[1] Department of Computer Science and Engineering, Incheon National University, Incheon, Republic of Korea
[2] Department of Computer Science and Engineering, Sungkyunkwan University, Suwon, Republic of Korea
[3] Department of Robotics, Hanyang University, Ansan, Republic of Korea
[4] Department of Artificial Intelligence, University of Seoul, Seoul, Republic of Korea
hbbaek359@gmail.com

Abstract. As we move towards a future where minimally invasive methods become the norm for surgeries and diagnostic procedures, it is increasingly vital to improve our strategies for viewing the organs and complex structures within our bodies. Image stitching presents an enticing solution, expanding our field of view by seamlessly weaving together a sequence of images. While existing stitching techniques do lean on the capabilities of endoscopy imaging, they, unfortunately, overlook the critical need for automated feedback when grappling with the complexities and challenges innate to endoscopy imaging. these methods struggle to stand firm against deformations and regions with low texture. In this paper, we introduce a robust endoscopic image-stitching algorithm designed to thrive in adversity. Its unique resilience to deformations and low-texture regions is reinforced by the inclusion of a radial basis function weighting that is paired harmoniously with location-dependent homography based on the corresponding locations of the strong features extracted by affine shape-adapted Hessian-Laplace detector. Crucially, this algorithm is steered by a sophisticated automatic feedback mechanism. This feedback system makes astute evaluations based on an image quality metric and the structural comparison between the sequences of endoscopy images. We have thoroughly validated the efficacy of our new approach using two public datasets, namely EndoSLAM and EndoAbS, under demanding conditions. The results eloquently illustrate the superior benefits of our technique. Our proposed method surpasses commonly employed techniques, delivering superior performance in quantitative metrics, including precision at 30.07%, recall at 114.89%, F1-score at 84.62%, and TRE at 46.07%.

Keywords: Endoscopy Imaging · Endoscopy Stitching · Homography · Feature Extraction · Feature matching.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15311, pp. 62–78, 2025.
https://doi.org/10.1007/978-3-031-78195-7_5

1 Introduction

Endoscopy is a common and essential tool in medical diagnostics and research, playing a pivotal role in identifying and treating diseases, particularly tumors [29]. Yet, there are limitations such as the inability to obtain the best field of view and magnify an image simultaneously [15]. Increasing magnification enhances detail but reduces the visual field, complicating comprehensive organ examination and disease evaluation [26]. With the advancement of computer technology, medical images can be shared online, facilitating remote consultations [27]. However, a single endoscopic image might not suffice for accurate judgment. Thus, the challenge lies in stitching together successive images for complete visualization, which is crucial for remote diagnosis and data measurement.

Image stitching involves addressing several interconnected problems. Firstly, there is the issue of data association, which involves finding common scene elements across various views, a topic discussed in-depth by Huang [9]. Secondly, we have the task of estimating a geometric transformation. This transformation should align with the data association and unify disparate views into a single, seamlessly stitched image [8]. These two tasks are usually undertaken concurrently [4]. Lastly, it is necessary to blend the individual images' intensities to ensure a consistent, smooth final image [28].

Of these, the sub-problem of data association is the most complex, particularly in surgical scenarios, and hence attracts significant research focus. A classic strategy for addressing this issue is to identify and extract image point features that correspond to unique landmarks within the scene and then match them across different views. This approach, known as feature-based stitching [18], has been thoroughly researched in recent years. Various well-known hand-crafted feature methodologies, such as Harris [20], SIFT [14], SURF [24], ORB [7], and FAST [12] have been applied to it. In more recent developments, data-driven features derived from deep neural networks are being employed for image stitching [2].

There are also other strategies for stitching that do not hinge on feature extraction. Direct and dense pixel-based registration techniques can be developed as an iterative optimization problem, with the goal of maximising similarity as calculated by mutual information [19] or other photometric similarity/difference measures [13]. With the increased use of deep learning across various fields, there are now proposed end-to-end stitching algorithms that rely on deep learning to deduce registration parameters [3].

Technology's evolution has substantially enhanced endoscopic instruments. Luo et al. [16] proposed a method that uses an Auto-Regressive (AR) model to predict human motion intentions for teleoperated tasks. This algorithm anticipates human movements, updating and adjusting robotic actions during Human-Robot Interactions (HRI) as necessary. Moreover, Su et al. [25] developed an innovative technique. They employed a swivel motion reconstruction method to mimic human-like behavior through kinematic mapping in robot redundancy.

Their groundbreaking framework merges an incremental learning approach with a deep convolutional neural network, facilitating efficient and speedy learning.

However, these kinds of tasks require automated system feedback regarding image quality. Owing to the endoscope's movement, limited field of view, and the complexity of human body structures, there is a high likelihood of missing crucial features in the image sequence. Therefore, it is essential to have real-time updates from the imaging process if a vital structure is not captured in the images due to the endoscope camera's unpredictable movement. For this reason, we provide automated system feedback during endoscopy regarding image quality, which is otherwise impossible to process manually.

Based on the analysis of existing literature, it becomes evident that traditional image stitching techniques have been primarily used for endoscopic imagery, despite their inherent limitations. Such approaches do not take into account the specific attributes of endoscopic images, resulting in significant difficulties in solving the stitching problem. In scenarios such as large motion, a blurred environment, low texture, and deformations, the existing techniques are prone to stitching failures, often without providing any insightful feedback.

In Section 3, we substantiate our proposed methodology by employing publicly accessible datasets. The comparison of stitched images in Figure 4 and 6 with recent techniques attests to the advantages gained from extracting a substantial number of robust features in texture-less and deformed regions. These features are then leveraged to compute a weighted local homography for stitching image sequences. The improvements compared to best performers in quantitative metrics, as depicted in Table 1–precision at 30.07%, recall at 114.89%, F1-score at 84.62%, and TRE at 46.07% undeniably affirm that our technique significantly outperforms commonly employed methods.

This paper offers several unique contributions:

- The introduction of a fully automated, feedback-oriented, robust stitching algorithm designed specifically for endoscopic image stitching.
- Our algorithm not only provides feedback to surgeons but also executes blur removal on endoscopic images based on image quality assessment.
- The implementation of an affine-shape adapted Hessian-Laplace detector to extract robust features from the images.
- Our approach uses these robust features to apply a weighted local homography designed with endoscopic imaging properties in mind.

2 Approach

In this section, we introduce our proposed approach for robust endoscopic image stitching with automated feedback. Figure 1 illustrates the various components of our technique.

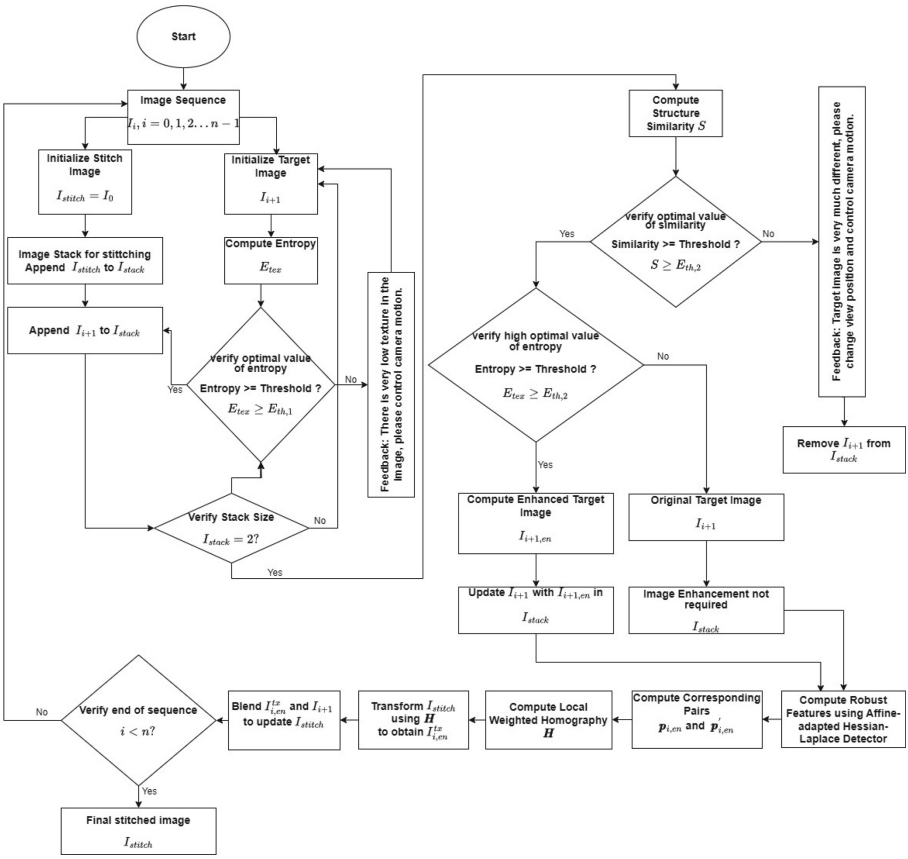


Fig. 1. Schematic of the proposed technique

2.1 Statistical Analysis-driven Automated Feedback

We make the assumption that the endoscopic images, denoted as I_i with $i = 1, 2, 3, ..., n$, are already aligned. Our processing of these images follows a sequential approach. To assess the level of texture availability, we employ a measure based on uncertainty or randomness within an image. This measure, denoted as $E_{i,tex}$, is calculated as follows:

$$E_{i,tex} = \sum_{k=1}^{255} (-p_k \log_2 p_k). \tag{1}$$

In Equation (1), p_k represents the probability associated with the gray level k of the image.

In order to assess the structural similarity between two consecutive images, I_i and I_{i+1}, we consider their luminance, contrast, and structure. This comparison is based on the following equations:

$$S(I_i, I_{i+1}) = [l(I_i, I_{i+1})]^{\alpha} \cdot [c(I_i, I_{i+1})]^{\beta} \cdot [s(I_i, I_{i+1})]^{\gamma}. \tag{2}$$

Here, the terms are defined as follows:

$$l(I_i, I_{i+1}) = \frac{2\mu_{I_i}\mu_{I_{i+1}} + \epsilon_1}{\mu_{I_i}^2 + \mu_{I_{i+1}}^2 + \epsilon_1},$$
$$c(I_i, I_{i+1}) = \frac{2\sigma_{I_i}\sigma_{I_{i+1}} + \epsilon_2}{\sigma_{I_i}^2 + \sigma_{I_{i+1}}^2 + \epsilon_2},$$
$$s(I_i, I_{i+1}) = \frac{\sigma_{I_i I_{i+1}} + \epsilon_3}{\sigma_{I_i}\sigma_{I_{i+1}} + \epsilon_3}$$

where μ_{I_i} and $\mu_{I_{i+1}}$, σ_{I_i} and $\sigma_{I_{i+1}}$, and $\sigma_{I_i I_{i+1}}$ represent the local means, standard deviations, and cross-covariance for images I_i and I_{i+1}.

2.2 Blur Removal for Enhanced Image Quality

To ensure that the texture information of a good quality endoscopic image is not lost during the blur removal process, we adopt a selective approach based on the score obtained from Equation (1). Randomly removing blur may result in the degradation of image quality and loss of texture details. The blur removal technique discussed in [5] is employed for this purpose. Consequently, the computation of the enhanced images is carried out as follows:

$$I_{i+1,en} = \frac{I_{i+1} - [1 - t(x)]A}{t(x)}. \tag{3}$$

Here, $t(x)$ represents the transmission map, and A denotes the single color in the image where the transmission map is 0.

2.3 Robust Feature estimation using Affine-Shape Adapted Hessian-Laplace Detector

To initiate our procedure, we employ a detector based on the Hessian matrix to identify salient points within the scale space. The Hessian matrix is essentially a matrix of second-order partial derivatives, obtained from the Taylor series expansion. It is articulated as:

$$H_{i,en} = \begin{bmatrix} I_{i,en,xx}(x;\sigma_d) & I_{i,en,xy}(x;\sigma_d) \\ I_{i,en,yx}(x;\sigma_d) & I_{i,en,yy}(x;\sigma_d) \end{bmatrix} \tag{4}$$

Here, $I_{i,en,xx}$, $I_{i,en,xy}$, and $I_{i,en,yy}$ denote second-order derivatives that are calculated using Gaussian kernels of scale σ_d.

Next, we employ a scale-normalized Laplacian, which helps us choose the appropriate scale for a detected point. It is expressed as:

$$Laplacian(x;\sigma_d) = \sigma_d^2 \left| I_{i,en,xx}(x;\sigma_d) + I_{i,en,yy}(x;\sigma_d) \right| \tag{5}$$

We can estimate the affine shape of the neighborhood surrounding the detected point by leveraging the eigenvalues of the second-moment matrix. This matrix is represented as:

$$M_{i,en} = \sigma_d^2 g(\sigma_I) * \begin{bmatrix} I_{i,en,x}^2(x;\sigma_d) & I_{i,en,x}I_{i,en,y}(x;\sigma_d) \\ I_{i,en,x}I_{i,en,y}(x;\sigma_d) & I_{i,en,y}^2(x;\sigma_d) \end{bmatrix} \tag{6}$$

In this scenario, the derivatives are averaged over the detected point's neighborhood by applying a smoothing process using a Gaussian window of scale σ_I.

2.4 Location-Based Weighted Homography Estimation for Feature Correspondence

Consider reference and target images denoted as $I_{i,en}$ and $I_{i+1,en}$. We compute matching pairs $\boldsymbol{p}_{i,en} = [x, y]^T$ and $\boldsymbol{p}_{i,en}^{'} = [x^{'}, y^{'}]^T$ from detected features x_r in previous section using these images. The pairs are derived from robustly identified feature points discussed in the previous section, using the Brute-Force (BF) matching algorithm [10]. The homographies between these pairs can be defined as

$$\boldsymbol{p}_{i,en}^{'} = \boldsymbol{h}(\boldsymbol{p}_{i,en}) \tag{7}$$

From Equation (7), we can perform a simple matrix expansion resulting in the following expressions:

$$x^{'} = \frac{h_1 x + h_2 y + h_3}{h_7 x + h_8 y + h_9} \tag{8}$$

$$y^{'} = \frac{h_4 x + h_5 y + h_6}{h_7 x + h_8 y + h_9} \tag{9}$$

In homogeneous coordinates, Equation 7 is represented as

$$\tilde{\boldsymbol{p}}_{i,en}^{'} = \boldsymbol{H}(\tilde{\boldsymbol{p}}_{i,en}) \tag{10}$$

where $\tilde{\boldsymbol{p}}_{i,en}^{'} = [x^{'}, y^{'}, 1]^T$, $\tilde{\boldsymbol{p}}_{i,en} = [x, y, 1]^T$ and $\boldsymbol{H}$ is a 3×3 matrix.

The columns of $\boldsymbol{H}$ are denoted by $\boldsymbol{h}_1 = [h_1, h_4, h_7]^T$, $\boldsymbol{h}_2 = [h_2, h_5, h_8]^T$, and $\boldsymbol{h}_3 = [h_3, h_6, h_9]^T$. By taking a cross-product on both sides of Equation 10, we get:

$$\boldsymbol{0}_{3\times1} = \tilde{\boldsymbol{p}}_{i,en}^{'} \times \boldsymbol{H}\tilde{\boldsymbol{p}}_{i,en} \tag{11}$$

This equation (11) can be reformulated as follows:

$$\boldsymbol{0}_{3\times1} = \begin{bmatrix} \boldsymbol{0}_{3\times1} & -\tilde{\boldsymbol{p}}_{i,en}^T & y^{'}\tilde{\boldsymbol{p}}_{i,en}^T \\ -\tilde{\boldsymbol{p}}_{i,en}^T & \boldsymbol{0}_{3\times1} & -x^{'}\tilde{\boldsymbol{p}}_{i,en}^T \\ -y^{'}\tilde{\boldsymbol{p}}_{i,en}^T & x^{'}\tilde{\boldsymbol{p}}_{i,en}^T & \boldsymbol{0}_{3\times1} \end{bmatrix} \begin{bmatrix} \boldsymbol{h}_1 \\ \boldsymbol{h}_2 \\ \boldsymbol{h}_3 \end{bmatrix} \tag{12}$$

The 9×1 vector in Equation 12 is referred to as $\boldsymbol{h}$. Considering that only two rows of the 3×9 matrix in Equation 12 are linearly independent, we can determine $\boldsymbol{h}$ utilizing a collection of N corresponding points, denoted as $\tilde{\boldsymbol{p}}i, en, kk = 1^N$ and $\tilde{\boldsymbol{p}}i, en, k^{'}k = 1^N$.

$$\boldsymbol{h} = \underset{\boldsymbol{h}}{\operatorname{argmin}} \sum_{k=1}^{N} \left\| \begin{bmatrix} \boldsymbol{a}_{k,1} \\ \boldsymbol{a}_{k,2} \end{bmatrix} \boldsymbol{h} \right\|^2 = \underset{\boldsymbol{h}}{\operatorname{argmin}} \left\| \boldsymbol{A}\boldsymbol{h} \right\|^2 \tag{13}$$

The components $\boldsymbol{a}_{k,1}$ and $\boldsymbol{a}_{k,2}$ correspond to the two rows of the matrix specified in Equation 12. We will also apply a constraint of $\|\boldsymbol{h}\|^2 = 1$ since a homographic transformation possesses only 8 degrees of freedom.

The paper [30], introduced the Moving DLT (Direct Linear Transform) framework to approximate local homography. This method incorporates locality-enforcing weights into the objective of Equation 13. The estimation of local homography at the position $\boldsymbol{p}_{i,en,j}$ is conducted as follows:

$$\boldsymbol{h}_j = \underset{\boldsymbol{h}_j}{\operatorname{argmin}} \sum_{k=1}^{N} w_{j,k} \left\| \begin{bmatrix} \boldsymbol{a}_{k,1} \\ \boldsymbol{a}_{k,2} \end{bmatrix} \boldsymbol{h} \right\|^2 \tag{14}$$

The calculation of scalar weights, denoted as $w_{j,k}k = 1^N$, is determined by changes relative to $\boldsymbol{p}_{i,en,j}$.

$$w_{j,k} = \left(\left\| \boldsymbol{p}_{i,en,k} - \boldsymbol{p}_{i,en,j} \right\|^2 * \log \left\| \boldsymbol{p}_{i,en,k} - \boldsymbol{p}_{i,en,j} \right\| \right) / \sigma^2 \tag{15}$$

We can reformulate Equation 14 as follows:

$$\boldsymbol{h}_j = \underset{\boldsymbol{h}_j}{\operatorname{argmin}} \left\| \boldsymbol{W}_j \boldsymbol{A}\boldsymbol{h} \right\|^2 \tag{16}$$

The weight matrix $\boldsymbol{W}_j \in \mathbb{R}^{2N \times 2N}$ is structured as follows:

$$\boldsymbol{W}_j = diag\left(\left[w_{1,j} w_{1,j} ... w_{N,j} w_{N,j} \right] \right) \tag{17}$$

In the research [30], they propose the concept of allotting increased weights to data that is closer to $\boldsymbol{p}_{i,en,k}$. This enhances the ability of the projective warp, $\boldsymbol{H}$, with respect to the local structure around $\boldsymbol{p}_{i,en,k}$. They also incorporate an offset parameter, γ, to circumvent numerical issues.

A key aspect of this concept is that the computation of local homography is only applicable to the areas of the target image that overlap with the reference image. For each pixel in regions that do not overlap, the transformation is computed as a weighted linear combination of the local homographies in the overlapping areas. Thus, an appropriate set γ is crucial to prevent extrapolation.

This overarching idea is highly effective in the context of real-world images, which often contain abundant textures and rigid structures. However, complications arise during endoscopy imaging, due to the presence of low texture and numerous deformations. To address these challenges, we consider an endoscopy

imaging property adapted weight function in Equation 15 that tackles deformations and outliers caused by low texture. This is achieved by combining squared distance and log distance. This combined approach allows for the flexible fitting of deformation patterns, while the log distance alone aids in reducing outliers.

When $\boldsymbol{p}_{i,en,k}$ shifts continuously in the source image domain $I_{i,en}$, the corresponding homography estimation $\boldsymbol{H}$ adjusts smoothly. This process results in a dynamic warp that can flexibly adapt to data and transform the source image into $I^{tx}_{i,en}$.

Blending Warped Images

In this section, we merge the transformed source image $I^{tx}_{i,en}$ with the target image $I_{i+1,en}$ to generate the final stitched image I_{stitch}. The computation for the stitched image [23] is as follows:

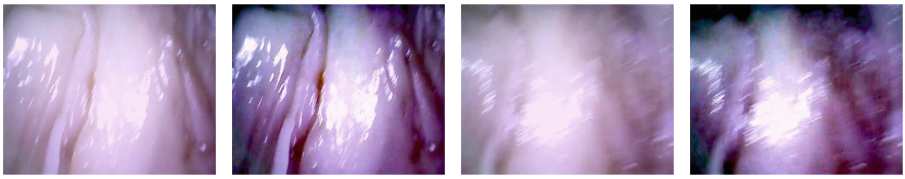
$$I_{stitch} = \alpha I^{tx}_{i,en} + (1 - \alpha) I_{i+1,en} \tag{18}$$

(a) (b) (c)

Fig. 2. Images with different entropy: (a) higher entropy (b) moderate entropy, and (c) lower entropy

3 Experiments

We have utilized three diverse, publicly accessible datasets: EndoSLAM [21], EndoAbs [22], and Hamlyn [16]. These datasets were selected purposefully to encompass a variety of conditions, including differing lighting circumstances, small fields of view, low-texture regions, and deformed areas, thereby providing a comprehensive evaluation of our method. The texture of an image is crucial for image-guided systems as it enhances feature extraction reliability and boosts computational precision.



(a) (b) (c) (d)

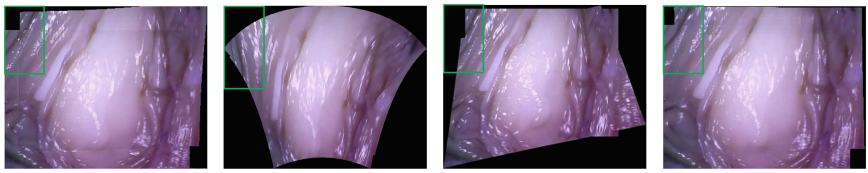
Fig. 3. Image Enhancement: (a) image with moderate blur, (b) enhanced image from moderate blur, (c) image with extreme blur and (d) enhanced image from extreme blur

3.1 Qualitative Evaluation

Figure 2 depicts three images from the EndoSLAM dataset, each presenting varying levels of texture $E_{i,tex}$. In Figure 3(a), the image displays a moderate blur with an entropy value of 7.21. After deblurring, an enhanced image is produced, as seen in Figure 3(b), which possesses a richer texture and an improved entropy value of 7.87. This refined image is more conducive to robust feature extraction. Conversely, the image in Figure 3(c) exhibits an intense blur with an entropy of 7.13. Following deblurring, the resultant image on Figure 3(d) shows only a slight improvement in texture quality, with an entropy of approximately 7.17. Despite the enhancement, this image remains below the optimal threshold and is not ideal for stitching.

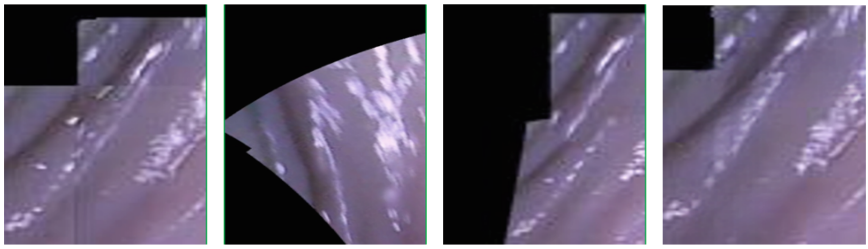
Figure 4 illustrates that our proposed method generates higher quality and more accurate stitching in comparison to other cutting-edge techniques. To further scrutinize the performance, a region in Figure 5 where robust feature extraction and subsequent image frame warping proved challenging was cropped. The APAP technique, as illustrated in Figure 5(a) and outlined in [30], presents a comparable outcome; specifically, it fails to accurately merge the tissue area, resulting in a warped region. The AutoStitch method [6] generates a malformed area and even excludes some parts entirely from the original image sequence. As we shift our focus to Figure 5(c), the latest wide parallax technique [11] misplaces the tissue region entirely due to the lack of robust features and incorrect warping in that particular area. In stark contrast, our proposed method delivers perfect alignment of the tissue region. This is achieved by successfully extracting robust features in these areas thanks to the affine adaptation of the Hessian-Laplace detector. Subsequently, these features are used to calculate a weighted homography, which enables precise warping. This is made possible by the appropriate distribution of weight based on whether a randomly located feature is sufficiently close to the target feature location.

Figure 6 and 7 showcase the exemplary results achieved when stitching using the stereo pairs from the Hamlyn and EndoAbs datasets, respectively. The Hamlyn dataset, while offering images of high resolution, introduces intricate challenges due to the presence of elements such as blood, deformations, and surgical instruments. Impressively, our proposed method managed to stitch the Hamlyn stereo pairs with remarkable accuracy, highlighting its resilience and effec-



(a) (b) (c) (d)

Fig. 4. Image stitching using state of art techniques: (a) APAP (b) AutoStitch (c) Wide Parallax, and (d) Proposed



(a) (b) (c) (d)

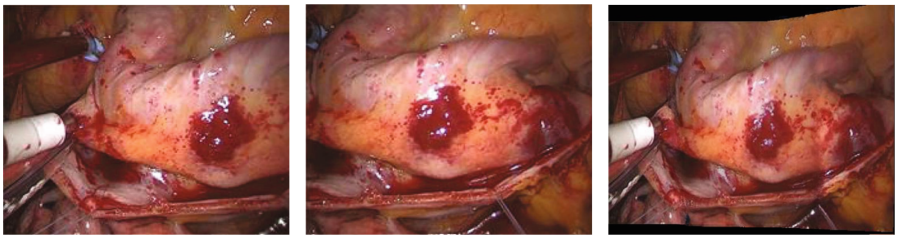
Fig. 5. Magnified stitched region using state-of-the-art feature based techniques: (a) APAP (b) AutoStitch (c) wide parallax, and (d) Proposed

tiveness even when faced with occlusions caused by surgical tools. Turning our attention to the EndoAbs dataset, it's worth noting that our technique consistently delivered despite the challenging low-light conditions. Such challenging scenarios are not uncommon in surgical environments. The successes displayed herein emphasize the potential of our method to furnish an expanded field of view, characterized by both precision and quality, from a stereo image sequence. This expanded perspective can be invaluable in enhancing surgical operations and diagnostic procedures.

Figure 8 depicts an instance of a failure during the image stitching process. As evident from the figure, the extreme blur in Figure 8(a) (the left image) impedes the algorithm's capability to establish correspondences between image sequences. The particular image lacks sufficient texture information, making it challenging to extract robust features. Consequently, the stitching process fails noticeably as shown in Figure 8(b). In such scenarios, our algorithm offers feedback to the experts, suggesting adjustments in the endoscope's movement to capture higher-quality images.

Figure 9 illustrates the scenario of the maximum angular limit in our proposed technique. Beyond this limit, the stitching process is affected as the angle increases. To assess the maximum allowable angle for flawless stitching, we set the scale to 0.5 and varied the angle values to 5, 10, 15, 45, and 50 degrees. Table 1 includes the maximum limit angle of 45 degrees, beyond which the stitching process begins to degrade. In Figure 9(a)(b)(c), the green box region demonstrates

nearly perfect stitching. However, in Figure 9(d)(e), we observe a degradation in the stitching process within the green box region as the angle increases from 45 to 50 degrees. These quantitative results are reflected in Table 1, where the Target Registration Error (TRE) is 0.05 for angles 5, 10, and 15 and increases to 1.3 for the 45-degree angle.



(a) (b) (c)

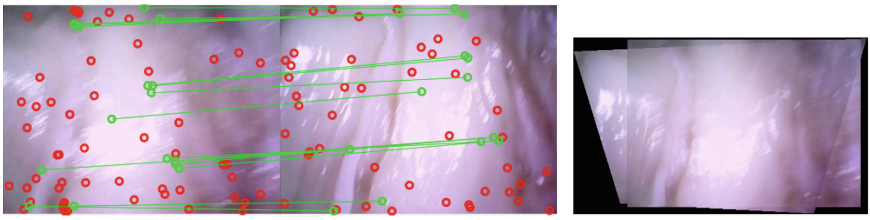
Fig. 6. Image Stitching using stereo pairs: (a) Hamlyn left view (b) Hamlyn right view (c) Hamlyin stitched

(a) (b) (c)

Fig. 7. Image Stitching using stereo pairs: (a) EndoAbs left view (b) EndoAbs right view, and (c) EndoAbs stitched

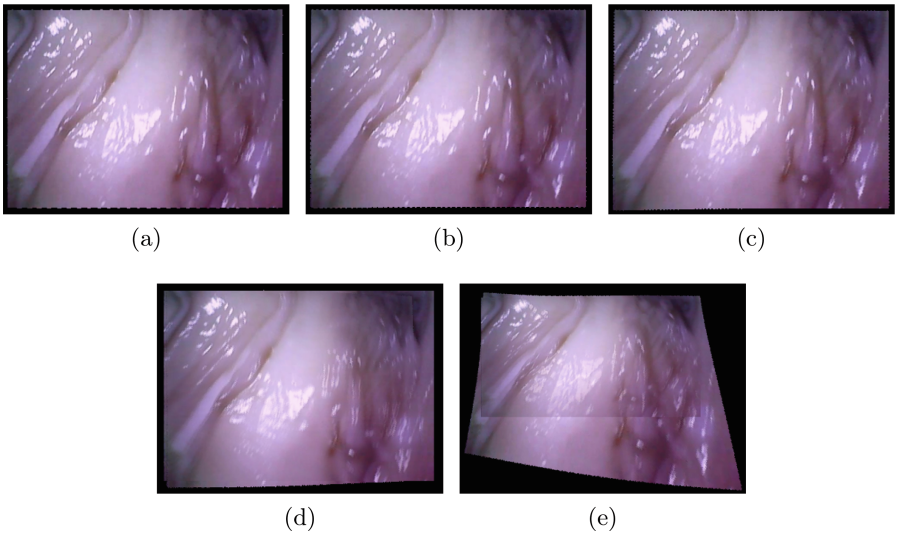
3.2 Quantitative Evaluation

To evaluate the efficacy of the affine-adapted Hessian-Laplace detector, we compared it with two recently adapted feature detectors for endoscopy imaging: Pyramid ORB [31] and Improved SIFT [16]. We employed the same image sequence these detectors use for feature extraction. When compared to [31] with 800 key points, our proposed method extracts 17,261 key points, which is approximately 21.5 times greater. Furthermore, when compared to [16] with 113 corresponding points, our technique maintains 1,505 correspondences, which is 13.3 times greater. The large number of key points that our affine-adapted detector extracts features from most regions, and the high number of correspondences allows for dense matching.



(a) (b)

Fig. 8. Failure case of image stitching: (a) matching pairs, and (b) stitched image



(a) (b) (c)

(d) (e)

Fig. 9. Stitching images using the proposed method at a scale of 0.5 with various rotations: (a) 5-degree rotation (b) 10-degree rotation (c) 15-degree rotation (d) 45-degree rotation, and (e) 50-degree rotation

Feedback Evaluation In our research, we utilized the EndoSLAM dataset to determine optimal feedback threshold values. After a thorough examination of multiple trials, we identified optimal thresholds for both entropy and structural similarity, which stood at 7.25 and 0.76, respectively. These identified thresholds play a pivotal role in facilitating feedback for our proposed algorithm. Specifically, an endoscopy image with entropy below the threshold suggests that the image might lack sufficient texture, while a lower structural similarity value compared to the set threshold suggests a lack of adequate structure similarity in the consecutive image sequence for successful stitching. Consider the images shown in Figure 2(a)(b)(c) with entropies of 7.45, 7.21, and 7.13 respectively. The progressive increase in entropy value signifies an image enriched in texture. Based on these observations, our method leverages these entropy values to provide the

surgeon with feedback regarding the optimal positioning and movement of the endoscope. Additionally, the structural similarity value obtained from equation 2 assists in determining when there is an inadequate overlapping region for effective comparison and feature extraction. For instance, Figure 3(b) depicts how improved image quality can accentuate the texture in an image. Yet, Figure 3(d) illustrates that even with enhanced image quality, extracting meaningful information for subsequent processes remains a challenge.

We need the feedback system to operate in real time during the endoscopy procedure to decide whether to accept or reject an image. However, real-time stitching is not necessary, as it is time-consuming and not crucial for the surgeon's immediate evaluation. The priority is to obtain high-quality images during the procedure, ensuring they can be stitched later without information loss. This will provide the surgeon with a high-quality stitched image for better diagnosis. Our feedback system can evaluate and correct endoscopy images at a rate of 10 frames per second.

Table 1. Evaluating the Efficacy of Various Techniques

Orientation		000000 Methods																			
Scale	Rotation	SIFT				SURF				ORB				AKAZE				Proposed			
		P	R	F1	TRE	P	R	F1	TRE	P	R	F1	TRE	P	R	F1	TRE	P	R	F1	TRE
0.9	5	0.65	0.17	0.27	0.45	0.59	0.29	0.39	8.88	0.59	0.15	0.23	3.87	0.75	0.32	0.3	3.70	0.95	0.77	0.85	0.06
0.9	10	0.64	0.15	0.25	0.85	0.60	0.15	0.23	10.7	0.69	0.12	0.20	6.71	0.82	0.34	0.48	8.02	0.93	0.66	0.77	0.06
0.9	15	0.65	0.12	0.20	1.15	0.50	0.11	0.19	17.8	0.72	0.11	0.20	7.03	0.81	0.35	0.49	10.6	0.88	0.56	0.68	0.05
0.8	5	0.48	0.10	0.13	0.41	0.41	0.19	0.26	9.46	0.73	0.22	0.34	3.62	0.67	0.25	0.36	3.53	0.95	0.89	0.92	0.06
0.8	10	0.47	0.10	0.13	0.79	0.48	0.11	0.18	10.7	0.68	0.20	0.31	7.01	0.69	0.25	0.37	7.31	0.93	0.75	0.84	0.06
0.8	15	0.51	0.10	0.13	1.11	0.46	0.10	0.16	16.6	0.58	0.16	0.25	7.20	0.73	0.25	0.38	10.8	0.89	0.63	0.74	0.05
0.7	5	0.40	0.10	0.10	0.36	0.60	0.23	0.33	8.21	0.67	0.14	0.23	3.13	0.71	0.21	0.33	3.66	0.93	0.89	0.91	0.05
0.7	10	0.39	0.10	0.11	0.70	0.55	0.14	0.23	10.2	0.61	0.13	0.22	6.18	0.76	0.23	0.35	7.02	0.92	0.87	0.90	0.06
0.7	15	0.41	0.10	0.11	1.10	0.49	0.11	0.18	15.8	0.55	0.14	0.22	7.22	0.80	0.22	0.34	10.4	0.90	0.76	0.83	0.05
0.6	5	0.39	0.10	0.11	0.31	0.33	0.11	0.17	7.49	0.73	0.10	0.16	2.88	0.55	0.11	0.19	3.32	0.93	0.88	0.90	0.04
0.6	10	0.44	0.10	0.12	0.61	0.37	0.11	0.16	14.5	0.56	0.10	0.11	5.33	0.66	0.14	0.23	6.22	0.93	0.86	0.89	0.05
0.6	15	0.45	0.10	0.13	0.89	0.44	0.10	0.17	14.9	0.69	0.10	0.16	7.84	0.66	0.14	0.23	6.24	0.90	0.82	0.86	0.05
0.5	5	0.48	0.10	0.15	0.27	0.60	0.12	0.21	6.13	0.71	0.10	0.10	2.33	0.46	0.10	0.10	2.52	0.84	0.78	0.81	0.04
0.5	10	0.45	0.10	0.13	0.51	0.51	0.10	0.17	11.8	0.73	0.10	0.12	4.43	0.43	0.10	0.10	4.57	0.93	0.86	0.89	0.05
0.5	15	0.47	0.10	0.14	0.74	0.53	0.10	0.16	17.1	0.70	0.10	0.10	6.73	0.41	0.10	0.10	7.64	0.86	0.77	0.81	0.05
0.5	45	0.36	0.08	0.11	2.10	0.29	0.08	0.06	18.3	0.42	0.05	0.08	7.95	0.38	0.04	0.07	8.71	0.31	0.26	0.28	1.30
Average		0.47	0.10	0.14	0.77	0.48	0.13	0.20	12.4	0.64	0.12	0.19	5.58	0.65	0.19	0.28	6.51	**0.88**	**0.74**	**0.79**	**0.12**

Assessment Through Simulated Transformations Endoscopy imaging encounters various obstacles, with the random motion of the endoscope being one of the primary challenges. This motion may introduce minor rotations, and the tissues encountered can also undergo deformation. Bearing this scenario in mind, we select an arbitrary frame from the EndoSLAM dataset and subject it to transformations such as rotation followed by scaling. This procedure yields 15

unique transformation combinations denoted as T_{GT}. The transformation matrix T_{est} can then be estimated based on the matched feature points identified in the feature correspondence section.

We used a specific threshold value of 0.8 for all methods, including our proposed method, as suggested by [17] in Table 1 to evaluate quantitative performance. After numerous observations and experiments with different thresholds on endoscopy images, we determined that a threshold value of 0.8 was the most effective.

For quantitative evaluation with simulated data, we adopt an approach akin to that presented in [1]. To assess the results, we utilized metrics such as precision, recall, and F1-score to quantify feature detection and matching accuracy. Additionally, the TRE was used to gauge the precision of the registration.

Table 1 displays the quality evaluation metrics for various methods. It is evident from the table that our proposed method significantly surpasses other best performers in commonly used techniques, delivering superior performance in quantitative metrics. Specifically, we achieved a precision of 30.07%, recall of 114.89%, F1-score of 84.62%, and TRE of 46.07%. The superior performance of our method is primarily attributed to the innovative affine-shape adapted detector, which excels in detecting a substantial number of robust features, especially in low-texture regions. In contrast, other methods under comparison tend to identify features primarily around specular reflection-affected areas, leading them to estimate less accurate matches. Remarkably, our method managed to extract a commendable 78 percent of accurate matches from the vast number of features identified. Such a substantial figure is pivotal for our stitching algorithm, as it relies on local homography.

(a) APAP (b) Proposed (c) Proposed-Corrected

Fig. 10. Ablation Study

4 Ablation Study

We conducted an ablation study to verify the importance of reducing image blur, which can significantly affect stitching and feature extraction. In both

Table 2. Ablation Study

methods	P	R	F1	TER
SIFT	0.41	0.53	0.55	1.04
SURF	0.21	0.016	0.03	NA
ORB	0.19	0.01	0.02	NA
Proposed	0.65	0.71	0.69	0.08
Proposed-Corrected	0.79	0.76	0.77	0.02

Figure 10 and Table 2, "Proposed-Corrected" refers to the case of using reduced-blur images. Figures 10a and 10b demonstrate that the APAP and our proposed methods perform poorly without blur reduction. Conversely, after reducing blur, the "Proposed-Corrected" method in Figure 10c stitches images almost perfectly. Wide Parallax and Autostitch results are not included because these methods failed to stitch the images due to insufficient feature detection in blurred images.

Additionally, we performed a quantitative evaluation of feature extraction techniques using both blurred and corrected images, as shown in Table 2. The metrics indicate that SIFT, SURF, and ORB methods perform poorly with blurred images. Even our proposed technique performs averagely with blurred images. However, after reducing blur, our "Proposed-Corrected" method shows significantly improved performance.

5 Conclusions and Future Work

In our study, we introduce an innovative automatic feedback system critical to contemporary endoscopic tools used in medical surgeries and diagnostics. This automatic feedback notifies the surgeon and assists the robot in adjusting the endoscope's motion during the imaging process. Moreover, it provides crucial information to experts, enabling them to acquire meaningful images necessary for diagnosis and surgery.

Our technique successfully eliminates blur based on feedback, ensuring our algorithm does not degrade the quality of high-grade images by inappropriately applying the smoke removal technique. The affine shape-adapted Hessian-Laplace detector incorporated in our approach also extracts robust features from complex endoscopic images, particularly in deformed and low-texture areas. Finally, we implemented an endoscopy property-adapted weighting to estimate local homography, enhancing the stitching in deformed and low-texture regions.

Acknowledgement. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2023-00250742, 2022R1A4A3018824, RS-2024-00438248, RS-2022-00155885). This research was also supported by the MSIT(Ministry of Science and ICT), Korea under the ITRC(Information Technology Research Center) support program(IITP-2023-RS-2023-00259061) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation).

References

1. Ali, S., Rohr, K., Axer, M., Amunts, K., Eils, R., Wörz, S.: Registration of ultra-high resolution 3d pli data of human brain sections to their corresponding high-resolution counterpart. In: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017). pp. 415–419. IEEE (2017)
2. Bano, S., Vasconcelos, F., Shepherd, L.M., Vander Poorten, E., Vercauteren, T., Ourselin, S., David, A.L., Deprest, J., Stoyanov, D.: Deep Placental Vessel Segmentation for Fetoscopic Mosaicking. In: Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racoceanu, D., Joskowicz, L. (eds.) MICCAI 2020. LNCS, vol. 12263, pp. 763–773. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59716-0_73
3. Bano, S., Vasconcelos, F., Tella Amo, M., Dwyer, G., Gruijthuijsen, C., Deprest, J., Ourselin, S., Vander Poorten, E., Vercauteren, T., Stoyanov, D.: Deep Sequential Mosaicking of Fetoscopic Videos. In: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.-T., Khan, A. (eds.) MICCAI 2019. LNCS, vol. 11764, pp. 311–319. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32239-7_35
4. Bartoli, A.: Groupwise geometric and photometric direct image registration. IEEE Trans. Pattern Anal. Mach. Intell. **30**(12), 2098–2108 (2008)
5. Berman, D., Avidan, S., et al.: Non-local image dehazing. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1674–1682 (2016)
6. Brown, M., Lowe, D.G.: Automatic panoramic image stitching using invariant features. Int. J. Comput. Vision **74**, 59–73 (2007)
7. Chaudhari, K., Garg, D., Kotecha, K.: An enhanced approach in image mosaicing using orb method with alpha blending technique. International Journal of Advanced Research in Computer Science **8**(5) (2017)
8. Chum, O., Matas, J.: Homography estimation from correspondences of local elliptical features. In: Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012). pp. 3236–3239. IEEE (2012)
9. Huang, T.S., Netravali, A.N.: Motion and structure from feature correspondences: A review. Proc. IEEE **82**(2), 252–268 (1994)
10. Jakubović, A., Velagić, J.: Image feature matching and object detection using brute-force matchers. In: 2018 International Symposium ELMAR. pp. 83–86. IEEE (2018)
11. Jia, Q., Li, Z., Fan, X., Zhao, H., Teng, S., Ye, X., Latecki, L.J.: Leveraging line-point consistence to preserve structures for wide parallax image stitching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12186–12195 (2021)
12. Konen, W., Tombrock, S., Scholz, M.: Robust registration procedures for endoscopic imaging. Med. Image Anal. **11**(6), 526–539 (2007)
13. Levin, A., Zomet, A., Peleg, S., Weiss, Y.: Seamless Image Stitching in the Gradient Domain. In: Pajdla, T., Matas, J. (eds.) ECCV 2004. LNCS, vol. 3024, pp. 377–389. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-24673-2_31
14. Li, Y., Wang, Y., Huang, W., Zhang, Z.: Automatic image stitching using sift. In: 2008 International Conference on Audio, Language and Image Processing. pp. 568–571. IEEE (2008)
15. Liu, S., Wang, L., Liu, H., Su, H., Li, X., Zheng, W.: Deriving bathymetry from optical images with a localized neural network algorithm. IEEE Trans. Geosci. Remote Sens. **56**(9), 5334–5342 (2018)

16. Liu, Y., Tian, J., Hu, R., Yang, B., Liu, S., Yin, L., Zheng, W.: Improved feature point pair purification algorithm based on sift during endoscope image stitching. Front. Neurorobot. **16**, 840594 (2022)
17. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision **60**, 91–110 (2004)
18. Milgram, D.L.: Computer methods for creating photomosaics. IEEE Trans. Comput. **100**(11), 1113–1119 (1975)
19. Miranda-Luna, R., Daul, C., Blondel, W.C., Hernandez-Mier, Y., Wolf, D., Guillemin, F.: Mosaicing of bladder endoscopic image sequences: Distortion calibration and registration algorithm. IEEE Trans. Biomed. Eng. **55**(2), 541–553 (2008)
20. Okumura, K.i., Raut, S., Gu, Q., Aoyama, T., Takaki, T., Ishii, I.: Real-time feature-based video mosaicing at 500 fps. In: 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 2665–2670. IEEE (2013)
21. Ozyoruk, K.B., Gokceler, G.I., Bobrow, T.L., Coskun, G., Incetan, K., Almalioglu, Y., Mahmood, F., Curto, E., Perdigoto, L., Oliveira, M., et al.: Endoslam dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos. Med. Image Anal. **71**, 102058 (2021)
22. Penza, V., Ciullo, A.S., Moccia, S., Mattos, L.S., De Momi, E.: Endoabs dataset: Endoscopic abdominal stereo image dataset for benchmarking 3d stereo reconstruction algorithms. The International Journal of Medical Robotics and Computer Assisted Surgery **14**(5), e1926 (2018)
23. Porter, T., Duff, T.: Compositing digital images. In: Proceedings of the 11th annual conference on Computer graphics and interactive techniques. pp. 253–259 (1984)
24. Rong, W., Chen, H., Liu, J., Xu, Y., Haeusler, R.: Mosaicing of microscope images based on surf. In: 2009 24th International Conference Image and Vision Computing New Zealand. pp. 271–275. IEEE (2009)
25. Su, H., Qi, W., Hu, Y., Karimi, H.R., Ferrigno, G., De Momi, E.: An incremental learning framework for human-like redundancy optimization of anthropomorphic manipulators. IEEE Trans. Industr. Inf. **18**(3), 1864–1872 (2020)
26. Tang, Y., Liu, S., Deng, Y., Zhang, Y., Yin, L., Zheng, W.: Construction of force haptic reappearance system based on geomagic touch haptic device. Comput. Methods Programs Biomed. **190**, 105344 (2020)
27. Tang, Y., Liu, S., Deng, Y., Zhang, Y., Yin, L., Zheng, W.: An improved method for soft tissue modeling. Biomed. Signal Process. Control **65**, 102367 (2021)
28. Tian, F., Shi, P.: Image mosaic using orb descriptor and improved blending algorithm. In: 2014 7th International Congress on Image and Signal Processing. pp. 693–698. IEEE (2014)
29. Yang, B., Liu, C., Zheng, W., Liu, S., Huang, K.: Reconstructing a 3d heart surface with stereo-endoscope by learning eigen-shapes. Biomed. Opt. Express **9**(12), 6222–6236 (2018)
30. Zaragoza, J., Chin, T.J., Brown, M.S., Suter, D.: As-projective-as-possible image stitching with moving dlt. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2339–2346 (2013)
31. Zhang, Z., Wang, L., Zheng, W., Yin, L., Hu, R., Yang, B.: Endoscope image mosaic based on pyramid orb. Biomed. Signal Process. Control **71**, 103261 (2022)

Deep Prior Based Limited-Angle Tomography

D. M. Bappy[1], Donghwa Kang[1], Jinkyu Lee[2], Youngmoon Lee[3], and Hyeongboo Baek[4(✉)]

[1] Department of Computer Science and Engineering, Incheon National University, Incheon, Republic of Korea
[2] Department of Computer Science and Engineering, Sungkyunkwan University, Suwon, Republic of Korea
[3] Department of Robotics, Hanyang University, Ansan, Republic of Korea
[4] Department of Artificial Intelligence, University of Seoul, Seoul, Republic of Korea
hbbaek359@gmail.com

Abstract. In the process of reconstructing images from data acquired within a limited angular range, we encounter what is termed limited-angle tomography. The deficiency of complete data in this context results in artifacts, commonly appearing as streaks or missing structures, which can significantly compromise the quality of the reconstructed slice. This degradation gives rise to issues such as boundary distortion, blurred edges, and intensity bias, potentially leading to misinterpretation of the images. Hence, addressing artifacts in limited-angle tomography is crucial for clinical applications. Although deep learning-based reconstruction has shown impressive results in recent times, concerns about its robustness persist. To bolster the robustness of our proposed technique, we integrate prior information from a modified U-net with preprocessed input into the Relative Variation - Simultaneous Algebraic Reconstruction Technique (RV-SART) to provide insights into unmeasured data. Subsequently, the method extracts structure from the initially reconstructed slice through structure-texture decomposition. This process facilitates the reconstruction of high-quality CT images while suppressing pattern-like artifacts. Extensive experiments demonstrate that our approach surpasses both traditional and state-of-the-art learning techniques in terms of reconstruction quality and preservation of fine structures in noisy limited-angle reconstruction problems. Our technique provides improvements over the recent LRIP-net for a 90-degree scanning range in quantitative metrics such as PSNR by 17.48%, RMSE by 46.36%, and SSIM by 6.18%.

Keywords: Limited Angle Tomography · Deep Prior · Artifacts.

1 Introduction

Computed Tomography (CT) finds applications in diverse fields, including medical examinations [13] [18], industrial nondestructive testing [16], and security

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15311, pp. 79–95, 2025.
https://doi.org/10.1007/978-3-031-78195-7_6

inspection [14]. In the context of medical CT, increasing attention is being devoted to mitigating X-ray radiation doses due to their potential harm to patients [15]. The reduction of radiation dose holds practical significance for patient well-being. Strategies for achieving this objective include lowering the X-ray tube current or decreasing the number of projection views. For instance, a straightforward and effective approach is limited-angle CT scanning, where projection views are constrained to a specific angular range. Limited-angle CT scanning is encountered not only in medical applications but also in scenarios such as the nondestructive testing of pipelines in service [31]. In this context, the scanning environment necessitates CT scanning within a restricted angular range. Furthermore, limited-angle CT scanning is employed to enhance temporal resolution in coronary computed tomography angiography [7]. Within micro-CT applications, certain identified objects exhibit distinctive structures, including elongated and discoid shapes. In such cases, the object's rotational range is constrained due to limitations in imaging geometry. Alternatively, in specific projection views, complete X-ray absorption occurs, leading to projection loss [29]. In summary, limited-angle CT has garnered increasing attention in recent years.

The rise of Convolutional Neural Networks (CNNs) across various computer vision tasks has led to a growing popularity of deep learning approaches in the field of medical imaging. Pelt and Batenburg [26] introduced an artificial neural network-based algorithm for fast limited-angle image reconstruction. This method is essentially a weighted combination of the Filtered Back Projection (FBP) technique and learned filters. Boublil et al. [5] applied a CNN-based model to integrate multiple reconstructed results, while Kang et al. [22] developed a deep CNN model in the wavelet domain, training wavelet coefficients from CT images using the contourlet transform. While the aforementioned approaches have shown promising outcomes and notably enhanced the quality of reconstructed images, certain limitations persist. One drawback is the absence of guaranteed worst-case performance from these algorithms. Another limitation is their reliance on post-processing methods, overlooking data consistency.

It is evident that iterative methods and learning-based methods exhibit a primary distinction. The former demonstrates flexibility in addressing diverse information retrieval tasks by straightforwardly specifying parameters and optimization. Conversely, the latter necessitates laborious training to acquire knowledge of the model before testing and typically faces limitations tied to specialized tasks. The practical implementation of deep learning techniques in limited-angle tomography faces significant challenges attributed to two major factors: First, the potential lack of generalization to unseen data due to inadequate training data; and second, sensitivity to noise. Consequently, images reconstructed solely through deep learning methods may appear suboptimal. Additionally, we observe that limited-angle reconstruction consists of pattern-like artifacts, causing small structures to be invisible. This issue can be resolved if we can separate structure and texture from the reconstructed slice.

We propose a hybrid approach that combines deep learning with iterative reconstruction. The objective is to leverage Relative Variation - Simultaneous Algebraic Reconstruction The technique (RV-SART) to impose constraints on the reconstructed images, ensuring consistency with the measured projection data, while unmeasured information is supplemented as prior information from FBPConvNet with preprocessed input. To achieve this, we introduce a method known as Deep Prior Based RV-SART. Firstly, we employ the Simultaneous Algebraic Reconstruction Technique - Total Variation (SART-TV) using limited-angle data. Following this, the resulting reconstructed output is input into the CNN to extract details about unmeasured data. In the subsequent phase, the output from the CNN is incorporated into RV-SART, providing crucial prior information. In this stage, the initially reconstructed slice undergoes decomposition into structure-texture components, aiding in the extraction of structural information and the mitigation of pattern-like artifacts. This process guarantees the reconstruction of high-quality CT images.

In Figure 4, we showcase the application of our proposed reconstruction technique alongside a recent learning-based approach. The visual representation illustrates the effective suppression of artifacts and the highly accurate reconstruction of intricate structures achieved by our method. Substantiating these visual results, we present a comparative analysis using three key quantitative metrics in Table 1. This table reveals remarkable enhancements, including a 17.48% increase in PSNR, a 46.36% reduction in RMSE, and a 6.18% improvement in SSIM when compared to the second-best performance listed in the table.

2 Related work

Numerous academics have pondered the limitations of post-processing approaches and have shifted their focus towards emphasizing data consistency to enhance the quality of reconstructed images. Certain methods leverage neural networks to incorporate prior information into existing iterative reconstruction algorithms [1,3,8,17]. Specific approaches, as described in references [6,9,10,20,24,33] embrace an end-to-end strategy that suggests the unfolding of an iterative algorithm and comprehensive training as a deep neural network. This approach facilitates the integration of physical information into the architecture through data consistency blocks, which are combined with trainable CNN regularizers.

DIOR [20] is a recent technique that employs a hybrid approach in the residual domain. In contrast, our hybrid approach operates in the image domain. Our method is simpler architecture and more effective in reconstructing high-quality CT images, with lower computational complexity compared to DIOR.According to our straightforward hybrid technique, structure-texture decomposition proves to be more effective compared to separating high and low-frequency components. This is because we deliberately extract texture components as pattern-like noise using relative variation, and then successfully remove them from the reconstructed CT images.

The technique we propose utilizes prior information that falls within the realm of deep learning-based reconstruction, specifically categorized as image-to-image reconstruction. In this approach, direct access to raw measurements is not necessary, indicating that image reconstruction is independent of manufacturing protocols. Through the utilization of a CNN, our method effectively maps low-quality images to their high-quality counterparts without requiring additional inputs. In the domain of post-processing sparse-view CT reconstruction, CNN plays a pivotal role, with FBPConvNet [21] standing out as a notable representative. FBPConvNet adopts the FBP algorithm for image reconstruction and integrates a CNN structure based on U-Net [27] for image segmentation that includes both an encoder and a decoder. Notably, FBPConvNet introduces a skip connection between the input and output, enhancing the network's ability to discern subtle distinctions. To address the limitations of U-Net, Han et al. [19] introduced Framing U-Net as an alternative post-processing method, specifically tailored to meet frame conditions. Another innovative approach comes from Lee et al. [23], who employed a multi-level wavelet U-Net, showcasing superior results compared to traditional U-Net. Despite the significant success of these post-processing techniques, it is crucial to note that they do not incorporate sinogram information into their CNN architectures. As a result, in extremely ill-posed problems where there are very limited numbers of projections, the aforementioned networks fail to reconstruct high-quality CT images and suppress artifacts.

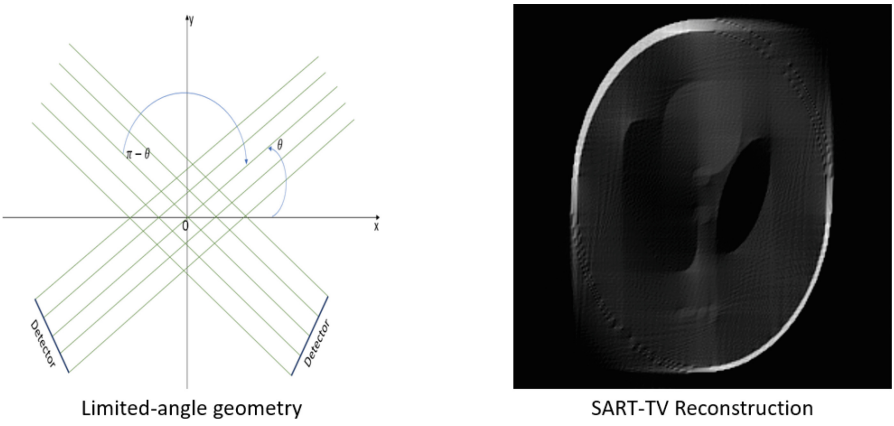


Fig. 1. Limited-angle parallel-beam scanning configuration

3 Approach

We examine the limited-angle parallel-beam CT in a two-dimensional setting. Figure 1 depicts the scanning configuration for this limited-angle parallel-beam

CT. Additionally, the initial reconstruction for the limited angle with a straightforward phantom model is presented in Figure 1. The reconstruction reveals artifacts, boundary distortion, and structure loss attributed to the insufficient number of projections obtained through limited-angle scanning.

3.1 Deep Prior Estimation

Jin et al. [21] introduced a post-processing image reconstruction technique named FBPConvNet, utilizing the FBP method to reconstruct the initial slice from sparse data and feed it into the CNN input layer. This approach exhibited impressive results in addressing sparse-view reconstruction in parallel beam X-ray CT.

FBPConvNet uses Filtered Back Projection (FBP) and a modified U-net for the reconstruction of sparse-view CT scans. The FBP reconstructs CT images using the available sparse projection data. Due to the incompleteness of the available data, artifacts are present in the reconstruction. To address this issue, the reconstructed CT scans from sparse views undergo training with a modified U-net, using full-view CT scans as the labeled images, aiming to eliminate the artifacts. The modified U-net incorporates two modifications. Firstly, zero padding is applied to ensure that the image size remains constant after each convolution. Secondly, the final layer is substituted with a convolutional layer, reducing the 64 channels to a singular output image. While this approach proved successful for sparse data, it encounters challenges in reconstructing data from a limited scan range. In such cases, artifacts not only encompass pattern-like noise but also involve boundary distortion. The FBP technique struggles to effectively reconstruct distorted boundaries. In contrast, the SART-TV method excels in reconstructing the initial boundary and effectively reducing noise from limited data when compared to FBP.

Building on their work, we leverage the SART-TV method to compute the initial slice and then insert it into the modified U-net input layer, specifically for limited-angle CT reconstruction. The rationale behind this choice lies in the superior performance of the SART-TV method when dealing with incomplete data, resulting in a higher-quality training set compared to the FBP. Considering the importance of training set quality in the modified U-net, this contributes to an enhanced efficacy for limited-angle reconstruction.

Our proposed method comprises the following steps: Firstly, we apply the SART-TV method to limited-angle projection data obtained from the LDCT dataset. Subsequently, we input the slice reconstructed by the SART-TV method into the modified U-net trained to minimize the disparity between the reconstructed image and the labeled image.

3.2 Initial Reconstruction

We utilize the discrete linear model for CT imaging:

$$\boldsymbol{A}\boldsymbol{f} = \boldsymbol{p} \tag{1}$$

where A is the system matrix, p is the vector of measured projection data, and f is the image to be reconstructed.

Our goal is to reconstruct f satisfying the following constraint:

$$\| \boldsymbol{Af} - \boldsymbol{p} \| < \mu \tag{2}$$

where μ is a parameter representing error tolerance. In the noise-free case, μ is ideally zero, while in the noisy case, it is set to a positive value.

Due to the severe ill-posedness of limited-angle tomography, the number of images satisfying the above constraint is not unique. We aim to reconstruct an image that satisfies this constraint and is simultaneously close to the modified U-net reconstruction $\boldsymbol{f}_{prior}$. To achieve this, we choose to initialize the reconstruction $\boldsymbol{f}$ iteratively, starting with $\boldsymbol{f}^{(0)} = \boldsymbol{f}_{prior}$, and solving it as follows:

$$\| \boldsymbol{Af} - \boldsymbol{p} \| < \mu \quad and \quad \boldsymbol{f}^{(0)} = \boldsymbol{f}_{prior} \tag{3}$$

Reconstructed slice $\boldsymbol{f}$ is computed as:

$$\boldsymbol{f}_j^{t+1} = \boldsymbol{f}_j^t + \gamma . \frac{\sum_{\boldsymbol{p}_i \epsilon \boldsymbol{P}_\alpha} \frac{S_\epsilon (\boldsymbol{p}_i - \sum_{k=1}^{M} \boldsymbol{A}_{i,k} . \boldsymbol{f}_k^l)}{\sum_{k=1}^{M} \boldsymbol{A}_{i,k}} \boldsymbol{A}_{i,j}}{\sum_{\boldsymbol{p}_i \epsilon \boldsymbol{P}_\alpha} \boldsymbol{A}_{i,j}}, \tag{4}$$

where i represents the projection ray index of $\boldsymbol{p}$, j denotes the pixel index of $\boldsymbol{f}$, $\boldsymbol{A}_{i,j}$ signifies the element of matrix $\boldsymbol{A}$ at the i-th row and j-th column, M stands for the total number of pixels, t denotes the iteration number, α represents the X-ray source rotation angle, γ is a relaxation factor typically set to 0.7, and $\boldsymbol{P}_\alpha$ denotes the set of projection rays corresponding to the X-ray source rotation angle α and S_ϵ denotes soft-thresholding operator with threshold ϵ.

3.3 Structure-Texture Decomposition

Subsequently, we engage in structure-texture decomposition for the reconstructed slice $\boldsymbol{f}$ to eliminate artifacts. We employ the concept of Relative Variation (RV) [30] to distinguish and extract structure and texture from the reconstructed slice. The expression for the RV of the reconstructed slice is as follows:

$$arg \ \ min \sum_{pix} (\boldsymbol{S}_{pix} - \boldsymbol{f}_{pix})^2 + \lambda . \left(\frac{D_x(pix)}{L_x(pix) + \varepsilon} + \frac{D_y(pix)}{L_y(pix) + \varepsilon} \right) \tag{5}$$

$\mathbf{S}$ represents the resulting structure image.The term $(\boldsymbol{S}_{pix} - \boldsymbol{f}_{pix})^2$ serves to stabilize the input and output, preventing wild deviations. The regularizer, denoted as relative total variation (RTV), accounts for the impact of texture removal in an image. This regularizer is defined as $\left(\frac{D_x(pix)}{L_x(pix)+\varepsilon} + \frac{D_y(pix)}{L_y(pix)+\varepsilon} \right)$, where λ in Equation (5) represents a weighting factor, and ε is a small positive number crucial for preventing division by zero in the element-wise operation. D is the pixel-wise windowed total variation in horizontal and vertical directions, and L denotes the windowed inherent variation that captures the overall spatial

variation. The solution of the loss function aims to make the extracted structures similar to those in the input image, while L and D provide information about the texture part. Textures, being the pattern-like effects, are undesirable in the CT image. Therefore, we will utilize D computed from Equation (5) to remove the pattern-like noise in our proposed RV-SART algorithm.

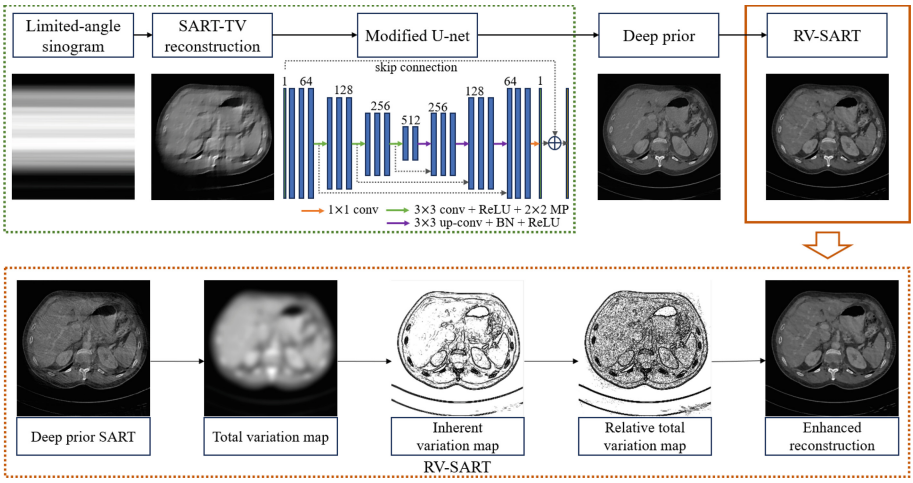


Fig. 2. Schematic of the proposed technique

3.4 Limited-Angle Reconstruction Algorithm

Figure 2 delineates our proposed methodology, providing a comprehensive overview of each section's output. In this figure, the reconstruction process part shows the overall reconstruction process, which includes two main parts: modified U-net with preprocessed input and RV-SART.

The first part shows the CNN model based on the U-net. It is composed of an encoder path and a decoder path. The encoder path consists of numerous 3×3 convolutions, rectified linear units, and 2×2 max pooling operations represented in the green arrow. The decoder path also consists of numerous 3×3 up-convolutions, batch normalizations, and rectified linear units represented in the purple arrow. The skip connection and the concatenation in the black dashed line arrow are available because of the loss of useful information in every convolution and max pooling. In the final layer of the CNN, a 1×1 convolution represented in an orange arrow is used to make the CNN output a single-channel image, which is the final reconstructed image.

RV-SART receives the CNN output and uses it as prior information about unmeasured data during reconstruction for data consistency with the measured data. During this stage, the initially reconstructed slice undergoes decomposition

Algorithm 1 Deep Prior Based RV-SART

```
Parameter initialization
λ = 0.0002, σ = 5, η = 15, k = 4 ;
Prior reconstruction
f_prior = initial reconstruction from modified U-net with preprocessed input
f = f_prior
f_0 = f
Relative Variation SART
for i = 1 : N_d do
    f = f + λA_i (p − A_i f)/(A_i.A_i)
    f_fil = median_filter(f)
 Enforce positivity
    if f_i < 0 then
        f_i = 0
    end if
    S_0 = f
    Structure-Texture decomposition
    solve D_x, D_y, L_x, L_y, S_pix using [33]
    return D_x, D_y, S_pix
    Enhance Reconstructed slice structure
    f = S_pix
    f_en = f − (D_x + D_y)
    p = A f_en
    compute l^2 norm
    N_i = √(p − A f_en)
    until {stopping criteria}
    if i ≠ 1&&N_i > N_{i−1} then
        return f_en
    end if
end for
return f_en
```

into structure-texture components through total variation and inherent variation, as shown in the RV-SART stage, facilitating the extraction of structural information and the suppression of pattern-like artifacts. This comprehensive process ensures the generation of high-quality CT images.

Considering the lower dose used to obtain the measured data introduces inherent noise. To mitigate this, we apply a small-sized filter according to the following expressions:

$$\boldsymbol{f} = \boldsymbol{f}_{prior} + \gamma A_i \frac{\boldsymbol{p} - A_i \boldsymbol{f}}{A_i . A_i} \tag{6}$$

$$\boldsymbol{f}_{fil}(s,t,v) = median\{\boldsymbol{f}(s+l, t+m, v+n) | (l,m,n) \epsilon R\} \tag{7}$$

Subsequent observations reveal artifacts in the reconstructed slice due to the significant amount of incomplete data. These artifacts exhibit a pattern resem-

bling stripes. To address this, we employ a structure-texture decomposition to separate pattern-like artifacts from the structure and enhance the reconstructed slice as expressed in Equation 5:

$$\boldsymbol{f}_{en} = \boldsymbol{f} - (D_x + D_y) \tag{8}$$

Algorithm 1 outlines the Deep Prior Based RV-SART in pseudocode. The process begins with the initialization of parameters in the parameter initialization phase (Line 1). Subsequently, the initial reconstruction is computed using a modified U-net with preprocessed input and established as a prior (Lines 2-3) to address the unmeasured region. The RV-SART phase (Lines 4-21) delineates the key steps involved in ensuring consistency between the unmeasured data and the measured projection data. The initially reconstructed slice, denoted as $\boldsymbol{f}_0$, acts as a temporary variable throughout the reconstruction process. An initial slice is reconstructed using prior information, and a small filter is employed to suppress noise while maintaining data consistency and positivity (Lines 5-10). The initially reconstructed slice is then assigned to the temporary variable $\boldsymbol{S}_0$ for further processing. To mitigate pattern-like artifacts stemming from limited data, structure-texture decomposition is performed on the reconstructed slice (Lines 11-13). The resulting slice, with suppressed pattern-like artifacts, becomes the current reconstructed slice, and its brightness is enhanced by subtracting windowed total variation (Lines 14-15). Subsequently, a forward projection is computed using the enhanced reconstructed slice, and the l^2 norm is calculated (Lines 16-17). The algorithm verifies the convergence criteria (Line 18). Finally, the high-quality CT slice is obtained (Line 22).

4 Experiments

Dataset. We employ the clinical dataset from the 2016 AAPM Low-Dose CT (LDCT) Grand Challenge [25], provided by the Mayo Clinic. We create the projection data within a parallel-beam geometry system using Siddon's ray-driven algorithm [28]. This dataset is already infused with realistic noise. Furthermore, we introduced a 10% Gaussian noise for comparison with state-of-the-art techniques.

Implementation. We employed a training dataset comprising 2 patient records from the AAPM dataset, and for testing, we employed data from a separate patient within the same AAPM dataset. The training process for modified U-net involved using pairs of limited-view SART-TV slices as input and full-view SART-TV slices as label. The limited-view SART-TV slices were computed from three distinct scanning angular ranges: 90, 120, and 150 degrees. In contrast, the full-view SART-TV slices were computed using the full 360-degree angular range. It is crucial to emphasize that this training approach makes the method applicable to real CT reconstructions, where access to an oracle reconstruction is unavailable. Following this, the network's output is fed into the iterative

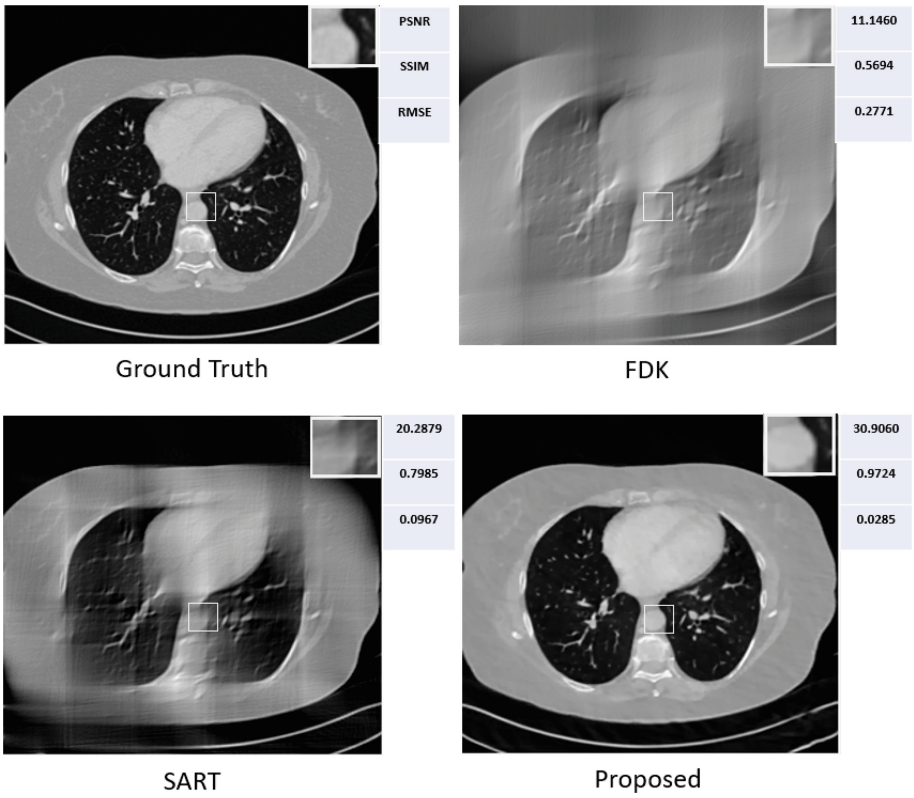


Fig. 3. Limited-angle reconstruction experiment of the AAPM dataset with 90° scanning angular range

model RV-SART as prior information for high-quality reconstruction. This process ensures accurate reconstruction of the structure while effectively suppressing artifacts. This limited-view reconstruction holds significant relevance in human imaging, as a substantial reduction in the number of views, for instance, a specific number of decrease in projections, corresponds to an equivalent reduction in the radiation dose administered to the patient.

Baseline methods. We compare two widely used classical techniques and six state-of-the-art learning-based techniques as our baselines, including traditional methods FDK [11] and SART [2], learning based methods FBPConvNet [21], LRIP-Net [12], SIPID [32], PD-net [1], IFSR-net [9], and SFSR-net [9]. To ensure a fair comparison, we employ parallel geometry for all techniques and apply the 2D network for slice-wise reconstruction.

Evaluation metrics. We assess the reconstructed CT slice using three quantitative metrics, namely peak signal-to-noise ratio (PSNR), structural similarity

(SSIM), and root-mean-square error (RMSE), following prior studies in [4]. Superior reconstruction quality is indicated by higher PSNR/SSIM values and lower RMSE values.

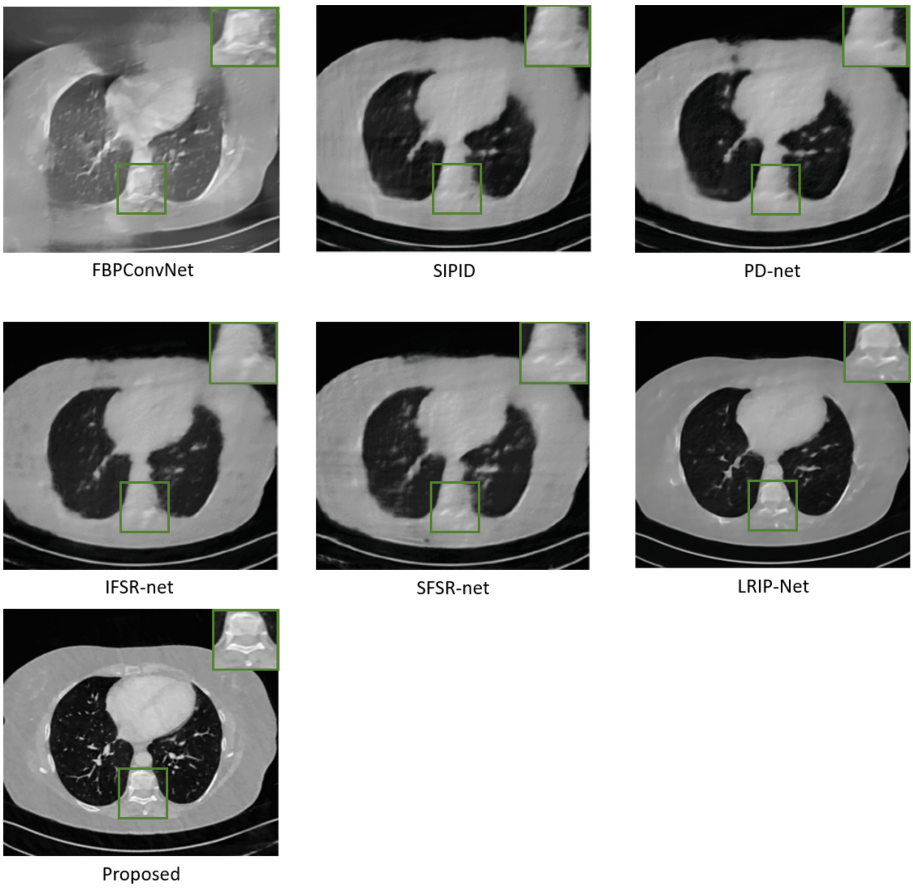


Fig. 4. Comparing Learning-Based Techniques for Limited-Angle Reconstruction Experiment to Preserve Structural Integrity in the AAPM Dataset with a 90° Scanning Angular Range

4.1 Results

Qualitative Evaluation. Figure 3 illustrates both the qualitative and quantitative performance of our proposed technique in comparison to frequently used traditional methods using the AAPM dataset with a 90-degree scanning angular range. In the reconstruction, it is evident that FDK completely failed to reconstruct the structure. SART performs reasonably well in reconstructing the region

Table 1. Evaluating Limited-Angle Data Distorted by 10% Gaussian Noise: A Comparative Analysis based on PSNR, RMSE, and SSIM Metrics

Noise	N_{view}	Method	PSNR	RMSE	SSIM
10%	150	FBPConvNet	21.8293	0.0810	0.7887
		SIPID	29.0276	0.0345	0.9193
		PD-net	29.0084	0.0354	0.9193
		SFSR-net	29.4543	0.0336	0.9199
		IFSR-net	29.6694	0.0328	0.9231
		LRIP-Net	30.8026	0.0288	0.9362
		Proposed	**31.6783**	**0.0261**	**0.9800**
	120	FBPConvNet	20.0065	0.0999	0.7465
		SIPID	26.6271	0.0461	0.8941
		PD-net	26.7667	0.0458	0.8944
		SFSR-net	27.2079	0.0436	0.9034
		IFSR-net	27.2853	0.0432	0.9032
		LRIP-Net	29.1261	0.0349	0.9256
		Proposed	**31.5577**	**0.0264**	**0.9740**
	90	FBPConvNet	18.7582	0.1153	0.7252
		SIPID	23.6216	0.0664	0.8607
		PD-net	23.6473	0.0657	0.8615
		SFSR-net	23.7253	0.0651	0.8591
		IFSR-net	24.2056	0.0616	0.8701
		LRIP-Net	25.9377	0.0457	0.9141
		Proposed	**30.9060**	**0.0285**	**0.9724**

with available data but struggles in areas with unavailable data. In comparison, our proposed technique demonstrates excellent performance in reconstructing fine structures and suppressing artifacts, particularly within the delineated white box. These qualitative results align closely with the quantitative values, as observed from significantly lower values in comparison to our proposed technique.

Figure 4 illustrates the outcomes of the reconstruction achieved by very recent learning-based approaches in the context of 90° limited-angle reconstruction. It is evident that learning-based methods demonstrate successful reconstruction in the missing angular region compared to classical techniques.

To thoroughly assess the preservation of structure and the removal of artifacts, we selected a small yet intricate feature within the image slice, demarcated by a green box. The FBPConvNet falls short in reconstructing even the outer structure of the slice, as it was expressly designed for sparse data and struggles to reconstruct adequately from limited input.

In comparison, SIPID [32], PD-net [1], IFSR-net [9], and SFSR-net [9] demonstrate commendable performance in reconstructing the overall structure. However, they still face challenges in accurately reconstructing small structures, and a significant portion of fine details is lost during the noise suppression process.

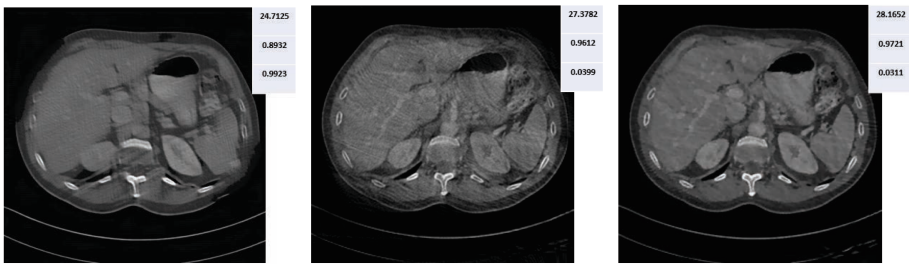
Conversely, LRIP-Net [12] exhibits proficient reconstruction of the majority of structures, yet struggles with the precise reconstruction of small features. In contrast, our proposed technique excels in reconstructing almost all fine structures flawlessly. This success is attributed to our approach of implementing structure-texture decomposition in the initial reconstruction phase, effectively eliminating pattern-like artifacts from the overall structure.

Notably, our proposed technique showcases superior preservation of image details and edges, as emphasized in the magnified region within the green box.

Quantitative Evaluation. Table 1 presents quantitative metrics for comparing recent learning-based techniques with our proposed approach. The evaluation encompasses three distinct limited-angle scanning configurations, each subjected to an additional 10% of Gaussian noise. The table illustrates that as the number of projections increases, the reconstruction quality improves for all techniques. Consistent with visual assessments, LRIP-Net consistently ranks second across all three scanning configurations, while our proposed technique secures the top position.

The quantitative evaluations align with the observed visual quality. Our proposed technique achieves the highest PSNR of 30.9060, SSIM of 0.9724, and the lowest RMSE of 0.0285. These metrics substantiate the ability of our technique to preserve high-quality structures, mitigate noise, and minimize suboptimal label errors compared to ground truth data.

Furthermore, our method outperforms alternatives by attaining the highest PSNR and SSIM, coupled with the lowest RMSE. This dual validation, both quantitative and qualitative, underscores the efficacy of our approach in addressing the challenges associated with limited-angle reconstruction, particularly in terms of maintaining structural integrity.



(a) Unet reconstruction (b) Proposed technique without Relative Variation (c) Proposed technique including Relative Variation

Fig. 5. ablation study

5 Ablation Study

We performed an ablation study to demonstrate how the proposed technique impacts reconstruction quality compared to the base U-Net. Initially, the base U-Net used FBP-reconstructed inputs, which resulted in very poor reconstruction quality. Instead, we propose using SART-TV reconstructed inputs to enhance the base U-Net architecture, thereby improving reconstruction quality.

The base U-Net's reconstructed output contained pattern-like noise, obstructing the visualization of small and thin structures. Figure 5 illustrates the performance of our proposed technique compared to the base U-Net. By providing SART-TV reconstructed inputs to the base U-Net, the reconstruction quality improves compared to using FBP-reconstructed inputs. However, as observed in Figure 5a, the U-Net's reconstruction quality remains poor due to its inability to reconstruct missing projection areas.

In contrast, our proposed technique without relative variation, shown in Figure 5b, successfully reconstructs the CT image in the missing projection areas, although it contains noise-like patterns that obscure the visualization of important organs. Finally, in Figure 5c, we utilize relative variation for structure decomposition, which removes these pattern-like artifacts and produces high-quality CT images with clear visualization of the organs.

Quantitative evaluation also suggests that including relative variation improves PSNR, SSIM, and RMSE to 28.1652, 0.9721, and 0.0311, respectively, compared to the base U-Net and the proposed technique without relative variation.

6 Conclusions

In this study, we introduce the Deep Prior Based RV-SART designed for reconstructing high-quality slices in extremely ill-posed conditions, particularly utilizing a 90-degree scanning angular range. Our approach incorporates prior information from a modified U-net with preprocessed input to feed the RV-SART algorithm about unmeasured data. Subsequently, it extracts structure from the initial reconstructed slice through structure-texture decomposition, facilitating the reconstruction of high-quality CT images while suppressing pattern-like artifacts.

Experiments demonstrate that Deep Prior Based RV-SART excels in high-quality reconstruction from limited-angle data, surpassing the performance of previous state-of-the-art techniques. Notably, prior information is constructed by adapting the efficient and simple architecture of sparse-based modified U-net for specific limited-angle data, providing crucial prior information regarding unmeasured data. This method not only sidesteps recent, computationally

complex deep learning approaches but also enhances accuracy in limited-angle reconstruction.

Acknowledgement. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2023-00250742, 2022R1A4A3018824, RS-2024-00438248, RS-2022-00155885). This research was also supported by the MSIT(Ministry of Science and ICT), Korea under the ITRC(Information Technology Research Center) support program(IITP-2023-RS-2023-00259061) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation).

References

1. Adler, J., Öktem, O.: Learned primal-dual reconstruction. IEEE Trans. Med. Imaging **37**(6), 1322–1332 (2018)
2. Andersen, A.H., Kak, A.C.: Simultaneous algebraic reconstruction technique (sart): a superior implementation of the art algorithm. Ultrason. Imaging **6**(1), 81–94 (1984)
3. Baguer, D.O., Leuschner, J., Schmidt, M.: Computed tomography reconstruction using deep image prior and learned reconstruction methods. Inverse Prob. **36**(9), 094004 (2020)
4. Bappy, D., Kang, D., Lee, J., Baek, H.: Prior-based enhanced asd-pocs for artifact suppression and structural preservation in sparse-view cbct. In: Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing. pp. 496–505 (2024)
5. Boublil, D., Elad, M., Shtok, J., Zibulevsky, M.: Spatially-adaptive reconstruction in computed tomography using neural networks. IEEE Trans. Med. Imaging **34**(7), 1474–1485 (2015)
6. Bubba, T.A., Galinier, M., Lassas, M., Prato, M., Ratti, L., Siltanen, S.: Deep neural networks for inverse problems with pseudodifferential operators: An application to limited-angle tomography (2021)
7. Chen, G.H., Tang, J., Hsieh, J.: Temporal resolution improvement using piccs in mdct cardiac imaging. Medical physics **36**(6Part1), 2130–2135 (2009)
8. Chen, H., Zhang, Y., Chen, Y., Zhang, J., Zhang, W., Sun, H., Lv, Y., Liao, P., Zhou, J., Wang, G.: Learn: Learned experts' assessment-based reconstruction network for sparse-data ct. IEEE Trans. Med. Imaging **37**(6), 1333–1347 (2018)
9. Cheng, W., Wang, Y., Li, H., Duan, Y.: Learned full-sampling reconstruction from incomplete data. IEEE Transactions on Computational Imaging **6**, 945–957 (2020)
10. Ding, Q., Chen, G., Zhang, X., Huang, Q., Ji, H., Gao, H.: Low-dose ct with deep learning regularization via proximal forward-backward splitting. Physics in Medicine & Biology **65**(12), 125009 (2020)
11. Feldkamp, L.A., Davis, L.C., Kress, J.W.: Practical cone-beam algorithm. Josa a **1**(6), 612–619 (1984)

12. Gao, Q., Ding, R., Wang, L., Xue, B., Duan, Y.: Lrip-net: Low-resolution image prior-based network for limited-angle ct reconstruction. IEEE Transactions on Radiation and Plasma Medical Sciences **7**(2), 163–174 (2022)
13. Gao, Y., Liang, Z., Moore, W., Zhang, H., Pomeroy, M.J., Ferretti, J.A., Bilfinger, T.V., Ma, J., Lu, H.: A feasibility study of extracting tissue textures from a previous full-dose ct database as prior knowledge for bayesian reconstruction of current low-dose ct images. IEEE Trans. Med. Imaging **38**(8), 1981–1992 (2019)
14. Gonzales, B., Spronk, D., Cheng, Y., Tucker, A.W., Beckman, M., Zhou, O., Lu, J.: Rectangular fixed-gantry ct prototype: combining cnt x-ray sources and accelerated compressed sensing-based reconstruction. IEEE Access **2**, 971–981 (2014)
15. de Gonzalez, A.B., Darby, S.: Risk of cancer from diagnostic x-rays: estimates for the uk and 14 other countries. The lancet **363**(9406), 345–351 (2004)
16. Guo, Y., Zeng, L., Wang, C., Zhang, L.: Image reconstruction model for the exterior problem of computed tomography based on weighted directional total variation. Appl. Math. Model. **52**, 358–377 (2017)
17. Gupta, H., Jin, K.H., Nguyen, H.Q., McCann, M.T., Unser, M.: Cnn-based projected gradient descent for consistent ct image reconstruction. IEEE Trans. Med. Imaging **37**(6), 1440–1453 (2018)
18. Han, H., Gao, H., Xing, L.: Low-dose 4d cone-beam ct via joint spatiotemporal regularization of tensor framelet and nonlocal total variation. Physics in Medicine & Biology **62**(16), 6408 (2017)
19. Han, Y., Ye, J.C.: Framing u-net via deep convolutional framelets: Application to sparse-view ct. IEEE Trans. Med. Imaging **37**(6), 1418–1429 (2018)
20. Hu, D., Zhang, Y., Liu, J., Luo, S., Chen, Y.: Dior: Deep iterative optimization-based residual-learning for limited-angle ct reconstruction. IEEE Trans. Med. Imaging **41**(7), 1778–1790 (2022)
21. Jin, K.H., McCann, M.T., Froustey, E., Unser, M.: Deep convolutional neural network for inverse problems in imaging. IEEE Trans. Image Process. **26**(9), 4509–4522 (2017)
22. Kang, E., Min, J., Ye, J.C.: A deep convolutional neural network using directional wavelets for low-dose x-ray ct reconstruction. Med. Phys. **44**(10), e360–e375 (2017)
23. Lee, M., Kim, H., Kim, H.J.: Sparse-view ct reconstruction based on multi-level wavelet convolution neural network. Physica Med. **80**, 352–362 (2020)
24. Lin, W.A., Liao, H., Peng, C., Sun, X., Zhang, J., Luo, J., Chellappa, R., Zhou, S.K.: Dudonet: Dual domain network for ct metal artifact reduction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10512–10521 (2019)
25. McCollough, C.H., Bartley, A.C., Carter, R.E., Chen, B., Drees, T.A., Edwards, P., Holmes, D.R., III., Huang, A.E., Khan, F., Leng, S., et al.: Low-dose ct for the detection and classification of metastatic liver lesions: results of the 2016 low dose ct grand challenge. Med. Phys. **44**(10), e339–e352 (2017)
26. Pelt, D.M., Batenburg, K.J.: Fast tomographic reconstruction from limited data using artificial neural networks. IEEE Trans. Image Process. **22**(12), 5238–5251 (2013)
27. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
28. Siddon, R.L.: Fast calculation of the exact radiological path for a three-dimensional ct array. Med. Phys. **12**(2), 252–255 (1985)

29. Xu, J., Zhao, Y., Li, H., Zhang, P.: An image reconstruction model regularized by edge-preserving diffusion and smoothing for limited-angle computed tomography. Inverse Prob. **35**(8), 085004 (2019)
30. Xu, L., Yan, Q., Xia, Y., Jia, J.: Structure extraction from texture via relative total variation. ACM transactions on graphics (TOG) **31**(6), 1–10 (2012)
31. Yu, W., Zeng, L.: Iterative image reconstruction for limited-angle inverse helical cone-beam computed tomography. Scanning **38**(1), 4–13 (2016)
32. Yuan, H., Jia, J., Zhu, Z.: Sipid: A deep learning framework for sinogram interpolation and image denoising in low-dose ct reconstruction. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). pp. 1521–1524. IEEE (2018)
33. Zhang, H., Dong, B., Liu, B.: Jsr-net: A deep network for joint spatial-radon domain ct reconstruction from incomplete data. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3657–3661. IEEE (2019)

CTextureFusion: Advanced Texture Transfer with Multi-head Attention for Improving Lung CT Super Resolution

Mario Amoros, Manuel Curado(✉), and Jose F. Vicent

Department of Computer Science and Artificial Intelligence, University of Alicante, Campus de San Vicente del Raspeig, Ap. Correos 99, 03080 Alicante, Spain
mar77@gcloud.ua.es, {manuel.curado,jvicent}@ua.es

Abstract. Lung diseases cause millions of deaths per year worldwide. For medical diagnosis, different medical images are available, with computed tomography (CT) standing out as effective tools for pathology detection. Recently, the use of artificial intelligence (AI) is an advance in early diagnosis, but the quality of the images in the acquisition greatly limits the result. Therefore, this paper proposes CTextureFusion, a new approach to improve the resolution of lung CT images by combining advanced super-resolution imaging techniques like the existing multi-modal and multi-head attention mechanisms and integrating low-resolution images with reference images and applying specific filters for edge detection and contrast enhancement, our model achieves high-quality, detailed image reconstruction. In experiments, the model demonstrates significant improvements in quantitative and qualitative terms. These results suggest a great potential of the model as digital preprocessing for further diagnostic enhancement.

Keywords: super-resolution · imaging · computed tomography · lung

1 Introduction

The analysis of medical images for the detection of diseases is the subject of numerous studies. The lung is an organ that can suffer from numerous diseases, such as pneumonia, cancer and chronic obstructive pulmonary disease (COPD), among others [9,19]. To reduce the impact of these pathologies, early diagnosis, proper assessment of disease progression, and effective treatment and surgical planning are important.

Among the different possibilities for capturing pathologies, CT scans offer a combination of high spatial resolution, three-dimensional visualization capability, increased clarity of lung lesions, visualization of soft tissues and pulmonary vasculature, or even the detection of details, such as small lung nodules that are not visible on other types of medical images (i.e. standard chest X-rays) [20].

However, in some cases, the standard resolution or low quality of a test acquisition may lead to a failure of detecting subtle features or minor lesions,

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15311, pp. 96–111, 2025.
https://doi.org/10.1007/978-3-031-78195-7_7

such as pulmonary nodules, infiltrative lesions, lung masses, embolisms or fibrosis, where improved visualization would allow early detection and monitoring. In recent years, many super-resolution imaging techniques based on computer vision, and more recently on deep neural networks [1], have emerged to improve the quality and resolution of images of all types, and specifically for medical imaging, serving as preprocessing tasks for further analysis. The advent of deep learning-based methods has rendered other methods obsolete.

2 Related work

Conventionally, super-resolution (SR) algorithms can be classified into three categories: those using interpolation-based methods [26], those based on reconstruction [8] and learning-based methods [13,21].

In recent years, deep learning-based SR models have been explored, from early models with convolutional neural networks [6] to more modern transform-based methods [29,30]. In [7], the authors report on the effectiveness of a CNN-based algorithm for reconstructing high-resolution images from low-resolution images.

The study of super-resolution in medical imaging (MedSR) differs from traditional SR techniques due to the unique characteristics of this type of imaging. In this context, in [5] the authors provide an overview of deep learning-based SR processing methods. In [28], the authors have successfully restored original images using a CNN-based algorithm from low-resolution chest computed tomography (CT) images. This CNN-based study shows promising results of super-resolution methods, however, the strategy has limitations as the CT images are artificially degraded and used as training data for the AI algorithm to restore the degraded images.

Medical image analysis using machine learning is being widely used in the diagnosis of various diseases, including lung diseases [24,27]. Among the state-of-the-art classifiers, used in medical applications, we have deep convolutional neural networks (DCNN) [2,12]. Thus, infection prediction using medical imaging is challenging and combining medical imaging with DL algorithms is a viable option in lung disease detection [15,23].

Deep learning-based SR techniques can improve the resolution and detail of X-ray and CT images, enabling more accurate diagnosis and detection of lung diseases for the development of personalized treatments, and reducing the noise inherent in this type of imaging due to the need not to expose the patient to excessive radiation.

Therefore, this article presents CTextureFusion, a novel model based on the transfer of features from a reference image combined with the use of multi-head attention blocks to improve the quality of lung CTs for the capture of small details to improve the characterization and visualization of possible pathologies with subsequent techniques.

3 Methodology

We propose a novel method, CTextureFusion, as a fusion technique of fundamental super-resolution principles, specifically adapted for CT scans field [4]. Its primary aim is to deduce the super-resolution image (I_SR) using that same image in low resolution, alongside reference images (I_ref), incorporating attention mechanisms. This includes the utilization of the existing multi-head[10] attention blocks and the similarity module to extract features of reference image, enhancing precision and detail in image reconstruction. In this article, tests have been performed increasing the resolution to x4 using images with size 64x64 as low-resolution images and images with 256x256 as target images in high resolution (HR).

The proposed attention mechanism is inspired by the Multi-Head methods implemented by Georgescu [10] for the attention blocks. Although it is inspired by this method, two predefined filters for edge detection and contrast enhancement have been included in the input instead of utilizing the variations of contrast used by them. This decision is grounded on the observation that current methodologies often result in excessively smoothed images with poorly defined edges. Thus, the incorporation of these filters seeks to improve the quality of the inputs of the network.

The methodology leverages one CT as a reference, using 2D slices in the Z dimension as a starting point. Through a similarity mechanism, the most similar slice is identified to maximize the capture of relevant features. Subsequently, these features, adjusted to various sizes, are integrated into the super-resolution network along with the application of both multi-head attention mechanisms. These mechanisms include spatial and channel attention, with a renewed focus on channel attention in successive stages, thus substantially improving the final image quality. The dependence of the methodology on this reference image is relative, as the goal is to achieve an accurate image reconstruction, not to increase the resolution beyond the known resolution.

3.1 Network Architecture

The proposed architecture for this network is based on the synergy of key modules specifically designed to enhance the resolution and quality of images. It features the **Multi-Head Attention Module** (MHCA) [10], which implements a variety of advanced attention mechanisms, addressing both spatial and channel aspects. The goal is to highlight and refine the significant features of the images. Subsequently, the **Texture Transfer Module** [32] is tasked with extracting and applying textural characteristics from reference images, to enrich the target low-resolution images (Fig. 1).

The process is further complemented by the **Reference Image Extraction Module**, which selects the most suitable images based on their similarity to the reference CT scan, thereby facilitating an effective comparison and a significant enhancement of the input images. A module dedicated to extracting features

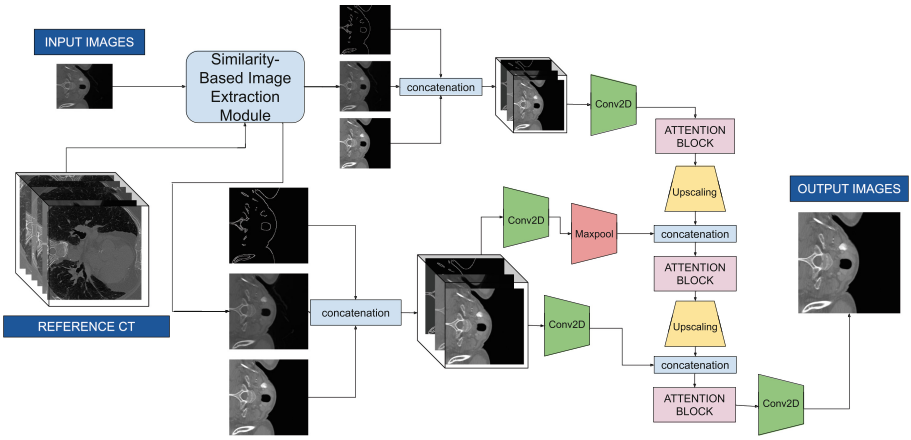


Fig. 1. Schema of Network architecture

from the input images themselves is also integrated, utilizing convolutional layers to extract the fundamental characteristics of the low-resolution images.

To optimize processing in the **Multi-Head Attention Module** (MHCA) [10], an innovative technique has been developed involving the use of input images with three channels: the original image in grayscale, complemented by two versions processed using specific edge detection and contrast enhancement filters. This triple image representation is subjected to concatenation, applied to both the low-resolution images and the reference images. This strategy allows for a deeper and differentiated analysis of the images, contributing to a more accurate and detailed high-resolution reconstruction.

Similarity-Based Image Extraction module The following module operates on the principle of comparing the input image against reference images based on the Peak Signal-to-Noise Ratio (PSNR) metric [14] (Fig. 2).

This metric is used to measure the quality of reconstructed images or videos compared to their originals (see eq. 2). It is commonly used in the field of image compression to evaluate the effectiveness of lossless or lossy compression algorithms. To calculate the PSNR between two images, it is first necessary to determine the Mean Squared Error (MSE) between the original and the reconstructed image. The MSE is calculated by averaging the square of the difference between the pixel values of the two images. Mathematically, the MSE is defined as:

$$MSE = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} (Img_1(i,j) - Img_2(i,j))^2 \tag{1}$$

where m and n are the dimensions of the images, and $Img_1(i,j)$ and $Img_2(i,j)$ are the pixel values at positions (i,j) in the original and reconstructed images, respectively.

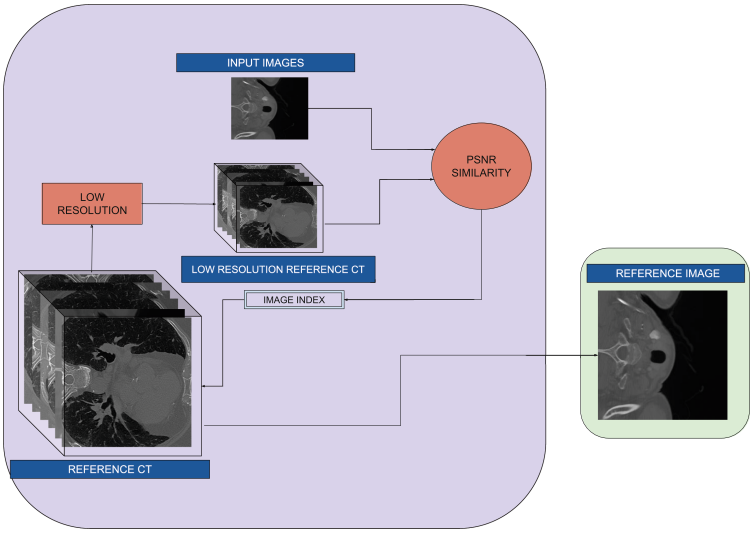


Fig. 2. Schema of similarity module.

Once the MSE is calculated, the PSNR can be calculated as:

$$PSNR = 10 \log_{10} \left(\frac{MAX_I^2}{MSE} \right) \tag{2}$$

Where MAX_I is the maximum possible value of a pixel in the image (in this case, being normalized, the maximum value is 1). In this case, it will be used as a metric of similarity between two images since it is a simple metric that will give us value to obtain the image with greater similarity. One of the options was to use SSIM (Structural Similarity Index Measure) but it is computationally more expensive [14] (Fig. 3).

MHCA module The *Multi-Head Attention* architecture was designed to process input with multiple attention heads, each focused on different spatial and channel features. The class definition can be described as follows:

- The number of attention heads is denoted as H.
- The sizes of the convolution kernels used are represented as a set $K = \{k_1, k_2, ..., k_n\}$.
- Each attention head sequentially applies two convolution operations, first halving the input channels and then restoring them to their original dimension, followed by a nonlinear activation.

The operation of each attention head can be described by the following equations:

$$A_i(x) = \sigma \left(\text{Conv}_{k_i, \text{out}=\frac{\text{in}}{2}} \left(\text{ReLU} \left(\text{Conv}_{k_i, \text{out}=\text{in}}(x) \right) \right) \right), \quad \forall k_i \in K \tag{3}$$

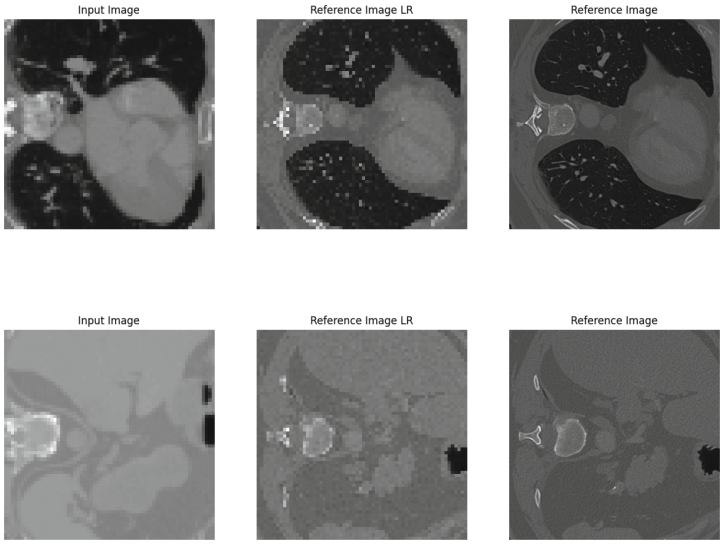


Fig. 3. Image extraction in reference module.

where *sigma* represents the sigmoid function, *ReLU* is the linear-rectified activation function, $Conv_{k_i,out}$ denotes a convolution operation with a kernel of size k_i and a specified number of output channels, and x is the input to the attention head. The convolution operation includes appropriate padding to maintain the spatial dimensions of the input.

The attention maps generated by each head, $A_i(x)$, are then combined by calculating the arithmetic mean:

$$A(x) = \frac{1}{H} \sum_{i=1}^{H} A_i(x) \tag{4}$$

Finally, the combined attention map is used to weight the original input, generating the output of the attention module:

$$Y(x) = x \cdot A(x) \tag{5}$$

This weighted output incorporates differential attention on the input features, allowing the model to focus on relevant aspects of the visual information.

3.2 Evaluation Metrics

In this study, we use PSNR (eq. 2) and SSIM (eq. 6) as key metrics to evaluate our CTextureFusion model. These metrics, widely recognized for their ability to measure the quality and structural similarity of super-resolved images, allow us to objectively compare the performance of our network to the state of the art.

PSNR assesses reconstruction fidelity, while SSIM [14]. measures visual quality relative to the original images, ensuring a complete evaluation of our approach.

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{6}$$

where x and y are the both compared images, μ_x and μ_y are the mean and σ_x^2 and σ_y^2 are the variances of x and y, respectively, σ_{xy} is the covariance between x e y, and $C_1 = (k_1 L)^2$ and $C_2 = (k_2 L)^2$ are two variables to stabilize the division with a small denominator; L is the dynamic range of the pixel values.

4 Experiments

4.1 Datasets

For the experiments, two datasets of lung CTs will be used:

Task06-Lung dataset: is a dataset of lung images consisted of individuals diagnosed with non-small cell lung cancer from Stanford University (located in Palo Alto, CA, USA), which was made publicly accessible through TCIA [3,11,25]. This dataset had been previously used to develop a radiogenomic signature.consists of 96 sets of preoperative thin-section of segmented 3D CT scans. The dataset was divided into two subsets: train with 64 input 3D volumes and test with 32 3D volumes.

The acquisition and reconstruction parameters for the CT scans included a section thickness of less than 1.5 mm, a tube voltage of 120 kVp, automatic tube current modulation with a range of 100-700 mA, a tube rotation speed of 0.5 s, a helical pitch between 0.9 and 1.0, and the use of a sharp reconstruction kernel. An expert thoracic radiologist identified the tumor region by marking it on a representative CT cross-section via software.

Covid dataset: this dataset comprises 20 labeled COVID-19 CT scans [17]. Two radiologists label the left lung, right lung, and infections, which are subsequently verified by an experienced radiologist. The segmentation of the left lung, right lung, and infections is done using COVID-19 CT scans, existing labeled lung CT datasets from other non-COVID-19 lung diseases, and heterogeneous datasets containing both COVID-19 and non-COVID-19 CT scans.

4.2 Training Settings

Training of the SR model was performed using an NVIDIA V100 GPU with 40 GB of memory. CTextureFusion was initialized with a reference CT consisting of 252 slices and a reduced version of the same, along with the scaling factor for resolution increase.

The loss function employed for training was the Mean Squared Error Loss (*MSELoss*), suitable for regression tasks such as super-resolution where the goal is to minimize the difference between predicted and actual pixel values. We have

used the Adam optimizer, and the chosen learning rate is 0.001, with a number of epochs of 100.

Finally, the training set split was 80% training data, 10% validation and 10% test.

4.3 Baselines

We compare our method with four SR models of the-state-of-the-art, which are usually used in the problem of CT lung imaging analysis:

- SRCNN [16] (Super-Resolution Convolutional Neural Network) is an image super-resolution algorithm based on deep convolutional neural networks. It is trained using pairs of low-resolution (LR) and high-resolution (HR) images to learn how to perform the transformation effectively. They are especially useful in CTs for better identification of anatomical structures, increased accuracy of early disease diagnosis due to the level of detail of areas of interest, and evaluation of small lesions.
- EDSR [22] (Enhanced Deep residual networks for single image Super-Resolution) is an image super-resolution algorithm that takes deep residual neural networks (ResNets) and includes more residual layers, reducing the sampling rate and using better filters to make it more computationally efficient. In this way, they enable better spatial resolution in CTs for greater anatomical structural clarity, the detection of minor lesions and relevant structures.
- SRResNet [18] (Super-Resolution Residual Network) is a super-resolution imaging algorithm that takes residual neural networks using residual connections to train deep networks and map directly from low to high resolution using adversarial with perceptual loss in model training. In this way, high-quality images with high detail are obtained, especially recommended for CT imaging, and are perceptually more realistic and have finer details compared to EDSR and SRCNN.
- RCAN [31] (Residual Channel Attention Networks) is a SR algorithm that uses attention mechanisms to highlight important image features, and is deeply useful in problems such as lung CTs.

4.4 Results and Analysis

At first, we compare our model in two datasets. Fig. 4 shows the different predictions of the networks for an image from each dataset. It is noticeable that the two datasets differ in the brightness of the CT scans, and additionally, in the second one, there are also symptoms of the COVID disease, where a certain amount of grey mass is visible within the lungs.

To evaluate our model CTextureFusion[1] a quantitative and qualitative comparison is performed, with the most efficient models in this type of state-of-the-art problems on two lung CT datasets (see Table 1).

[1] Code: https://github.com/Maristoteles02/CTextureFusion.

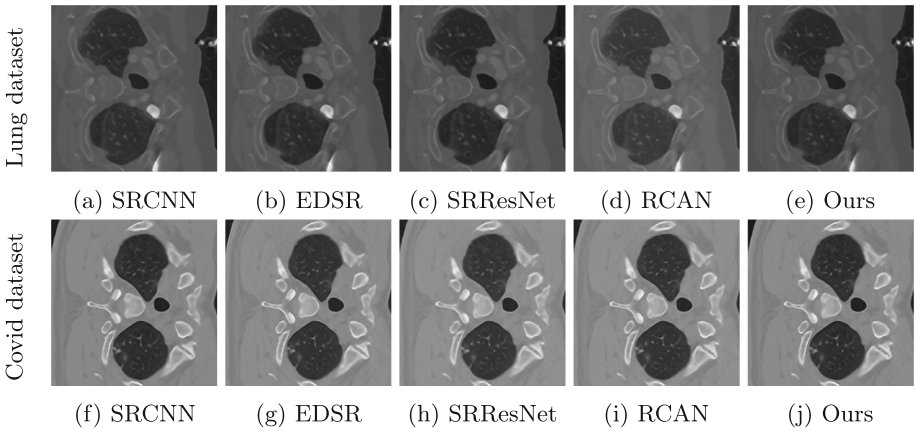


Fig. 4. Models comparison for an image from each dataset.

Table 1. SR models comparison.

Dataset	Model	Avg. SSIM	Avg. PSNR
Lung	SRCNN	0.6989	29.1972
	EDSR	0.7359	30.7072
	SRResNet	0.7298	30.4532
	RCAN	0.7356	30.6315
	CTextureFusion	**0.7399**	**30.7411**
Covid	SRCNN	0.7304	29.4400
	EDSR	0.7548	30.4755
	SRResNet	0.7497	30.3239
	RCAN	0.7557	30.5198
	CTextureFusion	0.7590	30.5552
	CTextureFusion Ref	**0.7616**	**30.5826**

In the reconstruction of these types of images to improve the quality of a low-resolution image, SSIM evaluates the structural similarity between two images from a human perception perspective (textures, edges, details, etc.) and PSNR measures the image quality according to signal and noise. As can be seen in Table 1, our model obtains better results than the state-of-the-art, both in SSIM and PSNR. The second and third best-performing models are EDSR and RCAN, respectively.

These metrics are often used in different superresolution problems, with one or the other being more relevant depending on the scope of the problem. In the case of lung CTs, the SSIM is very important to evaluate the quality of the image obtained by superresolution because many pathologies focus on small details, textures and edges, such as lung nodules, masses or small lesions.

In addition, in a CT scan, which usually lasts longer than 15 minutes, the radiation exposure is greater than a lung X-ray (around seconds). One way to reduce radiation exposure is to lower the amount of radiation, which results in the inclusion of noise in the CT scan. Therefore, the PSNR could be degraded and a balance between the two metrics is important. In Figure 5 it can be seen how our model improves in both metrics in the two datasets to the rest of the compared models, followed by ESDR and RCAN (with disparate performance depending on the dataset).

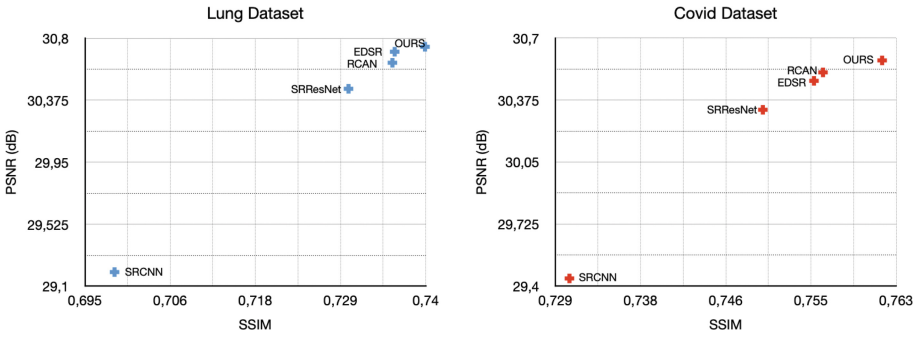


Fig. 5. Evaluation comparison of our model with the state-of-the-art in terms of PSNR and SSIM metrics.

In terms of performance, the inference time is not relevant because the quality of the medical image is more important than its speed. However, comparing our model with the next two in image quality (EDSR and RCAN), the inference time of the model is better (0.012) than the rest, 0.0161 and 0.0273, respectively. Moreover, the number of parameters of our model (1.5 M) is significantly lower than ESDR (3 M) and RCAN (5.4 M).

As for the validation of the visual perception of the model, three regions of interest from the same image of the COVID (GT - Ground Truth) dataset can be observed in Fig. 6.

In the first evaluation (Fig. 6-[a-h]), the region marked in green in (a) is selected, and a 64x64 pixel image is extracted, performing a 4x super-resolution (e). After the application of the different models, it can be observed how our model can define in greater detail the borders of the pulmonary pleura, distinguishing the visceral and parietal pleura. With this, diseases such as pleurodesis or pleuritis could be diagnosed.

In the second evaluation (Fig. 6-[i-p]), an 8x super-resolution (m) is applied, extracting a 32x32 pixel image (i). It can be seen how our model is able to capture with higher definition the pleural space, which are fluid-filled cavities that serve to lubricate and facilitate respiratory movements. The level of detail could help in the diagnosis of diseases such as pleural effusion, pneumothorax or hemothorax.

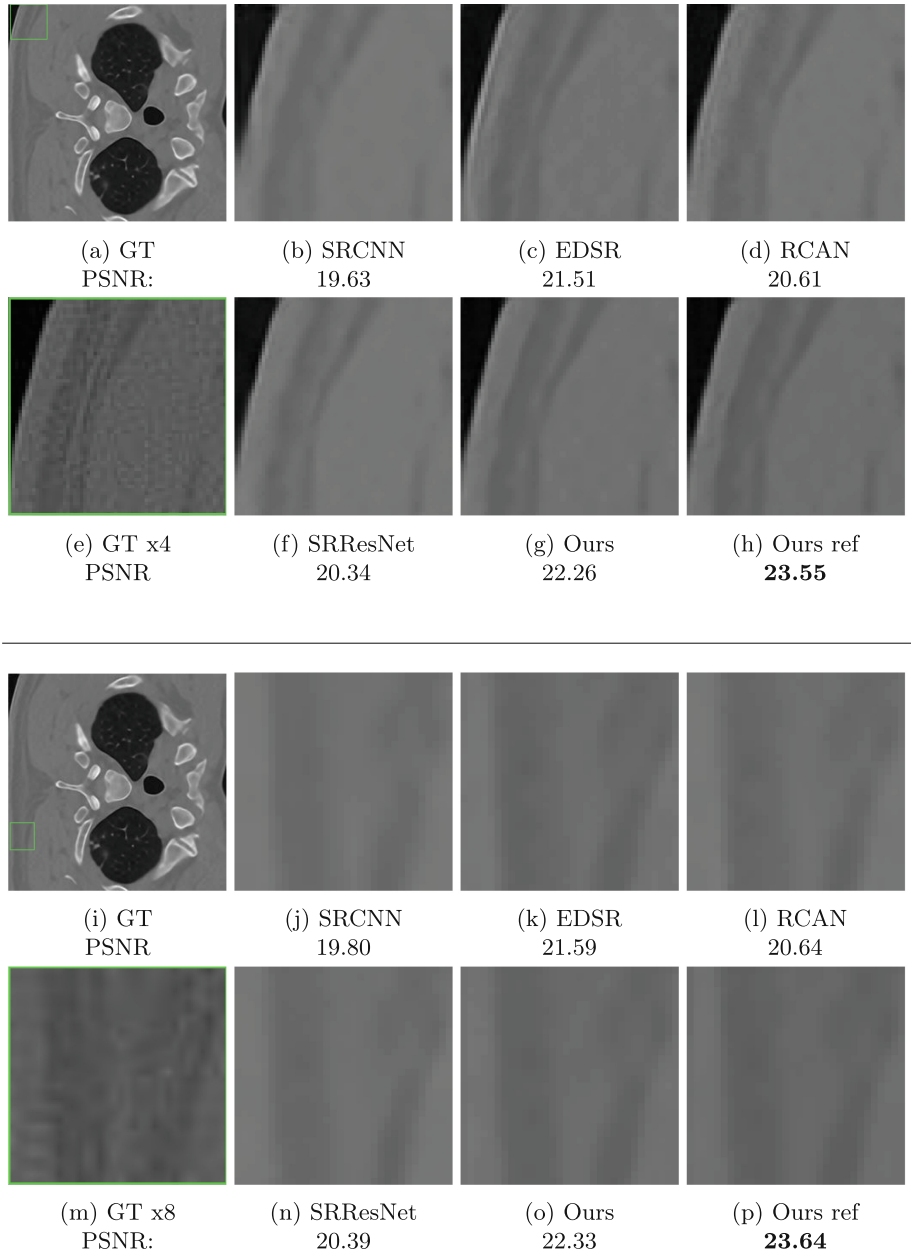
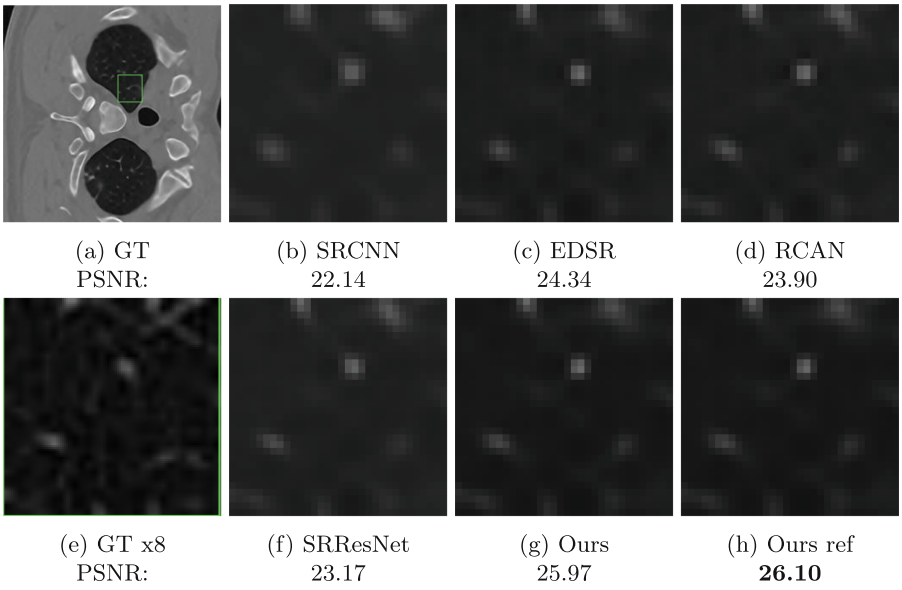


Fig. 6. Quantitative and qualitative comparison of two regions of interest (ROIs) of the pulmonary pleura in a CT scan.



(a) GT
PSNR:

(b) SRCNN
22.14

(c) EDSR
24.34

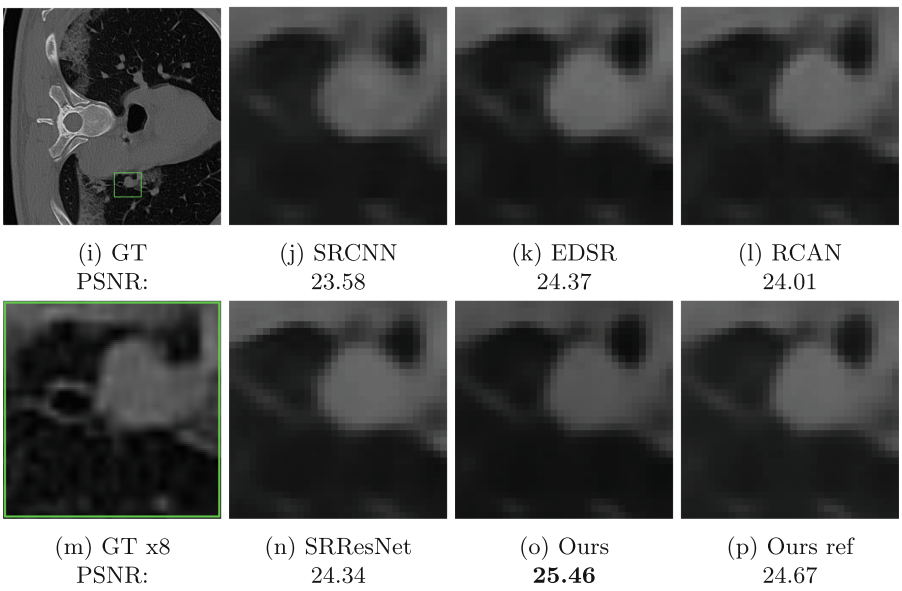
(d) RCAN
23.90

(e) GT x8
PSNR:

(f) SRResNet
23.17

(g) Ours
25.97

(h) Ours ref
26.10



(i) GT
PSNR:

(j) SRCNN
23.58

(k) EDSR
24.37

(l) RCAN
24.01

(m) GT x8
PSNR:

(n) SRResNet
24.34

(o) Ours
25.46

(p) Ours ref
24.67

Fig. 7. Quantitative and qualitative comparison of two regions of interest (ROIs) of the bronchi in a CT scan.

In the third evaluation (Fig. 7-[a-h]), an 8x super-resolution is applied (a), extracting a 32x32 pixel image (e). In it, the focus is on bronchi and bronchioles, obviating the blood vessels and cognitive tissue (links). Our model provides greater sharpness and contrast between these and the rest of the image, binarizing the background and points of interest. In this way, image post-processing techniques could be used to detect nodules or foreign masses in the lung, and to diagnose diseases such as cancer at an early stage.

Finally, in the fourth evaluation (Fig. 7-[i-p]), another 8x SR is performed to analyze the posterior basal segmental bronchus. It is observed how our model reconstructs its structure better than the other models.

In quantitative terms of these images, improvements in PSRN of up to 1.73% (pleura) and 2.05% (bronchi) are observed concerning the second-best model (EDSR).

The results show a slight increase in inference time compared to previous models. This increase is negligible as it is less than 1% compared to the previous network that took the longest (RCAN) [31].

5 Conclusion

In this work, an innovative method has been developed to enhance superresolution in computed tomography (CT) images, using a single reference CT. Through a similarity module, this method can identify those reference CT slices that present greater similarity with each specific slice. This process facilitates obtaining a detailed representation of the expected textures in the input image. The proposed network, with a reduced number of parameters compared to other models, has been applied on lung CT datasets with x4 and x8 scale super-resolution, surpassing previous state-of-the-art results with significant improvements (greater than 2% in some regions of interest), not only in terms of error (quantitative) but also qualitatively, significantly improving the super-resolution of ROIs, such as lung pleura or bronchi, of the images as a preprocessing step in the analysis of medical images, to be able to subsequently apply medical diagnostic tasks in different diseases. All this is achieved without the need to retrain the network or perform fine-tuning for each dataset.

Additionally, it has been demonstrated that the method is capable of generalizing diseases with some ease in the same type of image (chest CT). The aim is to lay the foundation for a general model applicable to all types of medical images, such as bone x-rays. Although there is a slight increase in the number of parameters and inference time compared to previous methods, the inference process remains sufficiently fast not to be a limiting factor.

Furthermore, reference images are used as feature extractors for the super-resolved image, enabling the enhancement of low-resolution images by always having a similar high-resolution image available as a reference. A comparison with non-DL based methods was not included as there is currently no ML based method with a good baseline for comparison with DL based methods. A validation dataset was used to evaluate the overfitting of all methods during the training process.

Acknowledgements. Financial support for this research has been provided under grant CIGE/2023/52 funded by Generalitat Valenciana (Spain). Moreover, the work has been supported by ValgrAI - Valencian Graduate School and Research Network for Artificial Intelligence and the Generalitat Valenciana.

References

1. Albawi, S., Mohammed, T.A., Al-Zawi, S.: Understanding of a convolutional neural network. In: 2017 International Conference on Engineering and Technology (ICET). pp. 1–6 (2017). https://doi.org/10.1109/ICEngTechnol.2017.8308186
2. Alqudah, A.M., Alqudah, A.: Improving machine learning recognition of colorectal cancer using 3d glcm applied to different color spaces. Multimedia Tools Appl. 81(8), 10839–10860 (2022). https://doi.org/10.1007/s11042-022-11946-9
3. Bakr, S., Gevaert, O., Echegaray, S., Ayers, K., Zhou, M., Shafiq, M., Zheng, H., Benson, J.A., Zhang, W., Leung, A.N., et al.: A radiogenomic dataset of non-small cell lung cancer. Scientific data **5**(1), 1–9 (2018)
4. Bhattacharyya, K.: Godfrey newbold hounsfield (1919-2004): The man who revolutionized neuroimaging (10 2016). https://doi.org/10.4103/0972-2327.194414
5. Chen, Z., Pawar, K., Ekanayake, M., Pain, C., Zhong, S., Egan, G.F.: Deep learning for image enhancement and correction in magnetic resonance imaging-state-of-the-art and challenges. J. Digit. Imaging **36**, 204–230 (2022)
6. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014. pp. 184–199. Springer Publishing (2014)
7. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell. **38**(2), 295–307 (2016). https://doi.org/10.1109/TPAMI.2015.2439281
8. Ebner, M., Patel, P.A., Atkinson, D., Caselton, L., Taylor, S., Bainbridge, A., Ourselin, S., Chouhan, M.D., Vercauteren, T.K.M.: Reconstruction-based super-resolution for high-resolution abdominal mri: A preliminary study (2019), https://api.semanticscholar.org/CorpusID:139978758
9. Feng, Y., W.Y.Z.C..M.H.: Artificial intelligence and machine learning in chronic airway diseases: Focus on asthma and chronic obstructive pulmonary disease. International journal of medical sciences **18**(13), 2871-2889 (2021). https://doi.org/10.7150/ijms.58191
10. Georgescu, M.I., Ionescu, R.T., Miron, A.I., Savencu, O., Ristea, N.C., Verga, N., Khan, F.S.: Multimodal multi-head convolutional attention with various kernel sizes for medical image super-resolution. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2195–2205 (2023)
11. Gevaert, O., Xu, J., Hoang, C.D., Leung, A.N., Xu, Y., Quon, A., Rubin, D.L., Napel, S., Plevritis, S.K.: Non-small cell lung cancer: identifying prognostic imaging biomarkers by leveraging public gene expression microarray data-methods and preliminary results. Radiology **264**(2), 387–396 (2012)
12. Hammad, M., ElAffendi, M., Ateya, A.A., Abd El-Latif, A.A.: Efficient brain tumor detection with lightweight end-to-end deep learning model. Cancers **15**(10) (2023)
13. Higaki, T., Nakamura, Y., Tatsugami, F., Nakaura, T., Awai, K.: Improvement of image quality at ct and mri using deep learning. Jpn. J Radiol 37, 73–80 (2019). https://doi.org/doi.org/10.1007/s11604-018-0796-2

14. Horé, A., Ziou, D.: Image quality metrics: Psnr vs. ssim. In: 2010 20th International Conference on Pattern Recognition. pp. 2366–2369 (2010). https://doi.org/10.1109/ICPR.2010.579
15. Hu, S., Gao, Y., Niu, Z., Jiang, Y., Li, L., Xiao, X., Wang, M., Fang, E.F., Menpes-Smith, W., Xia, J., Ye, H., Yang, G.: Weakly supervised deep learning for covid-19 infection detection and classification from ct images. IEEE Access **8**, 118869–118883 (2020). https://doi.org/10.1109/ACCESS.2020.3005510
16. Izonin, I., Tkachenko, R., Peleshko, D., Rak, T., Batyuk, D.: Learning-based image super-resolution using weight coefficients of synaptic connections. In: 2015 Xth International Scientific and Technical Conference" Computer Sciences and Information Technologies"(CSIT). pp. 25–29. IEEE (2015)
17. Jun, M., Cheng, G., Wang Yixin, e.a.: COVID-19 CT Lung and Infection Segmentation Dataset (Apr 2020). https://doi.org/10.5281/zenodo.3757476, https://doi.org/10.5281/zenodo.3757476
18. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4681–4690 (2017)
19. Lee, Y., Kim, E., Chae, K., Lee, C., Jin, G., Kim, S., Choi, J.: Machine learning predicts computed tomography (ct)-based normal regional lung function distribution in asthma and chronic obstructive pulmonary disease (copd). In: B68. A differente point of view: Lung imaging in COPD, pp. A4004–A4004. American Thoracic Society (2023)
20. Li, R., Xiao, C., Huang, Y., Hassan, H., Huang, B.: Deep learning applications in computed tomography images for pulmonary nodule detection and diagnosis: A review. Diagnostics **12**(2) (2022). https://doi.org/10.3390/diagnostics12020298, https://www.mdpi.com/2075-4418/12/2/298
21. Li, Y., Sixou, B., Peyrin, F.: A review of the deep learning methods for medical images super resolution problems (4 2021). https://doi.org/10.1016/j.irbm.2020.08.004
22. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 136–144 (2017)
23. Liu, Q., Leung, C.K., Hu, P.: A two-dimensional sparse matrix profile densenet for covid-19 diagnosis using chest ct images. IEEE Access **8**, 213718–213728 (2020). https://doi.org/10.1109/ACCESS.2020.3040245
24. Loizidou, K., Elia, R., Pitris, C.: Computer-aided breast cancer detection and classification in mammography: A comprehensive review. Comput. Biol. Med. **153**, 106554 (2023). https://doi.org/10.1016/j.compbiomed.2023.106554
25. Napel, S.: Nsclc radiogenomics: initial stanford study of 26 cases. The Cancer Imaging Archive. (2014)
26. Ru Xiang, Hui Yang, Z.Y.A.M.M.T.X.X., Wu, T.: Super-resolution reconstruction of gosat co2 products using bicubic interpolation. Geocarto International **37**(27), 15187–15211 (2022). https://doi.org/10.1080/10106049.2022.2096699
27. Talukder, M.A., Islam, M.M., Uddin, M.A., Akhter, A., Pramanik, M.A.J., Aryal, S., Almoyad, M.A.A., Hasan, K.F., Moni, M.A.: An efficient deep learning model to categorize brain tumor using reconstruction and fine-tuning. Expert Syst. Appl. **230**, 120534 (2023). https://doi.org/10.1016/j.eswa.2023.120534
28. Umehara, K., O.J..I.T.: Application of super-resolution convolutional neural network for enhancing image resolution in chest ct. Journal of Digital Imaging **31**, 441-450 (2018)

29. Wang, Z., Chen, J., Hoi, S.C.H.: Deep learning for image super-resolution: A survey. IEEE Trans. Pattern Anal. Mach. Intell. **43**(10), 3365–3387 (2021). https://doi.org/10.1109/TPAMI.2020.2982166
30. Yang, F., Yang, H., Fu, J., Lu, H., Guo, B.: Learning texture transformer network for image super-resolution. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5790–5799 (2020). https://doi.org/10.1109/CVPR42600.2020.00583
31. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European conference on computer vision (ECCV). pp. 286–301 (2018)
32. Zhang, Z., Wang, Z., Lin, Z., Qi, H.: Image super-resolution by neural texture transfer (3 2019), http://arxiv.org/abs/1903.00834

Cross-Modality Medical Image Registration with Local-Global Spatial Correlation

Souraja Kundu[1(✉)], Yuji Iwahori[2], M. K. Bhuyan[1], Manish Bhatt[1], Boonserm Kijsirikul[3], Aili Wang[4], Akira Ouchi[5], and Yasuhiro Shimizu[5]

[1] Indian Institute of Technology Guwahati, Guwahati 781039, Assam, India
{k.souraja,mkb,manishb}@iitg.ac.in
[2] Chubu University, Kasugai 487-8501, Japan
iwahori@isc.chubu.ac.jp
[3] Chulalongkorn University, Bangkok 10330, Thailand
Boonserm.K@chula.ac.th
[4] Harbin University of Science and Technology, Harbin 150080, China
aili925@hrbust.edu.cn
[5] Aichi Cancer Center Hospital, Nagoya 464-8681, Japan
{akira.ouc,yshimizu}@aichi-cc.jp

Abstract. In this paper, a translation-based cross-modality deformable medical image registration model is proposed. It focuses on preserving spatial correlation among local and global features of both modalities. This model uses a discriminator-free StyleGAN2 as the translation network and a U-Net-inspired architecture as the registration network to generate a deformation field that will warp the moving image to a fixed image using a spatial transformer network (STN). This registration network includes a CNN-based local and a BGRU-based global feature extraction module, a transformer-based local-global spatial correlation module, and a novel super-resolution loss function to register finer-level lymph node-like structures properly. The proposed model is evaluated on two pelvis datasets for MRI to CT registration. Experiments show a 36.7% increase in training speed, a 5.40% increase in structural similarity index, a 29.85% increase in normalized cross-correlation coefficient, and a 2.19% decrease in mean-squared error for cross-modality image registration compared to state-of-the-art translation-based registration models. This registration model has broader applications in multimodality image segmentation, lymph node classification, etc.

Keywords: Biomedical Imaging · Visual Transformer · Bidirectional Gated Recurrent Unit (BGRU)

1 Introduction

Image registration is a necessary pre-processing step for cross-modality image translation, segmentation, or classification tasks. Due to inherent tissue structure

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15311, pp. 112–126, 2025.
https://doi.org/10.1007/978-3-031-78195-7_8

and intensity differences in different imaging modalities, only affine transformation is not sufficient because this will just rotate, enlarge or shift the objects [1], rather non-linear deformable registration is required [2]. With multiple deformation fields appearing to align the images correctly, cross-modality image registration is an ill-posed problem due to a solution's nonuniqueness. For example, CT scans are known for their cost efficiency and widespread use, however, they often lack the necessary contrast for precise lymph node detection. On the other hand, MRIs stand out for their superior soft tissue contrast and the ability to provide detailed information on anatomy, function, and metabolism, making them more suitable for metastasis detection. Still, the high cost and lower accessibility of MRIs present a challenge. Therefore, cross-modality image registration can bridge the gap between CT and MRI imaging.

In recent years, cross-modality image registration has moved from using simple CNN models to more complex deep learning methods, like GANs and CycleGAN, which introduced the idea of learning without matched pairs. However, these learning techniques sometimes fail to perfectly align one modality's images with the other. This problem has led researchers to try learning with datasets with pairs of matching images to get more accurate results. The creation of synthetic images in one modality can augment the image datasets without additional scans and can facilitate cross-modality image registration successfully. Traditional Voxelmorph-based [3] networks use a transformer or a U-Net [4] registration network for deformation mapping across modalities, followed by STNs to learn spatial transformations. However, only Convolution Neural Network (CNN)-based registration networks can align local features well. Still, the global anatomical features may get dislocated, causing problems in identifying the location of lymph nodes or tumor-like local features amid the shifted anatomies in this registered image. On the other hand, only transformer-based registration networks can neglect to incorporate the local details at the correct position in the registered image and hence segmenting small structures like lymph nodes can fail.

In this context, our work focuses on integrating the local details within the global anatomies maintaining proper spatial correlation with the help of an innovative feature flow pipeline. Our contributions are summarized as:

1) We propose a novel deformable image registration network with a U-Net-inspired architecture. A convolutional encoder first extracts high-level local features. Next, global features are extracted by BGRU, fully connected layer, and squash layer. Then, positional correlations of global and local features are obtained with a visual transformer unit. Finally, features are reconstructed by a convolutional decoder.
2) We designed a new super-resolution loss function for registration to accurately retain lymph node positions and other spatial information in MRIs.
3) We achieved a 36.7% increase in training speed than traditional translation-based registration models by using dropout layers and StyleGAN2 generator instead of the more widely used Residual network generator.

Section 2 explores related work. Section 3 presents the registration algorithm, model architecture, datasets, comparison models, and modeling details. Section 4 describes the validation and ablation experiments. Section 5 reports statistical results followed by a conclusion in section 6.

2 Related Works

Image registration can be done by cross-modality image-to-image translation networks [5], [6]. However, choosing suitable metrics for comparing deformed and target images after registration across different modalities is challenging. Some commonly used metrics include mutual information [7] and local self-similarity [8].

Supervised or semi-supervised learning methods are common in multimodality image registration [9], [10]. However, they rely on having reference registered target images, which can be difficult to obtain and may struggle with structural variations. Unsupervised learning methods offer more flexibility and scalability by avoiding these limitations. Approaches like normalized mutual information (NMI) or modality-independent neighborhood descriptor (MIND) [11], [12] are used in unsupervised methods. However, NMI is better suited for global alignment, while MIND may struggle with severe deformations and achieving global alignment.

Unsupervised techniques for CT–MR deformable image registration have utilized CNNs, such as CycleGAN [13]. Zhao et al. [14] used a CNN to extract features and feed them to log-demons. Some studies employed U-Net [15], [16], but they often misalign global structures.

Recent unsupervised methods integrate spatial transform networks (STN) [17], like VoxelMorph-based networks [18], [19], [20]. However, CNN-based VoxelMorph struggles with anatomical variations, while transformer-based approaches may overlook local details like lymph nodes.

Our model tackles these issues by utilizing local and global feature extractors for fine-to-coarse feature extraction. A transformer in the local–global correlation module processes semantic classes, preserving critical features' positional relationships. Unlike KeyMorph [21], which gives good performance by relying on considerably large data and pre-trained models, our approach doesn't require them. Additionally, geometry is maintained by transforming multimodal registration into a monomodal task using simultaneous translation and spatial registration methods [22], [23], [24].

3 Proposed Methodology

3.1 Registration process

The translation-based registration algorithm is followed as mentioned in [25]. Hereafter, X denotes the source and Y denotes the target image sets. Concatenated (x_{real}, y_{real}) input data is fed to a registration network to learn a deformation field Φ to align Y to X. Simultaneously, x_{real} is fed to a cross-modality translation network to generate y_{gen}, and y_{real} is fed to a monomodality translation

network to generate y_{fake}. The super-resolution loss function is used to minimize the difference between y_{real} and registered source image $\Phi(x_{real})$. A structural loss function is used to minimize the structural dissimilarities between $\Phi(y_{gen})$ and y_{fake}. A precision loss is used to minimize the difference in minuscule features between $\Phi(y_{gen})$ and y_{real}. The translation network uses discriminator-free GAN with only the generator followed by a feature network. Since learning-based nonlinear image registration can result in nonunique and noninvertible solutions, the mapping from the deformed MRI to the original may be imprecise:

$$\phi_{MRI}(\phi_{CT}(MRI)) \approx MRI \tag{1}$$

3.2 Model Architecture

The translation network is the generator of a 6-block StyleGAN2 with a symmetric encoder and decoder. To reduce the introduction of non-physiological features incorrect for the moving image modality in the registered image, the discriminator is removed. The registration network consists of a local feature extractor module, a global feature extractor module, a local-global spatial correlation module, and a feature reconstruction module as shown in Fig. 1:

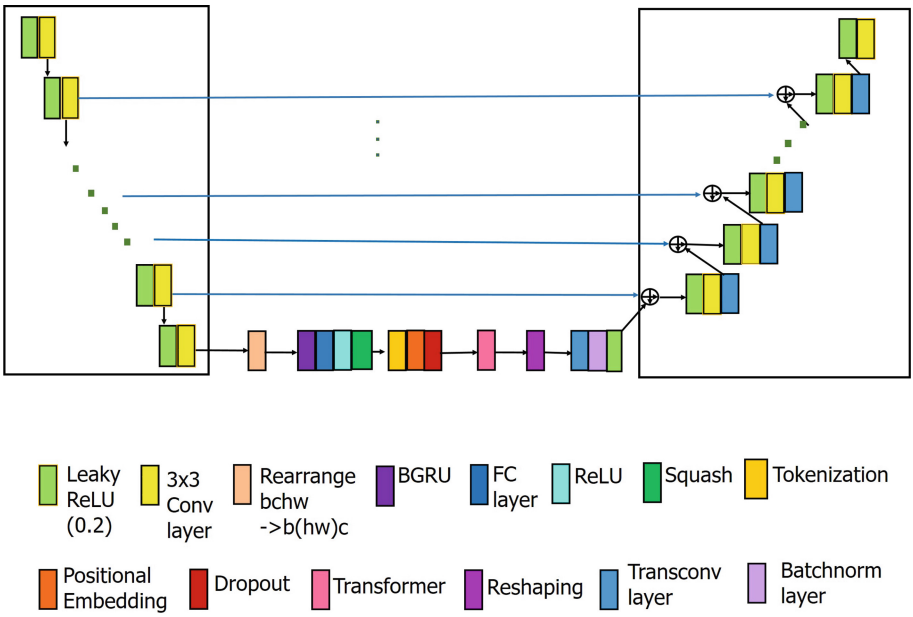


Fig. 1. The registration network architecture

Local Feature Extractor Module: CNNs are good at extracting local features like tissues, small organs, or lymph node-level details. So, concatenated moving

and fixed images of size $512 \times 512 \times 6$ are fed to 8 layers of 3×3 convolutional layer, and 0.2 negative slope Leaky-Relu stacks. This module extracts the local feature details and outputs a $8 \times 8 \times 64$ sized feature tensor enriched with local features along with their corresponding positional information [26].

Global Feature Extractor Module: Though this local feature tensor contains all the necessary fine details, it cannot give an idea of the larger organs, anatomies, or global structure. Hence, the global feature extractor module. This module consists of a BGRU, a fully connected layer, and a squash layer [27]. The resulting output tensor from the convolutional encoder, with dimensions $8 \times 8 \times 64$, comprises 64 feature representations. Each feature vector is of size $1 \times 1 \times 64$, denoted by $f_1, f_2, ..., f_{64}$, where $f \in \mathbf{R}^{64}$. These vectors are treated as time series data in the context of text processing and are fed to the BGRU layer. The output sequences of BGRU are then passed through a fully connected layer to adjust weights and a ReLU activation function:

$$G = ReLU(H \cdot W_{FC}) \tag{2}$$

where, $H = \{h_{1,BGRU}, h_{2,BGRU}, \ldots, h_{128,BGRU}\}$ are the concatenated hidden state vectors of forward and backward direction, and $W_{FC} \in 128 \times 64$ are the FC layer's weights.

Next, the concatenated tensor undergoes normalization through a squash function:

$$F = \frac{||G||^2}{1 + ||G||^2} \cdot \frac{G}{||G||} \tag{3}$$

The resulting feature vectors are then fed to a local-global spatial correlator.

Local-Global Spatial Correlation Module: A recurrent tokenizer splits the input tensor into manageable tokens representing only a handful of semantic concepts, thus reducing computational complexity, which are then combined with positional embeddings and a dropout layer for enhanced stability. Next, a transformer containing a sequence of attention and Multi-Layer Perceptron (MLP) layers identifies positional relationships among local and global features. This contrasts with the approach taken in [28], where the transformer focused on modeling semantic class-token connections.

Feature Reconstruction Module: The feature maps undergo decoding by stacks of convolutional layers, LeakyReLU activations, and trans-convolutional layers resembling the up-arm structure in a U-Net from a size of $8 \times 8 \times 64$ to $512 \times 512 \times 16$. Pixel-level features from the encoder feature maps are merged using concatenation, helping to restore detailed pixel-level information lost during tokenization.

Feature Network: In the StyleGAN2 setup, a feature network comes after the encoder. It improves patch sampling by processing encoded source and target features. This network consists of two sets of linear layers with GELU activation functions and dropout layers for regularization. Additionally, a novel self-attention-fused feature network enhances patch sampling from the hidden feature stack.

The modified MLP layer within the feature network can be represented as:

$$H'(z) = Dropout(GELU(Linear(z)))$$

Here, z denotes the input features, and the linear transformation is followed by GELU activation and dropout for regularization.

The translation network utilizes noise contrastive estimation (NCE) loss instead of a discriminator to maintain structural similarity between generated and target images. This NCE loss employs patch samples from the feature network.

3.3 Loss Functions

To increase registration accuracy, the following loss functions were used in the registration objective function:

Super-resolution Loss: A novel loss function called Super-Resolution Loss (SR) is proposed. This loss uses a simple super-resolution network [29] on cropped central regions of registered and real target domain images, where the probability of finding a lymph node is maximal. This cropping reduces computational complexity. The loss is calculated as the L2 norm of the super-resolved features of the real target image and the registered source image. Mathematically,

$$L_{SR} = \sqrt{||\psi_{real} - \psi_{registered}||_2} \tag{4}$$

Here, ψ_{real} and $\psi_{registered}$ represent the super-resolution features of the real target image y_{real} and registered source image $\Phi(x_{real})$, respectively.

Structural Loss: The L_2 losses between the generated target domain image y_{fake} from y_{real} and the warped version of the generated target domain image $\Phi(y_{gen})$ from x_{real} is calculated to maintain texture and structural similarity post-registration. This loss focuses on retaining structural features, so this term is called structural loss,

$$L_{struct} = ||\Phi(y_{gen}) - y_{fake}||_2 \tag{5}$$

Precision Preservation Loss: L_2 loss is also incorporated between $\Phi(y_{gen})$ and the real target image y_{real} to refine the deformation field. This loss term is called precision preservation loss,

$$L_{prec} = ||\Phi(y_{gen}) - y_{real}||_2 \tag{6}$$

Noise Contrastive Estimation (NCE) Loss: The registration objective function also includes NCE loss L_{NCE}, maximizing mutual information between the real target y_{real} and registered source images $\Phi(x_{real})$. The NCE loss maintains patch-wise structural similarity between the source and target domains.

The complete registration loss function is expressed as:

$$L_{reg} = \lambda_{struct} L_{struct} + \lambda_{prec} L_{prec} + \lambda_{NCE} L_{NCE}(\Phi(x_{\text{real}}), y_{\text{real}}) + \lambda_{SR} L_{SR} \tag{7}$$

where λ_{struct}, λ_{prec}, λ_{NCE}, and λ_{SR} represent the weights for the respective loss terms. The overall generator loss comprises this registration loss, GAN loss, NCE loss as reconstruction loss, and L_2 loss as smoothing loss.

3.4 Datasets

Dataset 1 - Aichi Cancer Centre Dataset: The dataset consists of Contrast-Enhanced CT and T2-weighted axial (4mm) MRI scans from 52 patients, approved by the institutional review boards of the Aichi Cancer Center Hospital (#H301076). CT-MRI pairing was done using RadiAnt DICOM Viewer. Before training and testing, background annotations were removed from DICOM slices to improve image quality. The dataset was split into 60 training, 599 testing, and 13 validation image pairs. The small training size allows for examining model performance with limited data, while the large testing dataset evaluates the model's generalization ability.

Dataset 2 - SynthRAD 2023 Grand Challenge Dataset: This dataset is a collaboration between three Dutch university medical centers: UMC Utrecht, UMC Groningen, and Radboud Nijmegen. It comprises paired CT and MRI pelvic data from 90 patients, acquired using standard imaging protocols. Both T1 and T2 weighted sequences are included to demonstrate the method's applicability across different imaging types. The dataset consists of 60 image pairs for training, 10 for validation, and 599 for testing.

3.5 Baseline Models

NeMAR Affine [22]***:*** NeMAR Affine is a competitive state-of-the-art model (published in 2020) that performs multi-modal registration using two networks - a spatial transformation network and a translation network. This is an unsupervised, multimodality image alignment model using CycleGAN [30] or Pix2PixGAN [31] for geometry-preserving translation. It utilizes an STN consisting of a convolutional network and a re-sampler layer in the registration network. Training includes L1 loss for reconstruction and conditional GAN loss for adversarial training. The affine STN model employs linear transformations for spatial alignment.

NeMAR U-Net [22]***:*** This variant of the above model employs a U-Net STN for nonlinear transformations. While affine transformations align global shifts, U-Net STNs align both local and global features.

DFMIR [25]***:*** Discriminator Free Medical Image Registration (DFMIR) model (published in 2022) surpassed several recent state-of-the-art multi-modal registration methods and some other well-established methods, such as, Synthesis-by-Registration [32], a VoxelMorph architecture with similarity metric MIND [3], CycleGAN [30] combined with VoxelMorph registration network using mono-modal similarity metric NCC etc. DFMIR uses an unsupervised joint framework with end-to-end optimization. It is composed of a discriminator-free Residual network for multimodality image translation and a U-Net for registration. It uses NCE loss and L1 losses in the objective function. For testing, DFMIR only uses the registration network.

3.6 Modelling Details

All models used recommended loss functions during training. Various learning rates were experimented with from 10^{-5} to 10^{-3} and epochs from 5 to 300, selecting the best combination via cross-validation. Equal weights were assigned to L_{struct} and L_{prec} to prioritize both precision and structure preservation, so $\lambda_{struct} = \lambda_{prec} = 1$. L_{NCE} was considered less important to reduce exact patch-wise replication and given a weighting parameter $\lambda_{NCE} = 0.25$. Our super-resolution loss had a weighting parameter $\lambda_{SR} = 1$. The Adam optimizer was employed with $\beta_1 = 0.5$ and $\beta_2 = 0.999$ and a constant learning rate of 2×10^{-4} for 150 epochs, followed by linear decay to 0 for the next 150 epochs. There are 11.378M parameters in the Generator network, 0.560M parameters in the feature network and 152.721M parameters in the registration network, resulting in average training time of 19 second per epoch for 60 training images. Implementation was done using PyTorch on an Nvidia RTX A6000 GPU. Registration quality was evaluated using the Structural Similarity Index Metric (SSIM), Mean Square Error (MSE), and Normalized Cross-Correlation Coefficient (NCC) on a separate test set.

4 Experiments

4.1 Validating Model Architecture Choices

Feature Network: Initially, a basic two-layer MLP patch-sampling network was used. Later, a self-attention layer was added for assessing patch contributions based on content and relationships. GELU (Gaussian Error Linear Unit) was then included for nonlinear mapping and dropout layers were added after each MLP layer.

Generator Network: Residual networks with 4, 6, 8, and 9 blocks, and StyleGAN2 with 2, 3, and 6 blocks were experimented with. By choosing a 6-block StyleGAN2 generator, complexity and training time were reduced. Using StyleGAN2 reduced all training losses and shortened training time by about 27%. The beneficial impact of noise injection on training StyleGAN2 was also analyzed.

Registration Network: Initially, only a transformer was explored, but a U-Net-like encoder-decoder architecture was found necessary for optimal performance. The contributions of BGRU, fully connected layers, ReLU, and Squash layers were assessed by selectively removing them. We experimented with passing hidden states in BGRU vertically but found it degraded performance due to weaker semantic connections in the vertical direction of the feature space.

Loss Functions: A structural similarity index loss was integrated into the registration network but it was found that SSIM is unsuitable even after varying weighting parameters. The super-resolution loss showed improvement in metrics. L_1 and L_2 losses were tried separately for both L_{struct} and L_{prec} and it was found that L_2 loss gives better performance.

4.2 Ablation Experiments

Besides experimenting with different choices of architecture and hyperparameters for each module in the registration network, the relevance of each module was also validated through ablation across both datasets. Initially, a model was tested with only the transformer between the U-Net encoder and decoder, thus removing the global feature extractor module. Next, the impacts of removing the local feature extractor, transformer, BGRU layer, GELU, self-attention, and dropout layer containing feature network and super-resolution loss were evaluated. Results emphasized the importance of the transformer, with all modules contributing incrementally to registration performance.

Additionally, different configurations for the generator network were experimented with and the number of style blocks in StyleGAN2 was varied. Results demonstrated the effectiveness of our generator and transformer setup. Variant models were experimented with by adjusting transformer initialization, encoder-decoder layers of StyleGAN2, and dropout layers at various stages to understand each element's contribution. Different combinations of loss-term weights were evaluated, finalizing our choice of loss function weights.

4.3 Model Comparison Experiments

Our cross-modality image registration network for MRI to CT registration was assessed and it was compared with other translation-based registration models like DFMIR, NeMAR affine, and NeMAR U-net. Both the dataset 1 and 2 were used for the experiments. For fairness in comparison, the same Adam optimizer was employed with $\beta_1 = 0.5$ and $\beta_2 = 0.999$ and a constant learning rate of 2×10^{-4} for 150 epochs, followed by linear decay to 0 for the next 150 epochs.

5 Results

5.1 Model Comparison Results

DFMIR was the best-performing baseline model. On the Aichi Cancer Center pelvis dataset, our model outperformed even DFMIR with a 2.19% reduction

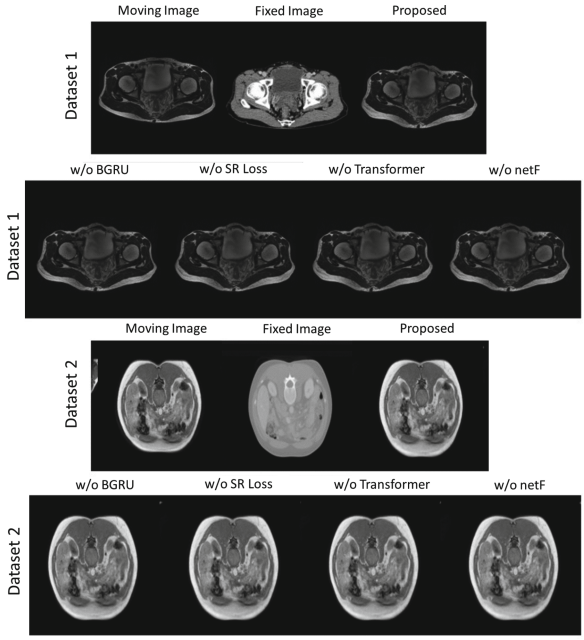


Fig. 2. Ablation study results on the two datasets for MRI to CT registration

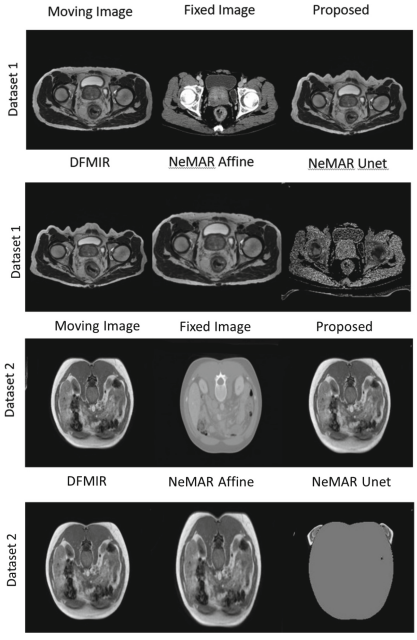


Fig. 3. Model comparison results on the two datasets for MRI to CT registration

in MSE loss, 5.40% improvement in SSIM, and a significant 29.85% increase in NCC. Similar results were observed on the SynthRAD challenge dataset, despite texture differences between CT and MRI. Here our model showed an average 0.54% decrement in MSE, 2.33% improvement in SSIM, and 3.44% increase in NCC compared to DFMIR. Our model reduced training time by 36.7% on average compared to DFMIR. The dropout layer contributed to a 9.7% improvement, while the StyleGAN2 generator accounted for a 27% decrease.

The results are given as mean ± standard deviation across 599 test image pairs in Table 1. Though the metric values for the NeMAR U-Net model seem superior, observing the corresponding registered images shows that it is struggling to generalize with limited data or overfitting in contrast to our model's robustness with only 60 training images. Since metric values are not fully reliable for model evaluation, the registered images are also shown for visual evaluation in Fig. 3.

Table 1. Cross-modality Image Registration Model Comparison

Dataset 1	**Aichi Cancer Centre Dataset**		
Models	**SSIM (↑)**	**MSE (↓)**	**NCC (↑)**
Proposed	**0.5517 ± 0.0541**	**44.62 ± 4.62**	**0.6929 ± 0.0583**
NeMAR Affine	0.3659 ± 0.0587	62.75 ± 7.18	0.3541 ± 0.0971
NeMAR U-Net	0.6416 ± 0.0572	39.78 ± 4.85	0.7739 ± 0.0594
DFMIR	0.5234 ± 0.0527	45.62 ± 4.59	0.5336 ± 0.0687
Dataset 2	**SynthRAD Challenge Dataset**		
Models	**SSIM (↑)**	**MSE (↓)**	**NCC (↑)**
Proposed	**0.4603 ± 0.0306**	**71.88 ± 6.20**	**0.8143 ± 0.0524**
NeMAR Affine	0.4576 ± 0.0306	71.99 ± 6.24	0.8091 ± 0.0568
NeMAR U-Net	0.4259 ± 0.0373	67.72 ± 7.48	0.8296 ± 0.0791
DFMIR	0.4498 ± 0.0391	72.27 ± 6.26	0.7872 ± 0.0886

5.2 Ablation Results

Key modules were systematically deactivated, like the global feature extractor block, local-global feature extraction module, and self-attention containing feature network and super-resolution loss function to assess their impact.

Across all datasets and variations, our model consistently outperformed others on SSIM, MSE, and NCC metrics. Ablation results are shown in Table 2 and Fig. 2. 'Without netF' refers to the model without GELU, dropout, and self-attention layers in the feature network.

Table 2. Cross-modality Image Registration Ablation Results

Dataset 1	**Aichi Cancer Centre Dataset**		
Models	**SSIM (↑)**	**MSE (↓)**	**NCC (↑)**
Proposed	**0.5517 ± 0.0541**	**44.62 ± 4.62**	0.6929 ± 0.0583
w/o BGRU	0.5463 ± 0.0544	45.14 ± 4.81	**0.7029 ± 0.0522**
w/o Transformer	0.5450 ± 0.0535	44.37 ± 4.66	0.6929 ± 0.0514
w/o netF	0.5461 ± 0.0543	44.41 ± 4.46	0.6951 ± 0.0511
w/o L_{SR}	0.5505 ± 0.0531	44.70 ± 4.53	0.6965 ± 0.0542
Dataset 2	**SynthRAD Challenge Dataset**		
Models	**SSIM (↑)**	**MSE (↓)**	**NCC (↑)**
Proposed	**0.4603 ± 0.0306**	**71.88 ± 6.20**	**0.8143 ± 0.0524**
w/o BGRU	0.4572 ± 0.0312	71.92 ± 6.21	0.8102 ± 0.0535
w/o Transformer	0.4598 ± 0.0302	71.96 ± 6.22	0.8107 ± 0.0559
w/o netF	0.4576 ± 0.0317	71.91 ± 6.19	0.8097 ± 0.0548
w/o L_{SR}	0.4576 ± 0.0306	71.99 ± 6.24	0.8091 ± 0.0568

6 Discussion

The StyleGAN2 generator improves the training time and loss performance compared to the ResNet-based generator because of training stabilization using progressive growing, and the initialization strategy for StyleGAN2. In the local feature extraction network, the convolutional encoder helps to extract intricate detailed local features at multiple scales. Positional information is preserved to some extent through padding in convolutions. However, a separate global feature extraction network block is added to integrate these local features into global features representing larger anatomical structures. In the local global feature correlation module, tokenization establishes a vocabulary representing a compact set of concepts, ensuring a comprehensive representation of global, contextual, and local features. In the feature network, the choice of GELU as the activation function is motivated by its effectiveness in capturing complex patterns in the data, the self-attention layer helps to induce contextual understanding in the embedding vectors generated from image patches during calculation of NCE loss.

7 Conclusion

In this paper, a new method is proposed for improving cross-modal image registration. Our model ensures precise alignment of local features like lymph nodes with global anatomical landmarks, crucial for tasks like segmentation or annotation. By integrating a local feature extractor, a BGRU-based global feature extractor, and a local-global correlation mechanism, our model achieves high

accuracy. A BGRU layer as a global feature extractor is introduced in the domain of medical image registration. Our unique super-resolution loss function further enhances registration precision. Compared to the best-performing state-of-the-art model, our model outperforms in terms of SSIM, MSE, and NCC metrics, with a 36.7% speed improvement.

As shown in this paper, we explored two different CT-MRI datasets of the pelvis region. This model is equally applicable to other paired datasets involving any two different modalities (e.g., CT, 2D or 3D MRI, PET) and images of any organ (e.g., brain, lungs). The main requirements for using this model and algorithm are that both images in the cross-modal pair must be from the same patient, capture the same organ from the same view (though they may be taken at different times and may have undergone natural drift, deformation, or minor imaging artifacts), and must be affine registered. One limitation of the model is it requires substantial computational resources in return of high accuracy. Our future work will be optimizing the algorithm. In the future, we plan to apply this model to tasks like cross-modality lymph node or tumor annotation.

Acknowledgements. Iwahori's research was supported by the Japan Society for the Promotion of Science (JSPS) Grant-in-Aid Scientific Research(C)(24K15019) and the Chubu University Grant.

References

1. Y Liu, R.R Martin, L Chen, X Ren, and L. Li. *Rigid Registration*, pages 681–691. Springer US, Boston, MA, 2014
2. Song Y Zou J, Gao B and Qin J. A review of deep learning-based deformable medical image registration. *Frontiers in Oncology*, 12:1047215, 12 2022
3. G Balakrishnan, A Zhao, M. R Sabuncu, J Guttag, and A. V. Dalca. Voxelmorph: A learning framework for deformable medical image registration. *IEEE Transactions on Medical Imaging*, PP:1–1, 02 2019
4. O Ronneberger, P Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, pages 234–241, Cham, 2015. Springer International Publishing
5. Nicolae-Cătălin Ristea, Andreea-Iuliana Miron, Olivian Savencu, Mariana-Iuliana Georgescu, Nicolae Verga, Fahad Shahbaz Khan, and Radu Tudor Ionescu. Cytran: A cycle-consistent transformer with multi-level consistency for non-contrast to contrast ct translation. *Neurocomputing*, 538:126211, 2023
6. S. Kundu, Y. Iwahori, M. Bhuyan, M. Bhatt, A. Ouchi, and Y. Shimizu. Improved residual vision transformer for ct to mri translation. In *2023 Fifth International Conference on Transdisciplinary AI (TransAI)*, pages 58–65, Los Alamitos, CA, USA, Sep. 2023. IEEE Computer Society
7. G Snaauw, M Sasdelli, G Maicas, S Lau, J Verjans, M Jenkinson, and G. Carneiro. Mutual information neural estimation for unsupervised multi-modal registration of brain images. volume 2022, pages 3510–3513, 07 2022
8. Atousa T. A lss-based registration of stereo thermal-visible videos of multiple people using belief propagation. Comput. Vis. Image Underst., 117(12):17361747, dec 2013.

9. H Sokooti, B De Vos, F Berendsen, M Ghafoorian, S Yousefi, B. P Lelieveldt, I Isgum, and M. Staring. 3d convolutional neural networks image registration based on efficient supervised learning from artificial deformations, 08 2019
10. Hu, Y., et al.: Weakly-supervised convolutional neural networks for multimodal image registration. Med. Image Anal. **49**, 1–13 (2018)
11. F Maes, D Loeckx, D Vandermeulen, and P. Suetens. Medical image registration using mutual information. *Proceedings of the IEEE*, 91:1699 – 1722, 11 2003
12. Heinrich MP, Jenkinson M, Bhushan M, Matin T, Gleeson FV, Brady SM, and Schnabel JA. Mind: Modality independent neighbourhood descriptor for multi-modal deformable registration. *Medical image analysis*, 16:1423–35, 05 2012
13. B Kim, D. H Kim, S. H Park, J Kim, J Lee, and J. C. Ye. Cyclemorph: Cycle consistent unsupervised deformable image registration. *Medical Image Analysis*, 71:102036, 03 2021
14. Liya Zhao and Kebin. Jia. Deep adaptive log-demons: Diffeomorphic image registration with very large deformations. *Computational and Mathematical Methods in Medicine*, 2015:1–16, 05 2015
15. T. Che, Y. Zheng, X. Sui, Y. Jiang, J. Cong, W. Jiao, and B. Zhao. Dgr-net: Deep groupwise registration of multispectral images. In *Information Processing in Medical Imaging*, 2019
16. A Hering, S Kuckertz, S Heldmann, and M.P. Heinrich. Enhancing label-driven deep deformable image registration with local distance metrics for state-of-the-art cardiac motion tracking. In *Bildverarbeitung für die Medizin 2019*, pages 309–314, Wiesbaden, 2019. Springer Fachmedien Wiesbaden
17. M Jaderberg, K Simonyan, A Zisserman, and K. Kavukcuoglu. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015
18. J Chen, E. C Frey, Y He, W. P Segars, Y Li, and Y. Du. Transmorph: Transformer for unsupervised medical image registration. *Medical Image Analysis*, 82:102615, 09 2022
19. Y Zhu and S. Lu. Swin-voxelmorph: A symmetric unsupervised learning model for deformable medical image registration using swin transformer. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 78–87, Cham, 2022. Springer Nature Switzerland
20. J Shi, Y He, Y Kong, J Coatrieux, H Shu, G Yang, and S. Li. *XMorpher: Full Transformer for Deformable Medical Image Registration via Cross Attention*, pages 217–226. 09 2022
21. A. Q Wang, E. M Yu, A. V Dalca, and M. R. Sabuncu. A robust and interpretable deep learning framework for multi-modal registration via keypoints, 04 2023
22. M Arar, Y Ginger, D Danon, I Leizerson, and A. Bermano. Unsupervised multi-modal image registration via geometry preserving image-to-image translation. pages 13407–13416, 06 2020
23. A Casamitjana, M Mancini, and J. E. Iglesias. Synth-by-reg (sbr): Contrastive learning for synthesis-based registration of paired images. *Simulation and synthesis in medical imaging: International Workshop, held in conjunction with MICCAI proceedings. SASHIMI*, 12965:44–54, 2021
24. Qin, C., Shi, B., Liao, R., Mansi, T., Rueckert, D., Kamen, A.: Unsupervised Deformable Registration for Multi-modal Images via Disentangled Representations. In: Chung, A.C.S., Gee, J.C., Yushkevich, P.A., Bao, S. (eds.) IPMI 2019. LNCS, vol. 11492, pp. 249–261. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-20351-1_19

25. Z Chen, J Wei, and R. Li. Unsupervised multi-modal medical image registration via discriminator-free image-to-image translation. *ArXiv*, abs/2204.13656, 2022
26. M. A Islam, M Kowal, S Jia, K. G Derpanis, and N. D. Bruce. Position, padding and predictions: A deeper look at position information in cnns. *ArXiv*, abs/2101.12322, 2021
27. J. Wang, Q. Chen, H. Pei, Q. Sun, and Y. Jia. Dynamic global-local attention network based on capsules for text classification. pages 1–8, 07 2020
28. B Wu, C Xu, X Dai, A Wan, P Zhang, Z Yan, M Tomizuka, J Gonzalez, K Keutzer, and P. Vajda. Visual transformers: Token-based image representation and processing for computer vision, 06 2020
29. K. He C. Dong, C. C. Loy and X. Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38:295–307, 2014
30. J Zhu, T Park, P Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networkss. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017
31. P Isola, J Zhu, T Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 2017
32. Adrià Casamitjana, Matteo Mancini, and Juan Eugenio Iglesias. Synth-by-reg (sbr): Contrastive learning for synthesis-based registration of paired images. In David Svoboda, Ninon Burgos, Jelmer M. Wolterink, and Can Zhao, editors, *Simulation and Synthesis in Medical Imaging*, pages 44–54, Cham, 2021. Springer International Publishing

A Hybrid Domain Encoder-Decoder Network for Progressive k-space Extrapolation in Super-Resolution MRI

Vazim Ibrahim[1,2,3] and Joseph Suresh Paul[2,3(✉)]

[1] Cochin University of Science and Technology, Ernakulam, Kerala, India
[2] Indian Institute of Information Technology and Management-Kerala, Trivandrum, Kerala, India
{vazim.i,j.paul}@iiitmk.ac.in
[3] Digital University Kerala, Trivandrum, Kerala, India
j.paul@duk.ac.in

Abstract. In this paper, we address the problem of Superresolution magnetic resonance imaging (SR-MRI) to reconstruct a high-resolution (HR) image from a low-resolution (LR) scan using progressive blocks of hybrid domain network. Each block consists of cascaded stages of an encoder-decoder based network model in the frequency domain (FDN) and spatial domain (SDN). We perform the SR-MRI reconstruction by estimating the signal components encoded into the higher frequencies sequentially using FDN-SDN pairs with progressively increasing k-space coverage. In this scheme, each block of hybrid-domain encoder-decoder type convolutional neural network (CNN) is sequentially trained to estimate the unknown signal values at k-space locations within the immediate neighborhood of the region covered by the previous block. A key advantage is the consistent PSNR improvement exceeding 3.2 dB at 4-fold down-sampling, achieved without employing multi-contrast training samples or multistreaming architecture.

Keywords: Super-resolution MRI · Deep Learning · convolutional neural networks · supervised learning · Normalized Cross-Correlation

1 Introduction

High spatial resolution in Magnetic Resonance Imaging (MRI) is often accompanied by reduced Signal-to-Noise Ratio (SNR), and attained at the cost of inherently long scan time. Although, it is possible to increase the SNR with simultaneous reduction in acquisition time by resorting to higher field strengths, the situation cannot be same with MR sequences of different organs and structures[20]. This is because high-field strength contributes to increased inhomogeneity effects, artifacts due to the cerebrospinal fluid (CSF) flow, magnetic

Supported by organization x.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15311, pp. 127–142, 2025.
https://doi.org/10.1007/978-3-031-78195-7_9

susceptibility artifacts; artifacts due to breathing, patient motion and swallowing. An easier approach is to retrospectively increase the spatial resolution using image super-resolution (SR) techniques applied to MRI (SR-MRI) [2–4,6,7,9–17,21,23,25,27,28,30].

Broadly, SR-MRI can be divided into two types based on the low-resolution (LR) image model. The first type (Type-I) attempts to increase the through-plane or in-plane resolution using a spatial domain degradation model that includes Gaussian blurring followed by down-sampling in the spatial domain. For in-plane super-resolution, the LR image is represented by a larger voxel size in the in-plane. In this case, an initial estimate of the high-resolution (HR) image is determined by linear or bicubic interpolation. In a recent Type-I SR-MRI deep learning model with mean square error (MSE) based reconstruction loss regularized using Laplacian and rank priors, the interpolated HR image forms the input to a convolutional neural network (CNN) with feedback connections from higher-level layers that compute feature maps based on variance of the Laplacian to capture the sharpness of fine local structures, and rank reduction for learning the global structures [2]. The spatial domain networks have also been used to improve the through-plane resolution for 3D acquisition performed with a matrix size set for high in-plane resolution and large slice thickness [8,14,15,17,27].

For a trajectory designed with a given matrix size and Bandwidth per pixel, the Type-II methods model the degradation resulting from acquisition performed using a subset of the low-frequency locations on the trajectory. In a Cartesian scan, this is equivalent to acquiring only the central portion of the kspace, with the scanner parameters set at the full matrix size and k-space coverage for a given Field-of-View (FOV). An initial estimate of the HR image is then obtained from the zero-padded k-space [9,11,12,24,26]. At higher down-sampling factors, this leads to the generation of oscillatory artifacts. The classic problem of removal of the undesirable oscillatory artifacts is often addressed by either pure filtering techniques [1], or k-space extrapolation techniques that exploit one or more forms of prior information about the HR image [18], or a combination thereof [9]. In addition to the undesirable high-frequency artifacts, reconstruction at higher down-sampling factors also involves estimation of signal values at more number of unaquired locations. Although Channel splitting network (CSN) [29] that performs residual and dense learning on the split features of a convolution layer is a preferred choice in this situation, the reduced number of low-frequency observations incapacitates the exploitation of redundant low-level information to retrieve complex high-frequency spatial patterns at the deeper part of the network. With encoder-decoder (EnDec) architecture, the propagation of undesirable high-frequency spatial features can be controlled by adversarial training the EnDec Generator network using a combination of MSE loss, perceptual loss, and Texture matching loss as shown in [11]. A significant contribution in [11] is the performance improvement at higher down-sampling factors obtained by concatenating multi-contrast features to the high-level feature maps in the decoder part using a pre-trained reference network (MCSR). At higher down-sampling factors, this network is progressively connected to have different levels required

for learning the complex spatial patterns. Our goal in this paper is to develop a hybrid-domain EnDec (Hd-EnDec) architecture that exploits the frequency domain information in the limited number of low-frequency observations, and thereby improve the performance of high-level feature extraction. In the absence of multi-contrast training features, the Hd-EnDec architecture provides higher PSNR performance in comparison to both the CSN and the EnDec Generator network employed in MCSR.

Because the oscillatory artifacts manifest as a result of inverse Fourier operation performed on the truncated k-space, the undesired spatial patterns can be avoided by directly training a frequency domain network (FDN) to learn the high-frequency signal values. The FDN is trained with the observed k-space zero-filled at the high-frequency locations forming the input and the fully acquired k-space as the reference. Although the architecture is similar to the FDN in compressed sensing MRI (CS-MRI) [5,19], the reduced number of observed samples in Type-II SR-MRI makes the problem heavily undetermined. Therefore, a direct estimation of the signal values becomes challenging. Instead, we resort to a progressive scheme by estimating the signal components encoded into the higher frequencies sequentially using FDN-SDN pairs with progressively increasing k-space coverage. In this scheme, each block of hybrid-domain encoder-decoder type CNN is sequentially trained to estimate the unknown signal values at k-space locations within the immediate neighborhood of the region covered by the previous block. To achieve this, we make use of the high spectral densities embodied in the observed low-frequency samples to learn the spectral densities in the neighboring high-frequency region by maximizing the Normalized Spectral Cross-correlation Energy loss function. In this process, the network relearns the low-frequency samples with higher spectral energy and preserves the spectral density of the reference k-space in the low-frequency region. This results in smoothed estimates of the high-frequency signal values in the neighboring regions, with smoothed spatial features in the absence of oscillatory effects. Therefore, the image reconstructed from FDN acts as a suitable training input to a second stage of spatial domain learning using the MSE loss function and the low-frequency consistency prior. The high-frequency information learned in SDN will thus be closer to the reference k-space. The cascaded block of pre-trained FDN and SDN (Hd-EnDec) gives 0.8-1.2dB improvement in PSNR at four-fold downsampling with reference to the 2-level progressive MCSR network in the absence of multi-contrast prior. In addition, the Hd-EnDec approach facilitates improved reconstruction compared to the conventional Super-Resolution Deep Learning (SRDL) frameworks used in ESRGAN, FA-GAN, W2AMSN, CSN and FP-GAN. These frameworks utilize deep enhancement modules emphasized with multi-streaming architecture or attention modules, with or without generative learning, to represent local and global high-frequency details in the high-level abstractions of the model.

With more high-frequency samples corresponding to large Tread or sampling rate, the number of layers required in a single FDN would be too large to ensure generalizability. To circumvent this, we divide the k-space into blocks of increas-

ing dimensions around the observed low-frequency samples. The innermost block with the least number of high-frequency samples are estimated first with a pre-trained hybrid network. The output k-space of this network is used to train a second hybrid network to estimate the high-frequency samples in the next outer block and so on. In this, the samples that are common to both blocks are relearned by the outer hybrid network with higher accuracy as compared to that estimated by the inner network. For a matrix size of 256 and down-sampling factor of 4, we typically use 2-3 hybrid blocks; with the inner block covering the central 128×128 region. The paper is organized into 4 sections. In the first section, we give a brief overview of the background on SR-MRI and related works. The theory section includes a brief description of the SR-MRI acquisition schemes for fast acquisition and partial k-space, the implementation of FDN, SDN and Hd-EnDec. The theory section also includes the combination of loss functions used to generalize the network models. The results section includes application of FDN, SDN and Hd-EnDec to publicly available brain data. The results are compared with ESRGAN [22], FA-GAN [7], W2AMSN [21], CSN [29] and FP-GAN [25] networks. The paper concludes with a detailed discussion of its implications and relative advantages.

2 Theory

2.1 SR-MRI Problem Formulation

In this paper, we broadly address the problem of SR-MRI using deep networks trained with the limited low-frequency samples obtained with Fast acquisition (FA) schemes in which the k-space coverage is limited to the central region for both Cartesian and non-Cartesian acquisition. SR-MRI designed for specified k-space coverage and matrix size aims to reconstruct a high-resolution MR image by acquiring only the central k-space points with reduced readout duration. In addition to the advantage of reducing the effective readout time, post-processing in SR-MRI achieves reconstruction without high-frequency oscillations due to k-space truncation, while retaining fine tissue details. The training data for a given imaging region (such as the brain, spine etc.) with the same resolution and contrast weighting, comprises of a set of HR images $\{f^{(k)}(x,y) \mid f^{(k)} \in R_+^{M\times N}$ $k = 1, 2, ..., N_t\}$ ($N_t \backsim 4000$ samples or more) obtained as the Fourier inverse of a fully acquired k-space $\{F^{(k)}(k_x,k_y) | F^{(k)} \in \mathbb{C}^{M\times N},\ k = 1, 2, ..., N_t\}$ for Cartesian acquisition and inverse NUFFT (iNUFFT) [26] for non-Cartesian trajectory. In the rest of this paper, notations without inclusion of the superscript denote samples used for prediction. The Type-II SR-MRI problem is to reconstruct $F(k_x,k_y)$ from the zero-filled LR k-space $G(k_x,k_y) \coloneqq Mask \circ F(k_x,k_y)$. The down-sampling operator determines the number and location of sampling points to be acquired on the specified trajectory. Based on the standard trajectory type and acquisition scheme, the mask to be chosen for SR-MRI takes the form:

$$Mask(k_x, k_y) := \begin{cases} 1 & \text{for } -\frac{k_{x_{max}}}{r} \leq k_x(t) \leq \frac{k_{x_{max}}}{,} - \frac{k_{y_{max}}}{\leq} k_y(t) \leq \frac{k_{y_{max}}}{,} t \leq \frac{T_{read}}{r} \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where r denotes an implicit down-sampling factor and T_{read} represents the read-out time. Unlike the Type-I SR-MRI, the implicit nature of the down-sampling factor effectively means that the spatial dimensions of both the LR and HR images are equal.

2.2 Proposed Learning Approach and Network Architecture

The hybrid architecture is composed of an FDN model (Γ_{Θ_F}) sequentially connected to the input of an SDN model (Γ_{Θ_I}) through a Fourier Transformation layer. The training process consists of first training Γ_{Θ_F} by optimizing the objective function $\mathcal{L}_F(.)$ with respect to the hyper-parameters:- Θ_F. $\Gamma_{\Theta_I}(.)$ is then trained with the pre-trained layers of Γ_{Θ_F} with respect to Θ_I .

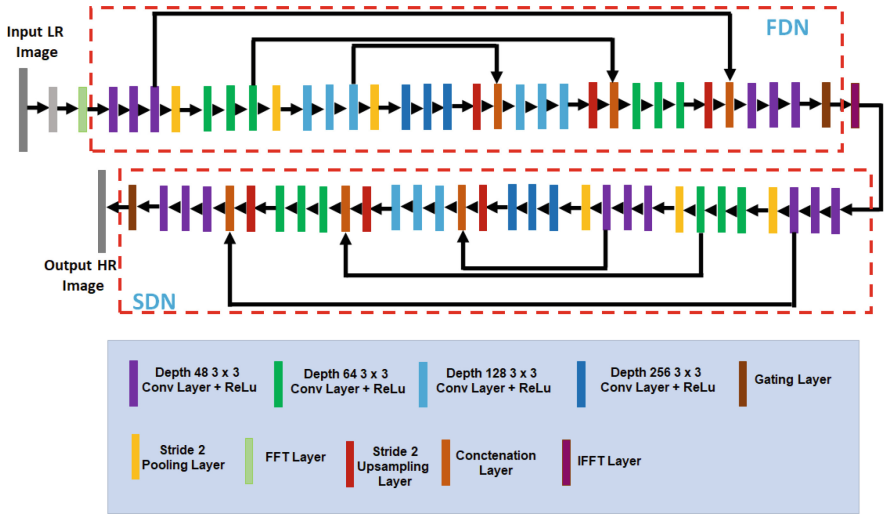


Fig. 1. A schematic depiction of the Hd-EnDec architecture. Red box highlights the EnDec architecture used for FDN and SDN models.

FDN model The feature map at the input layer of FDN can be represented as

$$\mathcal{G}_0 = \mathcal{T}(G(k_x, k_y))\mathcal{G}_0 \in \mathbb{R}^{M \times N \times 2}, \tag{2}$$

where $\mathcal{T}$ denotes a Tensor operator to represent the k-space as a tensor with two separate feature maps consisting of the real and imaginary parts. With N_t number of training samples, the notation for the feature map is extended to include

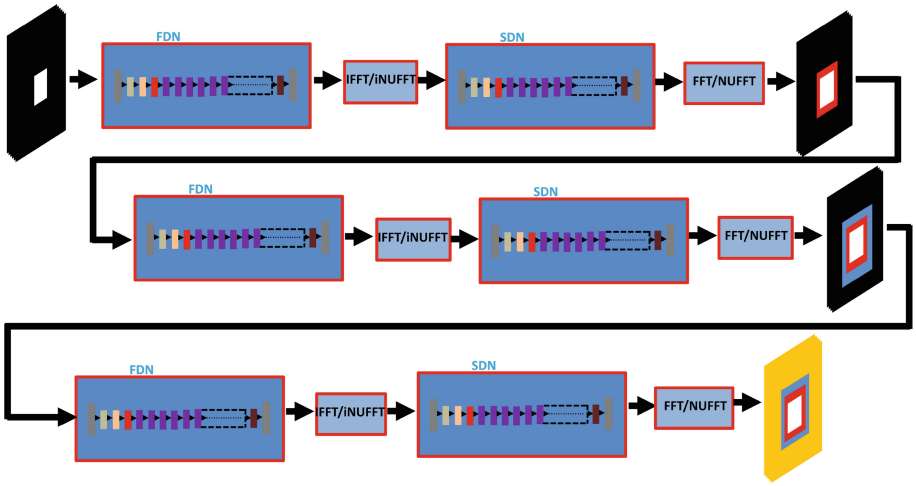


Fig. 2. A schematic depiction of the training process with progressive Hd-EnDec network. The progressive learning strategy involves stage-wise training of each Hd-EnDec model for a specific extent of k-space, followed by an inverse Fourier transform to provide input to the SDN model within the Hd-EnDec network. The inferred output image from the previous stage's SDN is upscaled to the next higher radix-2 dimension and then fed into the subsequent Hd-EnDec for training. In this stage, the previous Hd-EnDec is set to non-trainable mode. This process ultimately generalizes a hybrid domain sequential model to reconstruct the HR image.

the ensemble training data as $\mathfrak{G}_0 := \{\mathcal{G}_0^{(k)}\}_{k=1}^{N_t}$. Similar to a conventional CNN, the features maps in FDN are spatially connected with sparse weights in each successive layer. Similarly, the depth of each layer increases successively in the encoding path and reduces in the decoding path.The input and output feature maps of the s^{th} encoding layer is denoted by $\{\mathcal{G}_{s-1}^{(k)}\}_{k=1}^{N_t}$, $\mathcal{G}_{s-1}^{(k)} \in R^{M\times N\times D_{s-1}}$ and $\{\mathcal{G}_s^{(k)}\}_{k=1}^{N_t}$, $\mathcal{G}_s^{(k)} \in R^{M\times N\times D_s}$, $s = 1, 2, ..., S$, respectively. The depth D_s represents the number of feature maps in the s^{th} layer. The interconnections in each layer are obtained using a weight kernel $W_s \in R^{H_s\times V_s\times D_s}$ and bias $B_s \in R^{D_s\times 1}$. In CNN, the input and output feature maps are related by a non-linear mapping $\gamma(.)$ applied to a linear transformation of the input feature map. The linear transformation involves convolution of the input feature map containing weight kernel together with addition of a bias term. The input-output relationship for the s^{th} convolution layer can be expressed as

$$\begin{aligned} \mathcal{G}_s^{(k)} &= \gamma\big(W_s * \mathcal{G}_{s-1}^{(k)} + B_s\big), \\ \Theta_{F_s} &= (W_s, B_s). \end{aligned} \tag{3}$$

The mapping of a layer unit defined as in (3) can be extended to that of a U-net architecture [23] consisting of an encoder and decoder path in which the

output of $(s^{'})^{th}$ convolutional layer unit of the decoder path can be expressed as

$$\begin{aligned} \mathcal{G}_{s^{'}}^{(k)} &= \gamma\big(W_{s'} * \mathcal{G}_{s^{'}-1}^{(k)} + B_{s'}\big), \\ \Theta_{F_{s^{'}}} &= (W_{s'}, B_{s'}), s^{'} = 1, 2, ..., S^{'}, \end{aligned} \tag{4}$$

where $\mathcal{G}_{s'-1}^{(k)} \in \mathbb{R}^{M\times N\times D_{s'-1}}$, $\mathcal{G}_{s'}^{(k)} \in \mathbb{R}^{M\times N\times D_{s'}}$ are the input and output feature maps with depths $D_{s'-1}$ and $D_{s'}$. The output feature map $\mathcal{G}_{S}^{(k)}$ of S^{th} convolution layer in the encoder path is fed to the layer $s' = 1$ in the decoder path, i.e., $\mathcal{G}_{s'-1}^{(k)} \coloneqq \mathcal{G}_{S}^{(k)}$. $\Theta_{F_{s'}} \subset \Theta_F$ represents the hyper-parameter of the $(s^{'})^{th}$ convolutional layer unit in the decoder path that consists of weight kernel/filter $W_{s'} \in R^{H_s\times V_s\times D_{s'}}$ and bias $B_{s'} \in R^{D_{s'}\times 1}$ with reduced depth $D_{s'} < D_s$. The output feature map $\mathcal{G}_{S'}^{(k)}$ of $(S^{'})^{th}$ convolutional layer in the decoder is reduced to depth $D_0 = 2$ (corresponding to the real and imaginary components)using a gating convolution layer. With this, the final output of FDN can be represented as

$$G_{out}^{(k)} = \mathcal{T}^{\dagger}(\mathcal{G}_{\text{gating}}^{(k)}), G_0^{(k)} \in \mathbb{C}^{M\times N\times 1}, \tag{5}$$

where $\mathcal{T}^{\dagger}$ is the adjoint tensor operator, $\mathcal{G}_{\text{gating}}^{(k)} = \Psi(W_{\text{gating}} * \mathcal{G}_{s^{'}-1}^{(k)} + B_{\text{gating}})$ and Ψ represents the linear activation function applied to the gating layer.

Fourier domain Loss function To estimate the k-space signal in the immediate neighborhood, the spectral densities in the adjoining high-frequency region are learned by maximizing the Normalized Spectral Cross-correlation Energy loss function. During training, the network relearns the low-frequency samples with higher spectral energy while preserving the spectral density of the reference k-space in the low-frequency region. The spectral cross-correlation between the FDN output $G_{out}^{(k)}$ and the reference k-space $F^{(k)}$ can be expressed as

$$\begin{aligned} \mathcal{R}_{\text{FG}}(\omega_x, \omega_y) = \iint_{\Omega_x,\Omega_y=-\infty}^{\infty} & \overline{F}^{(k)}(\Omega_x, \Omega_y) \\ & G_{out}^{(k)}(\omega_x + \Omega_x, \omega_y + \Omega_y)\partial\Omega_x\partial\Omega_y, \end{aligned} \tag{6}$$

where $\overline{F}^{(k)}$ denotes the conjugate of the fully acquired k-space of the $k^{'th}$ training sample and (ω_x, ω_y) denotes the spectral lags. Since the cross-correlation is maximum at $(\omega_x \coloneqq 0, \omega_y \coloneqq 0)$), we define a metric using the discrete form of Spectral cross-correlation at $(\omega_x \coloneqq 0, \omega_y \coloneqq 0)$) as $\sum_{\Omega_x,\Omega_{y=1}}^{M-1,N-1} = \overline{F}^{(k)}(\Omega_x, \Omega_y)G_{out}^{(k)}(\Omega_x, \Omega_y)$. Therefore, maximizing the cross-spectral density would imply normalizing the spectral cross-correlation with the spectral energy function. This yields the Normalized spectral cross-correlation energy (NSCCE) loss function:

$$\mathcal{L}_{\text{FG}} = \frac{|\sum_{\Omega_x,\Omega_{y=1}}^{M-1,N-1} \overline{F}^{(k)}(\Omega_x, \Omega_y)\, G_{out}^{(k)}(\Omega_x, \Omega_y|^2}{Nor_{FG}^2}, \tag{7}$$

where $Nor_{FG}^2 = |F^{(k)}(\Omega_x, \Omega_y)|^2$.

SDN Model The SDN model Γ_{Θ_I} employs an architecture similar to that of FDN. The input training sample $g_0^{(k)}$ is obtained by inverse Fourier transforming the output feature map $G_{\text{out}}^{(k)}$ of the pre-trained model Γ_{Θ_F}. The SDN output in each forward pass is denoted as $\hat{f}^{(k)}$. The reference image $f^{(k)}$ used for computing the loss function $\mathcal{L}_I$ is obtained by inverse Fourier transforming $F^{(k)}$. The loss function $\mathcal{L}_I$ used to optimize the hyper-parameter Θ_I is the mean square error (MSE) expressed as

$$MSE(\hat{f}^{(k)}, f^{(k)}) = \frac{1}{MN} \|\hat{f}^{(k)} - f^{(k)}\|_{\text{F}}^2, \tag{8}$$

where $\hat{f}^{(k)}$ and $f^{(k)}$ denote the reconstructed and ground-truth HR images, respectively.

Hd-EnDec Model The proposed network architecture of the Hd-EnDec model is shown in Fig. 1. The network consists of an SDN cascaded to the output of a pre-trained FDN. Both the FDN and SDN components consist of a total of 4 convolutional blocks each in the encoder path. Each block has three convolutional layer units having the same depth. The depths of the convolutional layer units in consecutive blocks correspond to 48, 64, 128 and 256 filters. The decoder in each network component has a similar architecture, but with depths decreasing from 256 to 48. Each convolutional block in the encoder path also consists of a pooling layer with stride 2×2. Likewise, a convolutional block in the decoder path is preceded by an up-sampling layer with stride 2×2. The FDN and SDN models are interconnected with an IFFT/iNUFFT layer to enable domain transformation from FDN to SDN.

Progressive architecture of Hd-EnDec The signal values encoded at the higher frequencies contain more information regarding fine structural details. Therefore, the complexity of learning the high-frequency components using a single Hd-EnDec model increases with distance from the k-space origin. To address this problem, a progressive learning strategy is employed whereby the high-frequency components are progressively learned with multiple Hd-EnDec models as shown in Fig. 2. In the progressive learning approach, the k-space is segmented into blocks of increasing dimensions. The innermost block with the least number of high-frequency samples are estimated first with a pre-trained Hd-EnDec model. The estimated k-space is then used to train a second Hd-EnDec model to estimate the high-frequency samples in the next outer block and so on. The samples that are common to both blocks are re-learned by the higher level Hd-EnDec model, resulting in more accurate estimates as compared to the lower level model.

3 Experimental Evaluation

3.1 Datasets

IXI Dataset:-For the retrospective study, we have used the IXI dataset containing registered T2 weighted MR images of 578 patients made publicly available in (https://brain-development.org/ixi-dataset). A total of 3900 pairs of T2 weighted images from randomly selected 30 subjects were used for training and 1300 pairs from another randomly selected 15 subjects for validation with the size of original HR T2 weighted images as 256×256. LR images corresponding to down-sampling factors 2-4 are used as the input image.

3.2 Experimental details

The input for training is the zero-padded LR k-space with the same dimension as that of the target HR image. The training is performed with batch size set to 16. For each down-sampling factor, the learning rate used for the training process is determined from the respective validation loss curve characteristics. It is seen that for both down-sampling factors, the validation loss curves show improved convergence for the learning rate value of 1e-3.The loss functions are backpropagated in the training process using Adam optimizer. The general pipeline for the whole study consists of: 1) organizing LR k-spaces for training such that real and imaginary components are separated into two channels and serve as the input feature maps to the model, 2) the network is trained for a pre-fixed number of epochs with early stopping and check points made available based on the validation loss, 3) After generalizing the model, the test sample pairs consisting of LR and corresponding HR k-spaces are used to predict the super-resolved HR images, followed by PSNR computation referenced to the groundtruth HR images. All the work was implemented using Tensorflow on a NVIDIA Tesla V100-SXM2-16GB GPU in google colab platform and Quadro RTX 5000 GPU with 16GB of RAM.

4 Results

4.1 Ablation Study

This ablation study is meticulously designed to evaluate the effectiveness of progressive frequency domain learning in predicting high-frequency signal measurements and its influence on architectural choice. The experiment aims to reconstruct HR k-space with high-frequency measurements of size 256×256 from LR k-space of size 64×64, which is dominated by low-frequency signal values. The experiment also includes a sensitivity analysis to assess the impact of adding convolution layers in abstracting high-level features in frequency domain learning. Two network models are considered for the experiments: the Progressive Sequential Hybrid Domain Network (PSHDN), which is the proposed Hd-EnDec model, and the existing Sequential Hybrid Domain Network (SHDN), used to

address the CS-MRI problem [5]. The SHDN architecture is similar to the Hd-EnDec but lacks progressive upsampling and the EnDec architecture in FDN. Instead, it employs a cascade of convolution layers with a data consistency layer in FDN part of the hybrid model. For uniformity, the FDN part of PSHDN is also designed with a cascade architecture that includes progressive upsampling layers and data consistency at each progressive level of the HDN, making the FDN function as a decoder model. However, the SDN parts of both PSHDN and SHDN use the EnDec architecture. This architectural design effectively demonstrates the influence of progressive FDN learning compared to a conventional cascaded architecture in FDN. For the study, the convolution kernel size and the number of blocks in both PSHDN and SHDN are fixed at 3×3 and 4, respectively. Fixing the number of blocks at 4 means the first level of HDN in both PSHDN and SHDN will learn to predict k-space of size 128×128. Subsequently, they will predict k-space of size 192×192 and finally predict the HR k-space of size 256×256 from the previously predicted k-space of size 192×192. Additionally, PSHDN reduces the dimension at specific progressive blocks, unlike SHDN, where the dimension of the feature map or estimated k-space is consistently 256×256 at all HDN levels, starting with the zero-padded 64×64 LR k-space and its subsequent LR image. The zero-padding ensures the size is 256×256.The sensitivity analysis varies the number of convolution layers in the FDN part of the HDN model from 1 to 7. The SDN part uses the EnDec model with 3 convolution layers in each upsampling/downsampling module.

The visual quality of the reconstructed HR k-space is illustrated in Fig. 3. The left panel presents the ground truth HR k-space, HR image, and the LR k-space given as input to PSHDN/SHDN. The right side shows the reconstructed HR k-space, where the panels in rows 1 and 3 display the output of HR k-space from progressive levels 1 and 4 of PSHDN. Similarly, the panels in rows 2 and 4 show the output of HR k-space from progressive levels 1 and 4 of SHDN. The column-wise variation shows the output HR k-space obtained for different numbers of convolution layers in the FDN, denoted as N_d.It can be seen that the progressive upsampling architecture of PSHDN enables more number of accurate high-frequency predictions compared to the cascaded architecture. Additionally, when the number of convolution layers in the FDN increases from 3 to 7, there is an accumulation of signal values in the high-frequency region due to over-abstraction of high-level features with increased depth. This effect is also observed in PSHDN/SHDN, resulting in a reduced number of high-frequency signal predictions in the final progressive level as N_d increases from 3 to 7. From this, it can be inferred that the FDN model should be designed such that hierarchical representation of features in the frequency domain cannot be directly interpreted using high-level local modeling capabilities of convolution layers, as is done to realize hierarchical representation of low-level to high-level structural features extracted in the spatial domain.

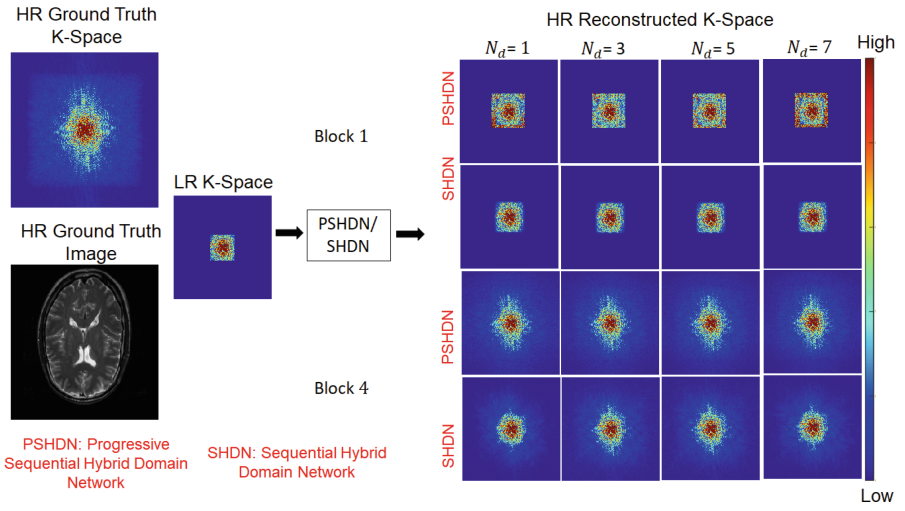


Fig. 3. Illustration of reconstructed HR k-space for varying the number of convolution layers in the FDN part of the Hd-EnDec model, compared to an existing sequential hybrid domain model.

4.2 Fast Acquisition

The PSNR performance of the Hd-EnDec architecture is compared against FDN and SDN in Fig. 4 for down-sampling factors 2, 3 and 4, respectively. The row-wise panel a) to c) shows the zoomed-in region of reconstructed HR image for the respective down-sampling factors 2-4. Left-Right column-wise panels 1-6 show the ground-truth, LR and HR images reconstructed using the FDN, SDN and Hd-EnDec deep networks. FDN is shown to result in PSNR gains of 1.38dB, 2.88dB and 2.32dB for the respective down-sampling factors, with improved visual quality of the reconstructed HR image compared against LR and FDN. In lieu of preservation of signal characteristics, the effect of remnant ringing present in FDN is reduced. Although application of attention mechanism in the Fourier domain results in only a fractional improvement in the overall PSNR, the undesirable high-frequency artifacts are suppressed by preserving the signal characteristics of k-space. In similar lines, the PSNR comparison of SDN architectures is shown in fifth column of Fig. 4. Although SDN achieves better PSNR performance (1.2dB, 2.1dB and 1.7dB for the three down-sampling factors) due to preservation of structural integrity gained from spatial domain processing, the inherent blur present in the reconstructed structures limits its standalone usage in a clinical setting. Moreover, the Hd-EnDec models achieves better PSNR performance (3.89dB, 4 dB, 3.72dB for the three down-sampling factors) that shows the improved reconstruction attained with the hybrid learning.

Table 1. Quantitative comparison with the state-of-the-art methods for $\times 2$, $\times 3$ and $\times 4$

Method	**IXI dataset**					
	$\times 2$		$\times 3$		$\times 4$	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
LR	36.18 ± 1.3	$0.964 \pm .010$	30.68 ± 1.1	$0.903 \pm .015$	27.63 ± 0.8	$0.824 \pm .017$
ESRGAN[22]	36.92 ± 1.1	$0.967 \pm .001$	32.14 ± 1.1	$0.917 \pm .001$	30.17 ± 1.1	$0.895 \pm .001$
FA-GAN[7]	37.45 ± 1.1	$0.974 \pm .001$	32.51 ± 1.1	$0.945 \pm .001$	30.47 ± 1.1	$0.902 \pm .001$
W2AMSN[21]	38.72 ± 1.1	$0.971 \pm .001$	32.78 ± 1.1	$0.952 \pm .001$	30.96 ± 1.1	$0.921 \pm .001$
CSN[29]	38.63 ± 1.1	$0.980 \pm .001$	32.85 ± 1.1	$0.949 \pm .001$	30.85 ± 1.1	$0.916 \pm .001$
MCSR[11]	38.20 ± 1.1	$0.976 \pm .001$	33.01 ± 1.1	$0.960 \pm .001$	31.24 ± 1.1	$0.943 \pm .001$
FP-GAN[25]	38.86 ± 1.1	$0.972 \pm .001$	33.24 ± 1.1	$0.958 \pm .001$	31.42 ± 1.1	$0.932 \pm .001$
Hd-EnDec	$\mathbf{39.63 \pm 1.1}$	$\mathbf{0.981 \pm .001}$	$\mathbf{34.13 \pm 1.1}$	$\mathbf{0.964 \pm .001}$	$\mathbf{31.57 \pm 1.1}$	$\mathbf{0.947 \pm .001}$

4.3 Comparison with state-of-the-art methods

For the case of fast acquisition, a comparative assessment of the hybrid network is performed with the state-of-the-art methods. The latter include MCSR [11] and CSN [29]. The images reconstructed using the above methods are illustrated in Fig. 5 for down-sampling factors 2, 3 and 4, respectively. The left-right column-wise panels represent the ground-truth, LR image; HR images reconstructed using: CSN, MCSR and Hd-EnDec (i.e., the proposed method) from two different sources in the IXI database. Row-wise panels show the respective images at different down-sampling factors. While for down-sampling factor 2, all the reconstructed images are shown to preserve structural integrity, our method alone is found to be capable of alleviating the ringing effects fully together with simultaneously retaining the sharpness of the fine structural elements as indicated by the yellow arrow in Fig. 5(b). Although the structures are relatively blurred and distorted with higher down-sampling factors as shown in Fig. 5, the relative sharpness is higher with the proposed method as indicated by the respective yellow arrows in panels-b), d) and f). In addition, the performance of the Hd-EnDec network is also evaluated quantitatively and compared against the state-of-the-art methods in Table 1. It is seen that SR-MRI using Hd-EnDec shows superior performance with PSNR gains of 3.89dB, 4 dB and 3.72dB at down-sampling factors 2, 3 and 4, respectively. Table 1 also reports an improvement in SSIM score attained by the Hd-EnDec network for all down-sampling factors as compared to other methods.

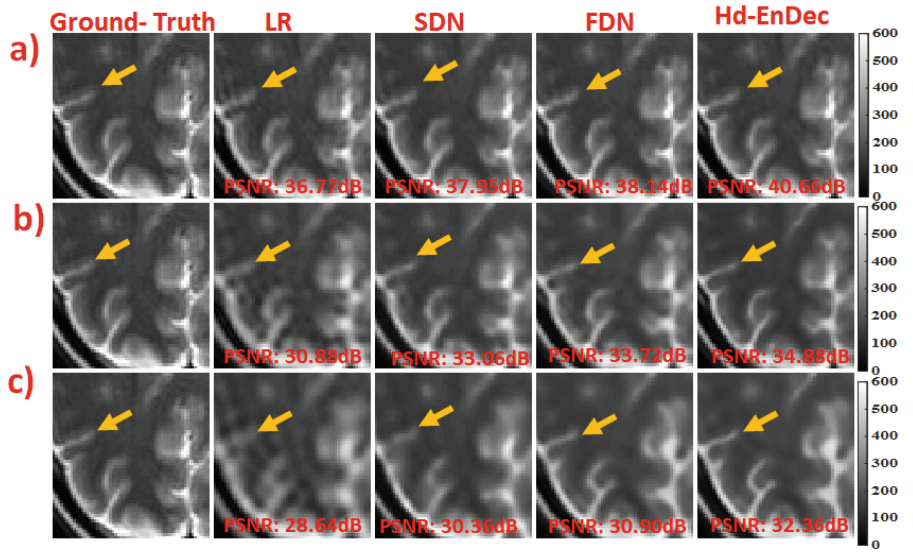


Fig. 4. Visual comparison of LR, FDN, SDN and Hd-EnDec architecture for down-sampling factors 2, 3 and 4 with the ground-truth HR image chosen from IXI dataset. Column-wise panels show the respective images and the corresponding k-space hot maps.The yellow and blue arrow in rows- 1, 3 and 5 show regions with blurred structures in the LR image.

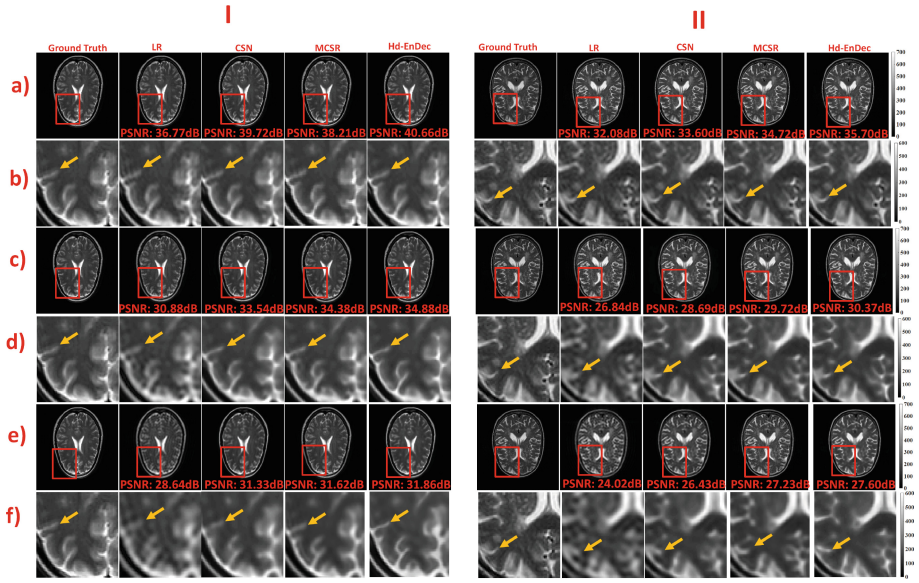


Fig. 5. Visual comparison of LR, CSN, MCSR and Hd-EnDec for down-sampling factors 2, 3 and 4 with the ground-truth HR image chosen from IXI dataset. Yellow arrow indicates regions with blurred structures in the LR image.

5 Conclusion

As opposed to the conventional form of spatial domain super-resolution for MRI, we have developed a hybrid domain deep network for performing retrospective super-resolution through robust k-space extrapolation; useful for fast MR acquisition including partial Fourier and non-Fourier acquisitions. The hybrid architecture designed for the specific type of SR-MRI application ensures simultaneous preservation of the signal characteristics of k-space and structural integrity in the spatial domain necessary for accurate retrieval of high-resolution information. With limited number of k-space samples, the efficiency of FDN can be improved using an attention mechanism introduced as a means to correct the higher-level Fourier domain feature maps with the higher-level features of a pre-trained teacher sub-network in the Fourier domain. The inclusion of a pre-trained teacher network entails the application of a knowledge distillation based transfer learning mechanism. Our experimental results demonstrate improved performance of the proposed hybrid domain neural network model for high down-sampling factors in comparison to other state-of-the-art methods.

Acknowledgements. Please place your acknowledgments at the end of the paper, preceded by an unnumbered run-in heading (i.e. 3rd-level heading).

References

1. Archibald, R., Gelb, A.: A method to reduce the gibbs ringing artifact in mri scans while keeping tissue boundary integrity. IEEE Trans. Med. Imaging **21**(4), 305–319 (2002)
2. Chaudhari, A.S., Fang, Z., Kogan, F., Wood, J., Stevens, K.J., Gibbons, E.K., Lee, J.H., Gold, G.E., Hargreaves, B.A.: Super-resolution musculoskeletal mri using deep learning. Magn. Reson. Med. **80**(5), 2139–2154 (2018)
3. Cherukuri, V., Guo, T., Schiff, S.J., Monga, V.: Deep mr brain image super-resolution using spatio-structural priors. IEEE Trans. Image Process. **29**, 1368–1383 (2019)
4. Chun, J., Zhang, H., Gach, H.M., Olberg, S., Mazur, T., Green, O., Kim, T., Kim, H., Kim, J.S., Mutic, S., et al.: Mri super-resolution reconstruction for mri-guided adaptive radiotherapy using cascaded deep learning: In the presence of limited training data and unknown translation model. Med. Phys. **46**(9), 4148–4164 (2019)
5. Eo, T., Jun, Y., Kim, T., Jang, J., Lee, H.J., Hwang, D.: Kiki-net: cross-domain convolutional neural networks for reconstructing undersampled magnetic resonance images. Magn. Reson. Med. **80**(5), 2188–2201 (2018)
6. Guerreiro, J., Tomás, P., Garcia, N., Aidos, H.: Super-resolution of magnetic resonance images using generative adversarial networks. Computerized Medical Imaging and Graphics p. 102280 (2023)
7. Jiang, M., Zhi, M., Wei, L., Yang, X., Zhang, J., Li, Y., Wang, P., Huang, J., Yang, G.: Fa-gan: Fused attentive generative adversarial networks for mri image super-resolution. Comput. Med. Imaging Graph. **92**, 101969 (2021)
8. Li, Y., Iwamoto, Y., Lin, L., Xu, R., Tong, R., Chen, Y.W.: Volumenet: A lightweight parallel network for super-resolution of mr and ct volumetric data. IEEE Trans. Image Process. **30**, 4840–4854 (2021)

9. Luo, J., Mou, Z., Qin, B., Li, W., Yang, F., Robini, M., Zhu, Y.: Fast single image super-resolution using estimated low-frequency k-space data in mri. Magn. Reson. Imaging **40**, 1–11 (2017)
10. Lyu, J., Li, G., Wang, C., Cai, Q., Dou, Q., Zhang, D., Qin, J.: Multicontrast mri super-resolution via transformer-empowered multiscale contextual matching and aggregation. IEEE Transactions on Neural Networks and Learning Systems (2023)
11. Lyu, Q., Shan, H., Steber, C., Helis, C., Whitlow, C., Chan, M., Wang, G.: Multi-contrast super-resolution mri through a progressive network. IEEE Trans. Med. Imaging **39**(9), 2738–2749 (2020)
12. Lyu, Q., Shan, H., Wang, G.: Mri super-resolution with ensemble learning and complementary priors. IEEE Transactions on Computational Imaging **6**, 615–624 (2020)
13. Masutani, E.M., Bahrami, N., Hsiao, A.: Deep learning single-frame and multi-frame super-resolution for cardiac mri. Radiology **295**(3), 552–561 (2020)
14. Peeters, R.R., Kornprobst, P., Nikolova, M., Sunaert, S., Vieville, T., Malandain, G., Deriche, R., Faugeras, O., Ng, M., Van Hecke, P.: The use of super-resolution techniques to reduce slice thickness in functional mri. Int. J. Imaging Syst. Technol. **14**(3), 131–138 (2004)
15. Peled, S., Yeshurun, Y.: Superresolution in mri: application to human white matter fiber tract visualization by diffusion tensor imaging. Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine **45**(1), 29–35 (2001)
16. Plenge, E., Poot, D.H., Bernsen, M., Kotek, G., Houston, G., Wielopolski, P., van der Weerd, L., Niessen, W.J., Meijering, E.: Super-resolution methods in mri: can they improve the trade-off between resolution, signal-to-noise ratio, and acquisition time? Magn. Reson. Med. **68**(6), 1983–1993 (2012)
17. Scheffler, K.: Superresolution in mri? Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine **48**(2), 408–408 (2002)
18. Shi, F., Cheng, J., Wang, L., Yap, P.T., Shen, D.: Lrtv: Mr image super-resolution with low-rank and total variation regularizations. IEEE Trans. Med. Imaging **34**(12), 2459–2466 (2015)
19. Souza, R., Lebel, R.M., Frayne, R.: A hybrid, dual domain, cascade of convolutional neural networks for magnetic resonance image reconstruction. In: International conference on medical imaging with deep learning. pp. 437–446. PMLR (2019)
20. Vargas, M., Delattre, B., Boto, J., Gariani, J., Dhouib, A., Fitsiori, A., Dietemann, J.: Advanced magnetic resonance imaging (mri) techniques of the spine and spinal cord in children and adults. Insights Imaging **9**, 549–557 (2018)
21. Wang, H., Hu, X., Zhao, X., Zhang, Y.: Wide weighted attention multi-scale network for accurate mr image super-resolution. IEEE Trans. Circuits Syst. Video Technol. **32**(3), 962–975 (2021)
22. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Change Loy, C.: Esrgan: Enhanced super-resolution generative adversarial networks. In: Proceedings of the European conference on computer vision (ECCV) workshops. pp. 0–0 (2018)
23. Wu, Z., Chen, X., Xie, S., Shen, J., Zeng, Y.: Super-resolution of brain mri images based on denoising diffusion probabilistic model. Biomed. Signal Process. Control **85**, 104901 (2023)
24. Xiong, J., Fox, P.T., Gao, J.H.: The effects of k-space data undersampling and discontinuities in keyhole functional mri. Magn. Reson. Imaging **17**(1), 109–119 (1999)

25. You, S., Lei, B., Wang, S., Chui, C.K., Cheung, A.C., Liu, Y., Gan, M., Wu, G., Shen, Y.: Fine perceptive gans for brain mr image super-resolution in wavelet domain. IEEE transactions on neural networks and learning systems (2022)
26. Zaitsev, M., Zilles, K., Shah, N.J.: Shared k-space echo planar imaging with keyhole. Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine **45**(1), 109–117 (2001)
27. Zhao, C., Dewey, B.E., Pham, D.L., Calabresi, P.A., Reich, D.S., Prince, J.L.: Smore: a self-supervised anti-aliasing and super-resolution algorithm for mri using deep learning. IEEE Trans. Med. Imaging **40**(3), 805–817 (2020)
28. Zhao, C., Shao, M., Carass, A., Li, H., Dewey, B.E., Ellingsen, L.M., Woo, J., Guttman, M.A., Blitz, A.M., Stone, M., et al.: Applications of a deep learning method for anti-aliasing and super-resolution in mri. Magn. Reson. Imaging **64**, 132–141 (2019)
29. Zhao, X., Zhang, Y., Zhang, T., Zou, X.: Channel splitting network for single mr image super-resolution. IEEE Trans. Image Process. **28**(11), 5649–5662 (2019)
30. Zhu, Q., Ren, Y., Qiu, Z., Wang, W.: Robust mr image super-resolution reconstruction with cross-modal edge-preserving regularization. Int. J. Imaging Syst. Technol. **29**(4), 491–500 (2019)

MAR-DTN: Metal Artifact Reduction Using Domain Transformation Network for Radiotherapy Planning

Belén Serrano-Antón[1,2,3(✉)], Mubashara Rehman[4,5], Niki Martinel[4], Michele Avanzo[5], Riccardo Spizzo[5], Giuseppe Fanetti[5], Alberto P. Muñuzuri[2,3], and Christian Micheloni[4]

[1] FlowReserve Labs S.L., 15782 Santiago de Compostela, Spain
belen.serrano.anton@rai.usc.es
[2] CITMAga, 15782 Santiago de Compostela, Spain
[3] Group of Nonlinear Physics, University of Santiago de Compostela, 15782 Santiago de Compostela, Spain
[4] Machine Learning and Perception Lab, Università degli Studi di Udine, 33100 Udine, UD, Italy
[5] Centro di Riferimento Oncologico di Aviano IRCCS, 33081 Aviano, PN, Italy

Abstract. For the planning of radiotherapy treatments for head and neck cancers, Computed Tomography (CT) scans of the patients are typically employed. However, in patients with head and neck cancer, the quality of standard CT scans generated using kilo-Voltage (kVCT) tube potentials is severely degraded by streak artifacts occurring in the presence of metallic implants such as dental fillings. Some radiotherapy devices offer the possibility of acquiring Mega-Voltage CT (MVCT) for daily patient setup verification, due to the higher energy of X-rays used, MVCT scans are almost entirely free from artifacts making them more suitable for radiotherapy treatment planning.

In this study, we leverage the advantages of kVCT scans with those of MVCT scans (artifact-free). We propose a deep learning-based approach capable of generating artifact-free MVCT images from acquired kVCT images. The outcome offers the benefits of artifact-free MVCT images with enhanced soft tissue contrast, harnessing valuable information obtained through kVCT technology for precise therapy calibration. Our proposed method employs UNet-inspired model, and is compared with adversarial learning and transformer networks. This first and unique approach achieves remarkable success, with PSNR of 30.02dB across the entire patient volume and 27.47dB in artifact-affected regions exclusively. It is worth noting that the PSNR calculation excludes the background, concentrating solely on the region of interest.

M. Rehman—Co-first author of the manuscript.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-78195-7_10.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15311, pp. 143–159, 2025.
https://doi.org/10.1007/978-3-031-78195-7_10

Keywords: kilo-Voltage-CT (kVCT) · Mega-Voltage-CT (MVCT) · Metal artifact reduction (MAR) · Artificial intelligence (AI).

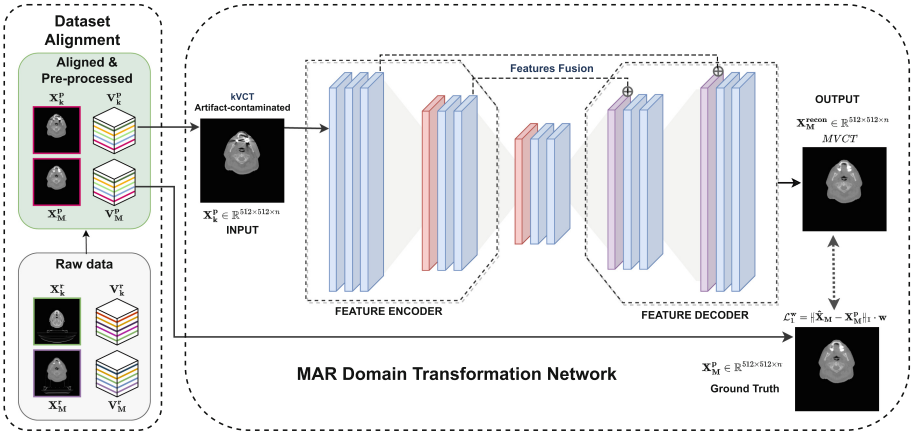
1 Introduction

Since their introduction in the 1970s, advanced medical imaging techniques, particularly high-resolution Computed Tomography (CT), have been crucial for computer-assisted diagnosis [5]. However, when patients with metal implants undergo imaging, such as dental fillings or hip prostheses, severe beam attenuation occurs, resulting in discernible streaks that compromise image fidelity and hampering clinical assessment [1].

Recent advancements in deep learning have shown promise in mitigating metal artifacts through supervised learning methodologies. However, obtaining ground truth images without artifacts is challenging. [9] tackles this issue by generating datasets with and without metal artifacts, enabling the development of numerous algorithms for Metal Artifact Reduction (MAR). Other approaches encompass a variety of image-to-image deep learning models, including deep residual architectures [6] and interpretable convolutional dictionary networks [24]. Numerous other methodologies utilize sinogram-to-sinogram deep learning models [20,30] or dual-domain deep learning models using both image and sinogram data [12,15,28]. These models can be further extended by incorporating state-of-the-art interpolation-based algorithm Normalized MAR corrected data as an extra input [4,14]. A combination of multiple supervised deep learning methods can be effective in reducing metal artifacts from complex cases of cardiac CT images [17]. Approach [25] uses pix2pix [7] for MAR, it introduces bandwise normalization method, which splits a CT image into three channels according to the intensity value and considerably improves the performance of the cGAN. CNN-based approach [29] is introduced to predict an artifact-suppressed prior image. Extending these concepts, [15] introduced DuDoNet, a dual-domain learning technique combining sinogram enhancement and image domain reconstruction. Improved version of DuDoNet [15], restores sinogram consistency and simultaneously enhance CT images by incorporating metal segmentation in both domains. In more recent work, [18] introduced an alternative dual-domain approach, emphasizing deep sinogram completion for improved MAR performance.

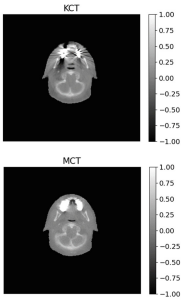
Mega-Voltage Computed Tomography (MVCT) is used for verification of patient positioning immediately before the radiotherapy treatment. It is less prone to streak artifacts from metallic implants because it uses high-energy beams produced by a radiotherapy linear accelerator, which are less attenuated by metal than conventional diagnostic X-rays. The main drawback of MVCT is that it is available only in some specialized radiotherapy machines [5]. [16] proposes to reduce metal artifacts in kVCT by using MVCT images as prior images. The iterative method proposed in [19] segments tissue regions in Megavoltage cone-beam CT images and the metal region in kVCT images for template creation. Forward projection of the templates generates sinograms. Artifact images are reconstructed from the sinograms. Finally, corrected images are obtained by subtracting artifact images from

original kVCT images. [21] utilizes the sinogram of kVCT and MVCT along with the corresponding metal trace to ultimately produce artifact-free kVCT images. Methodology proposed in [10], employing convolutional neural networks to obtain artifact-free kVCT images, by utilizing two networks where the first generates synthetic artifact-free kVCT images from MVCT, which are then used to train the second network. The second network takes kVCT images with artifacts as input and produces artifact-free kVCT images as output.



(a)

(b)



(c)

Fig. 1. (a) Abstract overview of proposed Domain Transformation Network. (b) Sagittal view of the body with distinct delineations of the head, neck, and body regions(blue). (c) kVCT (top) and MVCT (bottom) axial artifact slices after normalization and masking.

Different techniques have been utilized to mitigate metal artifacts in kVCT scans, yet they predominantly operate within the same domain. In contrast, our innovative approach involves transforming the CT domain from kVCT to MVCT, as MVCT is inherently less artifact-sensitive, henceforth preferred for its robustness in clinical applications. We hereby propose Metal Artifact Reduction using

Domain Transformation Network (MAR-DTN) to address metal artifacts in oncological imaging. Our approach generates MVCT using a model that employs a UNet architecture with skip connections, tailored for MAR from kVCT images, to systematically mitigate artifacts during the transformation process from kVCT to MVCT. Leveraging its encoder-decoder structure and spatial awareness, it effectively processes 512x512 pixel images to produce MVCT output (see Fig. 1a). This network is trained by employing 3858 kVCT slices of head and neck region (Fig. 1b).

It has achieved exceptional results, a noticeable point is that Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) in all tables concern the regions of interest only, not the background like the other methodologies. This approach aims to improve CT quality, interpretability, and analysis of medical images by transforming artifact-contaminated kVCT into artifact-free MVCT. This novel method enables radiation oncologists to gain insights into MVCT from kVCT alone, potentially avoiding repeated imaging and its implications for patients' health. In addition, our study entails a comparative evaluation of MAR-DTN's performance in relation to three current state-of-the-art methods, based on adversarial learning and transformers.

2 Methods

2.1 Dataset Collection and Processing

Due to the lack of available aligned kVCT and MVCT datasets, a new dataset consisting of 5469 images from 52 patients from the *National Cancer Institute (CRO) IRCCS*[1]. For each patient, we acquired kVCT and MVCT images; the kVCT images obtained have matrix size of 512 × 512, on the axial plane with a pixel size of 1.074 mm × 1.074 mm, and slice thickness of 2 mm, furthermore, the MVCT images obtained have a matrix size of 512 × 512, on the axial plane with a pixel spacing of 0.754 mm × 0.754 mm with the slice thickness of 2 mm or 4 mm.

Patients underwent intensity-modulated radiotherapy for oropharyngeal or nasopharyngeal cancer. Non-contrast-enhanced CT imaging was performed using a 32-slice scanner (Toshiba Aquilion LB, Toshiba Medical Systems Europe, Zoetermeer, the Netherlands) with parameters set at 120 kVp, 2-5 mm slice thickness, and 1.07-1.17 mm pixel size. Additionally, patients underwent scanning with helical tomotherapy (Hi-Art II Tomotherapy System, Tomotherapy Inc., Madison, Wisconsin, USA), utilizing a radiotherapy 6MV linear accelerator capable of acquiring MVCT images for daily patient setup verification. The imaging beam, produced by the same LINAC as the therapeutic beam, had a nominal energy of 3.5 MV, with slice thickness ranging from 2-5 mm and a pixel size of 0.75 mm.

The slices of each modality volume are manually categorized into three regions: head, neck, and body (see Fig. 1b). The head region comprises from the beginning of the cranial cavity to the chin, while the neck region spans from

[1] Centro di Riferimento Oncologico di Aviano IRCCS, Via F.Gallini 2, Aviano (PN), 33081, Italy

the chin to the shoulders. The remaining (body region) slices are not considered since we care about removing artifacts caused by metal implants in the teeth area. To separate the artifact-corrupted slices from artifact-free slices, we define artifacts in kVCT images as values exceeding 2000 Hounsfield Units (HU), while for MVCT images, the artifact threshold is set at 1000 HU. These thresholds were determined through visual inspection and following recommendations from [10,16].

Table 1. Number of patients and slices (images) in the acquired dataset.The head and neck region include the artifact slices since we work with artifacts caused by metallic dental implants.

Set	Number of patients	Number of slices of the head and neck regions	Number of slices with artifacts
Train	36	3858	560
Validation	10	1031	153
Test	6	580	96

For the training and subsequent evaluation of the proposed model, two datasets are constructed; the first is $\mathcal{D}_{\text{All}}$, the dataset comprises CT slices up to the neck region (including slices with and without artifacts), and the second dataset is $\mathcal{D}_{\text{Art}}$, which contains only artifact-contaminated CT slices. Out of the total number of slices in the dataset, 14.78% exhibit artifacts, hence belong to $\mathcal{D}_{\text{Art}}$. Both datasets are further sub-divided into three distinct datasets, as specifically, 70% of the patients are used for training ($\mathcal{D}^{Tr}_{\text{All}}$ and $\mathcal{D}^{Tr}_{\text{Art}}$), 20% for validation ($\mathcal{D}^{Val}_{\text{All}}$ and $\mathcal{D}^{Val}_{\text{Art}}$), and the remaining 10% for testing ($\mathcal{D}^{Ts}_{\text{All}}$ and $\mathcal{D}^{Ts}_{\text{Art}}$) (see Table 1).

2.2 kVCT-MVCT Alignment and Preprocessing

The primary goal is to create a dataset with aligned kVCT and MVCT images. Despite originating from the same patient and reference system (with the same origin point), both image volumes (kVCT and MVCT) were not pixel-aligned leading to increased challenges (*i.e.*, such as the need to address alignment and artifact reduction simultaneously). To achieve this, image alignment was performed using the Elastix module of 3D Slicer, open source software (version 5.6.1) [3,11].

The aligned kVCT and MVCT volumes undergo normalization to the range $[-1, 1]$. This process involves setting the lower threshold at -1000 for air and upper thresholds at 2000 for kVCT artifacts and 1000 for MVCT artifacts. Additionally, utilizing the segmentation provided by clinicians (depicted as green segmentation in Fig. 1b), the image background is standardized to the value of -1. The result of such a preprocessing on two sample kVCT and MVCT slices is shown in Fig. 1c.

2.3 Proposed Methodology

The objective is to project images acquired in the kVCT domain onto the MVCT domain while removing/reducing the artifacts induced by metallic implants.

In what follows, with $\mathbf{m}$ denoting the kVCT ($\mathbf{k}$) or MVCT ($\mathbf{M}$) modality, we let $\mathbf{X_m^r} \in \mathbb{R}^{d \times d}$ be a raw ($\mathbf{r}$) slice with $d = 512$ denoting the image resolution. The volume containing the n slices of a patient is $\mathbf{V_m^r} \in \mathbb{R}^{d \times d \times n}$. The original images undergo an alignment process (see Section 2.2), resulting in two new volumes, $\mathbf{V_k^a}, \mathbf{V_M^a} = \mathit{alignment}(\mathbf{V_k^r}, \mathbf{V_M^r})$, which are aligned pixel by pixel.

After this process, all the slices in a volume are preprocessed (see Section 2.2) to obtain $\mathbf{X_m^p} = \mathit{preprocess}(\mathbf{X_m^a}), \forall \mathbf{X_m^a} \in \mathbf{V_m^a}$ that collectively define the dataset for the experiments. The summary diagram is shown in Fig.2.

The input to our model is a preprocessed kVCT image, $\mathbf{X_k^p}$, while the ground truth is the corresponding preprocessed MVCT image, $\mathbf{X_M^p}$. The output of the model is the domain transferred kVCT to MVCT slice, denoted as $\hat{\mathbf{X}}_\mathbf{M} \in \mathbb{R}^{d \times d}$.

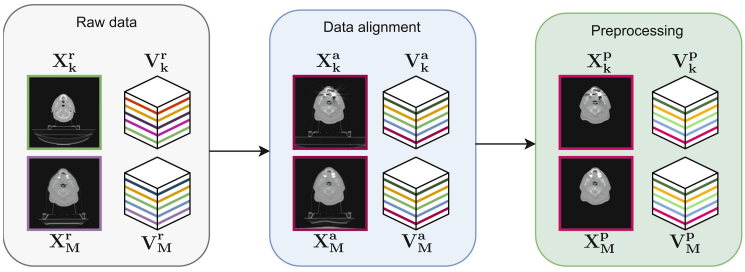


Fig. 2. Steps followed for dataset generation. We start with raw and unaligned kVCT and MVCT volumes –slices (lines in the cube) do not correspond. Then, volumes are pixel-aligned and so the slices correspond (Section 2.2). Finally, corresponding slices in kVCT and MVCT volumes are normalized and masked (Section 2.2).

Network architectures We propose a Metal Artifact Reduction using Domain Transformation Network (MAR-DTN), which closely aligns with the architectural principles of the UNet framework [22]. The UNet architecture has been widely used in previous works for pixel-to-pixel image tasks. In medical imaging, specifically, it has demonstrated excellent results in segmentation, denoising, and MAR [22,23,25]. The detailed architectural explanation of our proposed model, named MAR-DTN, can be found in the supplementary material (Section 1: Explanation of Proposed Model Architecture).

Our investigation involves a comparative analysis of the performance of MAR-DTN against three contemporary state-of-the-art methods. The first one is a Conditional Generative Adversarial Network (cGAN), named pix2pix [7]. Additionally, a modification of this network is included, replacing the original generator with MAR-DTN (referred to as custom-pix2pix).

In addition, a network leveraging transformers is implemented due to their demonstrated high performance in addressing pixel-to-pixel image tasks. The SwinIR architecture [13] is structured into three key components: shallow feature extraction, deep feature extraction, and high-quality image reconstruction. Notably, the deep feature extraction module integrates numerous residual Swin Transformer blocks (RSTB), each incorporating multiple Swin Transformer layers alongside a residual connection.

Finally, an architecture initially designed for medical image segmentation is included. This is the INet architecture, a network that does not perform downsampling. It simply enlarges receptive fields by increasing the kernel sizes of convolutional layers in steps (e.g., from 3×3 to 7×7 and then 15×15). In our case, the final activation is not performed in order to obtain a network capable of generating images. We used this architecture for image generation because INet maintains spatial information by fixing the sizes of feature maps and fuses multilevel semantics by concatenating feature maps of all preceding layers. This allows INet to enhance optimization capabilities.

Loss functions In addressing artifact reduction with neural networks, various loss functions such as L1, FFL (Focal Frequency Loss), MSE (Mean Squared Error), SSIM (Structural Similarity Index), and MS-SSIM (Multi-Scale Structural Similarity Index) offer distinct advantages.

- The weighted $\mathcal{L}_1^{\mathbf{w}}$ loss function is defined by: $\mathcal{L}_1^{\mathbf{w}} = \|\hat{\mathbf{X}}_{\mathbf{M}} - \mathbf{X}_{\mathbf{M}}^{\mathbf{p}}\|_1 \cdot \mathbf{w}$, where $\mathbf{w} \in \mathbb{R}^{d \times d}$ is the pixel weight. This loss emphasizes the absolute differences between predicted and ground truth values and penalizes outliers, contributing to robust artifact reduction.
- FFL [8] is defined by: $\mathcal{L}_{FFL}^{\beta,\alpha} = \frac{1}{d \cdot d} \sum_{u=0}^{d-1} \sum_{v=0}^{d-1} z(u,v) |F_{\hat{\mathbf{X}}_{\mathbf{M}}}(u,v) - F_{\mathbf{X}_{\mathbf{M}}^{\mathbf{p}}}(u,v)|^2 \cdot \beta$, where $z(u,v) = |F_{\hat{\mathbf{X}}_{\mathbf{M}}}(u,v) - F_{\mathbf{X}_{\mathbf{M}}^{\mathbf{p}}}(u,v)|^\alpha$, $\beta \in \mathbb{R}$ is the weight of spatial frequency, $\alpha \in \mathbb{R}$ is the scaling factor, and $F(u,v)$ is the spatial frequency value at the spectrum coordinate (u,v). This loss focuses on high-frequency artifacts, helps in preserving image details while suppressing artifacts, thus enhancing perceptual quality.
- MSE is defined by $\|\hat{\mathbf{X}}_{\mathbf{M_{ij}}} - \mathbf{X}_{\mathbf{M_{ij}}}\|_2^2$ which measures the average squared distance, and provides simplicity and ease of interpretation.
- SSIM [26], evaluates luminance, contrast and structure, ensuring preservation of perceptual features, making them suitable for maintaining image fidelity during artifact reduction tasks.
- MS-SSIM divides images into multiple scales and computes SSIM for each scale separately. Then, it averages these SSIM values to get a single value representing structural similarity. This method offers a more comprehensive evaluation, considering structural similarity across different resolutions.

Implementation details and evaluation metrics All networks have the same input and output shape, 512×512, corresponding to the size of $\mathbf{X}_{\mathbf{k}}^{\mathbf{p}}$ and $\mathbf{X}_{\mathbf{M}}^{\mathbf{p}}$. Models were optimized using Adam with learning rate and weight decay set

to 0.001 and $5e^{-4}$, respectively. The batch size was set to 4 for all networks except SwinIR for which we used 2 samples (due to computational memory issues). We trained for 20 epochs with early stopping with a patience of 5 epochs. Data augmentation [2] includes horizontal flip with a probability of 0.5 and shift, scale, and rotate with a probability of 0.8 (*shift_limit*= 0.0625, *scale_limit*= 0.1, *rotate_limit*= 5). This introduces variability into our dataset by applying transformation probabilities to alter the dataset in each epoch, thus aiding in the mitigation of data limitation.

Models were trained on an Intel Xeon Server with 188GB of RAM and an Nvidia A100 GPU. We evaluated our methodology using PSNR and SSIM metrics.

3 Experimental Results

3.1 Loss function analysis

The impact of different loss functions, whether used individually or in combination, is analyzed in this study. We excluded the INet network from our evaluation because its performance, as detailed in Section 3.1, is significantly lower compared to the other architectures. Including INet could skew the comparative analysis and potentially introduce biases, thus detracting from a fair assessment of the loss functions' effects on more competitive networks.

First, we explore the impact of using an L_1 loss function with weights ($\mathcal{L}_1^{\mathbf{w}}$) on images containing artifacts. Weight assignment is based on body segmentation provided by clinicians (see Fig. 1b), where $\mathbf{w}[i, j]$ is set to 0.1 outside the body segment and varies within the set $\{1, 25, 50, 100\}$ inside the segment for slices with artifacts. Slices without artifacts maintain a weight of 1 throughout the body segment. Since the only variable is the weight within the body segment, we simplify the notation in the following sections and denote this value as w. Therefore, $\mathcal{L}_1^{100}$ indicates a weight of 100 within the body segment for slices with artifacts.

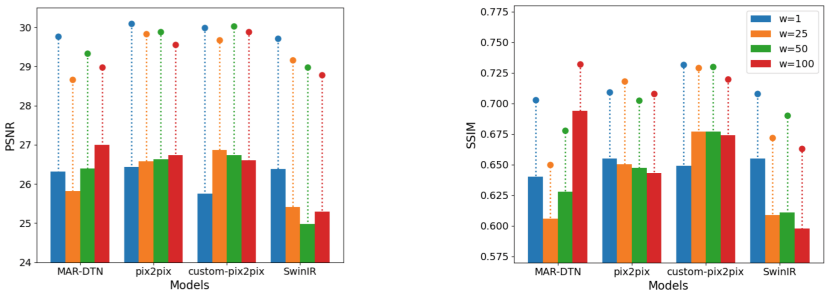
Additionally, the parameters β and α of the $\mathcal{L}_{FFL}^{\beta,\alpha}$ are discretely varied in the set of values $\{0.5, 1, 1.5\}$. This variation allows for exploring different weightings and contributions of both parameters in the neural network's learning process, particularly in handling images with artifacts.

$\mathcal{L}_1^{\mathbf{w}}$ Analysis Fig. 3a and Fig. 3b show the PSNR and SSIM values obtained by the networks of the study after training with $\mathcal{L}_1^w$ loss when $w \in \{1, 25, 50, 100\}$ using $\mathcal{D}_{\text{All}}$.

The first thing to note is the limited variability of results obtained when modifying the parameter w. In terms of PSNR, the results do not vary by more than 3dB, while for SSIM, the results demonstrate a variance of no more than 10%.

In the case of MAR-DTN, a positive trend in the artifact set is observed when $w > 25$ increases. Conversely, with SwinIR, better results are achieved when no supplementary weight is allocated to the artifact class. Moreover, when $w > 1$ parameters have similar results. For pix2pix, no significant difference in results is observed, and the same holds true for custom-pix2pix when $w > 1$. However, increasing the value of w for the artifact class does lead to an improvement in the artifact set results.

Regarding the results in the $\mathcal{D}_{\text{All}}$, represented by dots in Fig. 3a and Fig. 3b, we observe slightly inferior results when giving more weight to the $\mathcal{D}_{\text{Art}}$ set. However, we are not concerned as the focus is on the artifact region. Therefore, for the remaining experiments, we will use $w = 100$, as it yields the best result for MAR-DTN and similar values to the state-of-the-art L_1 for the rest of the networks.



(a) (b)

Fig. 3. PSNR (a) and SSIM (b) values evaluated on the $\mathcal{D}_{\text{All}}$. The dots represent the mean value of all slices in the dataset, while the bars represent the mean value of slices with artifacts. Values obtained using the four considered networks (MAR-DTN, pix2pix, custom-pix2pix and SwinIR) trained on the $\mathcal{D}_{\text{All}}$ with the $\mathcal{L}_1^w$ loss function only.

$\boldsymbol{\mathcal{L}_{FFL}^{\beta,\alpha}}$ **Analysis** β and α were varied within the set $\{0.5, 1, 1.5\}$. The average value of the metrics evaluated on the $\mathcal{D}_{\text{All}}^{Ts}$ set can be seen in Fig. 4a and Fig. 4b.

As with the prior study, we observe some variability in the results with less than 2dB in PSNR and 10% in SSIM.

However, it is observed that increasing the value of alpha decreases the metrics. For the minimum α value, $\alpha = 0.5$, the best PSNR result is for $\beta = 1$, with a PSNR value of 27.81dB. On the other hand, the mean SSIM value when $alpha = 0.5$ and $\beta = 1$ is 0.64, very close to the best value, which is 0.65. Taking this into account, we conclude that the best combination of values is $\beta = 1$ and $\alpha = 0.5$.

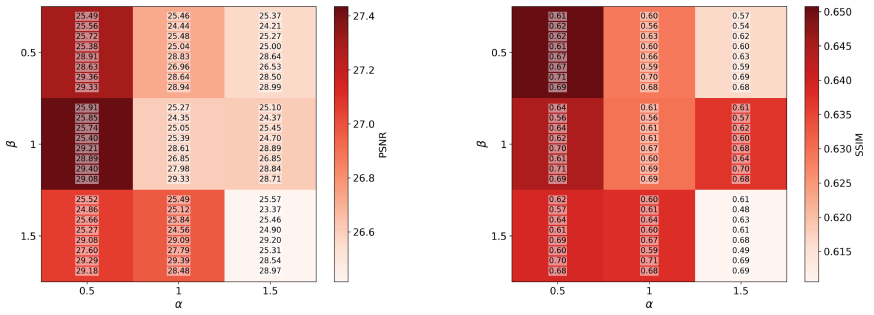


Fig. 4. Heatmaps with the mean values of PSNR (a) and SSIM (b) evaluated on the test dataset after training the networks using the $\mathcal{L}_{FFL}^{\beta,\alpha}$ loss function with various combinations of the parameters α and β (x and y-axis, respectively). Each cell represents the mean of 8 values, the first 4 corresponding to the parameter value evaluated on $\mathcal{D}_{\text{Art}}^{Ts}$, and the last 4 corresponding to the parameter value evaluated on the $\mathcal{D}_{\text{All}}^{Ts}$, for each neural network in the study, MAR-DTN, pix2pix, custom-pix2pix, and SwinIR, respectively.

Loss function comparison Table 2 compares the results obtained considering different loss functions combinations and datasets.

The $\mathcal{L}_1^{100}$ loss function achieves the best results on $\mathcal{D}_{\text{Art}}$ for MAR-DTN, with a PSNR of 27.17dB, and it is the second best result for pix2pix, with a PSNR of 26.31dB. However, for both custom-pix2pix and SwinIR, the performance is reduced by almost 2dB.

Also the $\mathcal{L}_{FFL}^{1,0.5}$ loss function has been tested alone. It yields less accurate results than $\mathcal{L}_1^{100}$, decreasing the PSNR value by up to 2dB for pix2pix. However, custom-pix2pix maintains a PSNR value of 26.15dB, competitive with the rest of the loss functions. Regarding the rest of the loss functions, the improvement in using $\mathcal{L}_{MS-SSIM}$ instead of $\mathcal{L}_{SSIM}$ stands out, especially notable in custom-pix2pix. The most complex loss function ($\mathcal{L}_1^{100} + \mathcal{L}_{MS-SSIM} + \mathcal{L}_{FFL}^{1,0.5}$) introduces noise during training and fails to surpass the metric value achieved by simpler functions. Nonetheless, SwinIR achieves the best result, with a PSNR of 26.42dB. An example of the reconstruction of a slice with artifacts can be seen in Fig. 5.

However, the results in the $\mathcal{D}_{\text{All}}$ improve when other loss functions are added to $\mathcal{L}_1^{100}$. In the case of MAR-DTN and custom-pix2pix, the best combination is $\mathcal{L}_1^{100} + \mathcal{L}_{SSIM}$, reaching a PSNR value of 30.02dB. For pix2pix and SwinIR, the best combination is $\mathcal{L}_1^{100} + \mathcal{L}_{MSE}$ with a PSNR of 28.92 and 29.39dB, respectively.

In general, the metric values for slices with artifacts are lower when trained on $\mathcal{D}_{\text{All}}$. This is consistent with having an unbalanced dataset and means that the loss functions are not entirely capable of addressing the issue of class imbalance.

On the other hand, the results obtained with INet can be found in the last column of Table 2. The highest PSNR achieved is 12.67dB for the artifact set,

Table 2. Comparative analysis for different networks and loss function combinations, indicated with a check mark which sum of loss functions have been used for training. For the Pix2Pix networks, it indicates the loss function of the generator. The dataset column indicates the dataset with which the network has been trained and evaluated; where dataset is $\mathcal{D}_{\text{All}}$ then model is trained on $\mathcal{D}_{\text{All}}^{Tr}$ and tested on $\mathcal{D}_{\text{All}}^{Ts}$, and in case of $\mathcal{D}_{\text{Art}}$ then model is trained on $\mathcal{D}_{\text{Art}}^{Tr}$, and tested on $\mathcal{D}_{\text{Art}}^{Ts}$. Finally, the remaining columns show the PSNR and SSIM values obtained for the test sets. Where the dataset is the $\mathcal{D}_{\text{All}}$, we report both on the performance obtained on artifact slices from within the $\mathcal{D}_{\text{Art}}^{Ts}$, and the mean of PSNR and SSIM on whole dataset $\mathcal{D}_{\text{All}}^{Ts}$ (in parentheses). Underlined values indicate the highest performance for each network with certain loss function combinations, while highlighted values indicate the highest overall performing model across all configurations.

Loss combination					Dataset	MAR-DTN		pix2pix [7]		custom-pix2pix		SwinIR [13]		INet [27]	
$\mathcal{L}_1^{100}$	$\mathcal{L}_{SSIM}$	$\mathcal{L}_{MS\text{-}SSIM}$	$\mathcal{L}_{MSE}$	$\mathcal{L}_{FFL}^{1,0.5}$		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
✓					$\mathcal{D}_{\text{Art}}$	27.17	0.69	26.31	0.64	25.24	0.68	25.46	0.61	11.61	0.04
					$\mathcal{D}_{\text{All}}$	26.99 (28.97)	0.71 (0.73)	26.36 (28.7)	0.63 (0.69)	26.61 (29.08)	0.67 (0.7)	25.29 (28.79)	0.59 (0.66)	12.02 (12.03)	0.04 (0.04)
✓	✓				$\mathcal{D}_{\text{Art}}$	27.11	0.69	26.21	0.63	26.98	0.71	26.16	0.62	10.94	0.08
					$\mathcal{D}_{\text{All}}$	27.09 (30.02)	0.69 (0.73)	26.39 (28.58)	0.64 (0.68)	27.13 (29.85)	0.68 (0.70)	24.90 (28.97)	0.68 (0.68)	12.01 (12.27)	0.03 (0.03)
✓		✓			$\mathcal{D}_{\text{Art}}$	27.46	0.69	26.32	0.65	27.06	0.67	26.26	0.63	11.96	0.01
					$\mathcal{D}_{\text{All}}$	27.08 (29.97)	0.69 (0.73)	26.37 (28.69)	0.58 (0.68)	27.04 (29.35)	0.64 (0.70)	26.25 (29.39)	0.63 (0.67)	12.67 (12.95)	0.05 (0.04)
✓			✓		$\mathcal{D}_{\text{Art}}$	26.94	0.68	26.25	0.64	26.55	0.64	26.18	0.63	11.70	0.02
					$\mathcal{D}_{\text{All}}$	27.11 (29.89)	0.69 (0.72)	26.42 (28.92)	0.68 (0.7)	26.98 (29.22)	0.64 (0.68)	25.24 (29.00)	0.60 (0.66)	11.23 (10.96)	0.01 (0.01)
				✓	$\mathcal{D}_{\text{Art}}$	26.35	0.64	24.03	0.51	26.15	0.61	24.58	0.59	9.05	0.08
					$\mathcal{D}_{\text{All}}$	26.52 (29.51)	0.66 (0.70)	25.85 (28.88)	0.56 (0.61)	26.02 (29.04)	0.60 (0.69)	25.39 (29.33)	0.61 (0.69)	10.03 (10.01)	0.02 (0.01)
✓				✓	$\mathcal{D}_{\text{Art}}$	27.06	0.69	25.66	0.63	26.66	0.60	25.40	0.61	11.95	0.03
					$\mathcal{D}_{\text{All}}$	26.99 (29.85)	0.69 (0.72)	25.99 (28.56)	0.59 (0.65)	26.48 (29.06)	0.62 (0.69)	25.55 (29.18)	0.60 (0.68)	11.54 (11.86)	0.08 (0.07)
✓	✓			✓	$\mathcal{D}_{\text{Art}}$	27.08	0.69	25.66	0.63	26.18	0.56	26.42	0.64	9.58	0.05
					$\mathcal{D}_{\text{All}}$	25.65 (28.65)	0.63 (0.69)	26.41 (28.74)	0.68 (0.64)	27.08 (28.04)	0.64 (0.68)	25.45 (28.79)	0.61 (0.66)	12.53 (12.21)	0.06 (0.05)

highlighting the architecture's inability to compete with the other architectures in the study. For the SSIM, we observe a similar behavior, with a maximum value of 8%. In Fig. 5, the results obtained by INet can be seen. The artifacts not only have not been reduced, but they also acquire higher contrast, along with the rest of the image. However, new artifacts appear, which blur and deform other structures; for example, noticeable in the gird where streaks obtained with a combination of loss functions $\mathcal{L}_1^{100} + \mathcal{L}_{MS-SSIM} + \mathcal{L}_{FFL}^{1,0.5}$.

3.2 State-of-the-Art Comparison

The comparison between MAR-DTN and state-of-the-art networks (pix2pix, custom-pix2pix, and SwingIR using $\mathcal{L}_1^1$ as loss function) shows MAR-DTN achieves the best result for the $\mathcal{D}_{\text{Art}}$, with a PSNR of 26.99dB and an SSIM of 0.69 points. However, custom-pix2pix and MAR-DTN achieve the best overall result for the $\mathcal{D}_{\text{All}}$, with a PSNR of 29.88dB and an SSIM of 0.73 points for pix2pix. SwinIR exhibits a decrease of up to 0.76dB for PSNR and 0.07 points for SSIM across the $\mathcal{D}_{\text{All}}$, with a larger decrease observed within the $\mathcal{D}_{\text{Art}}$, reaching a difference of 1.7dB compared to MAR-DTN. Table 3 presents a comparison between the performance time and complexity of the networks.

Table 3. Comparison of trainable parameters, number of multiplications and additions (MACs), training time computed for the $\mathcal{D}_{\text{All}}$ in 1 epoch and patient reconstruction time (in this case 170 slices) for state-of-the-art methods under study.

Network	Parameters (M)	MACs (G)	Training time (s)	Patient reconstruction time (s)
MAR-DTN	1.882	116.686	65.32	3.56
pix2pix	54.413	77.99	80.02	3.75
custom-pix2pix	4.646	123.277	67.42	4.25
SwinIR	1.614	425.034	2,774.76	47.27
INet	2.96	896.31	807.38	5.31

3.3 Clinical evaluation

Initial feedback from clinicians indicates that the quality of the MVCT images generated through our proposed method is highly regarded. Clinicians have noted that synthetic MVCT images exhibit excellent contrast for both soft tissues and bones, which is essential for accurate diagnosis and treatment planning in clinical practice. These qualitative observations suggest promising outcomes in terms of image quality and clinical utility, laying a strong foundation for further quantitative evaluation and validation studies in the future.

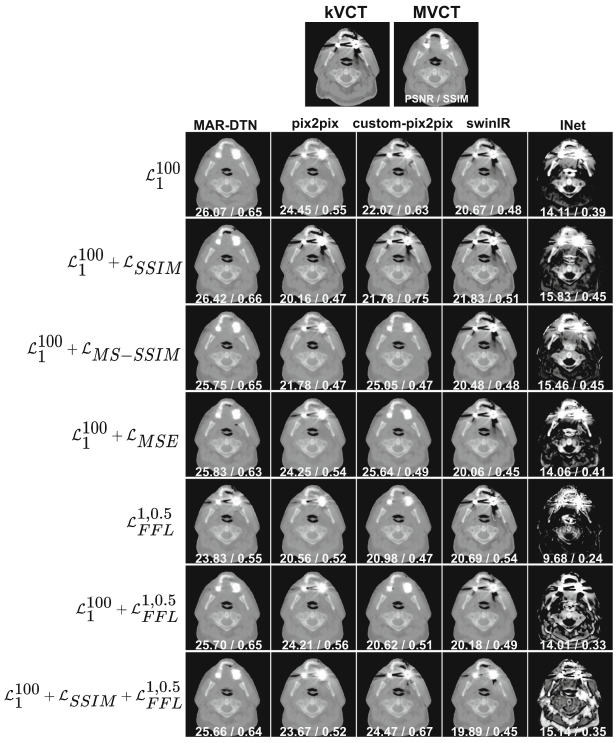


Fig. 5. Reconstruction of a slice with artifacts by the different models and loss functions. First row shows preprocessed kVCT and MVCT images (ground truth). First column indicates the loss function, and the following ones indicate the model used. Networks have been trained on the $\mathcal{D}_{\text{Art}}$.

4 Discussion and Conclusion

In this study, we compared our proposed domain transformation methodology with some state-of-the-art methods, where kVCT images serve as input and MVCT images as output. Our results demonstrate that a lightweight model like MAR-DTN can effectively reduce artifacts with the appropriate combination of loss functions, even with a reasonable dataset size. The performance of the models is evaluated on two datasets: $\mathcal{D}_{\text{Art}}$, which contains only images with artifacts, and $\mathcal{D}_{\text{All}}$ includes both artifact-affected and non-affected images.

Numerous combinations of loss functions were tested, though only a select few are presented in Table 2 due to space constraints. Consequently, a deliberate choice was made to include those combinations yielding more promising results within the allocated space.

As we compare the performance of models trained on $\mathcal{D}_{\text{All}}$, MAR-DTN shows the best performance in several cases, especially with the combination of $\mathcal{L}_1^{100} + \mathcal{L}_{SSIM}$ when tested on $\mathcal{D}_{\text{All}}^{Ts}$, achieving the highest PSNR of 30.02 dB and a high

SSIM of 0.73 on over all patient volume, in addition to that when tested on $\mathcal{D}_{\text{Art}}^{Ts}$ still achieves competitive results. Model pix2pix and custom-pix2pix show similar performance, with custom-pix2pix slightly outperforming pix2pix in most cases. custom-pix2pix performs best on $\mathcal{D}_{\text{All}}$ with the loss combination of $\mathcal{L}_1^{100}+\mathcal{L}_{SSIM}$ and pix2pix show fair performance using $\mathcal{L}_1^{100}+\mathcal{L}_{MSE}$. SwinIR exhibits decent performance but is generally outperformed by MAR-DTN, particularly in terms of PSNR. However, it shows competitive SSIM values. Model INet performs the worst among all models, with significantly lower PSNR (max 12.67 dB) and SSIM (max 0.08) values, highlighting its inability to effectively reduce artifacts or maintain structural similarity.

Furthermore, as we compare the performance of models trained on $\mathcal{D}_{\text{Art}}$, MAR-DTN achieves the highest performance on this dataset with the combination of $\mathcal{L}_1^{100}+\mathcal{L}_{MS-SSIM}$, achieving a PSNR of 27.46 dB and an SSIM of 0.69 when tested on $\mathcal{D}_{\text{Art}}^{Ts}$. Overall, MAR-DTN performs better than all other models across various loss combinations, particularly on the $\mathcal{D}_{\text{All}}$ dataset. Model pix2pix and custom-pix2pix show similar PSNR and SSIM values, typically around 26-27 dB for PSNR and 0.64-0.68 for SSIM, depending on the loss function combination used. Model custom-pix2pix slightly outperforms pix2pix in most of the combinations. Model SwinIR performs reasonably well, achieving PSNR values around 25-26 dB and SSIM values around 0.64-0.67, depending on the loss function combination but it is outperformed by MAR-DTN in most combinations. INet shows the poorest performance on $\mathcal{D}_{\text{Art}}$. We can conclude that it is not capable of eliminating artifacts using the loss functions in this study, falling far behind its competitors. What is achieved, however, is an increase in contrast between different bone and muscle structures. Nevertheless, it also introduces new artifacts, which hinder the correct evaluation of the images. It is important to note that INet's initial goal is image segmentation, not image generation. Additionally, INet performs better with low-resolution images, making it less appropriate for our dataset.

Despite achieving satisfactory results, it is worth noting that our study's considered dataset is relatively small. Nevertheless, our approach demonstrates significant potential, as evidenced by MAR-DTN's robust performance metrics achieved across various network architectures and loss functions. To further enhance the impact of our findings, we plan to incorporate systematic qualitative evaluations by clinical staff. Additionally, we aim to expand our dataset to include more types of artifacts across different body regions. These steps will provide deeper insights and potentially lead to even more improved results, reinforcing the efficacy and applicability of our methodology in broader contexts. Moreover, our future work aims to develop a generalized model for the entire body. This extension will significantly broaden the applicability and robustness of our approach, paving the way for more comprehensive and versatile artifact management in medical imaging.

Acknowledgements. This work is supported by the Italian Ministry of Health (Ricerca Corrente). APM and BSA acknowledge financial support by the Spanish Ministerio de Economía y Competitividad and European Regional Development Fund,

MCIN/AEI/ 10.13039/501100011033 and by "ERDF A way of making Europe". Xunta de Galicia funded research under Research Grant No. 2021-PG036 and the Spanish Ministerio de Ciencia e Innovación MCIN/AEI/10.13039/501100011033 through the Industrial Doctorates Grant.

References

1. Boas, F.E., Fleischmann, D.: Ct artifacts: causes and reduction techniques. Imaging in Medicine **4**(2), 229–240 (04 2012), copyright - © 2012 Future Medicine Ltd
2. Buslaev, A., Iglovikov, V.I., Khvedchenya, E., Parinov, A., Druzhinin, M., Kalinin, A.A.: Albumentations: Fast and flexible image augmentations. Information **11**(2), 125 (2020)
3. Fedorov, A., Beichel, R., Kalpathy-Cramer, J., Finet, J., Fillion-Robin, J.C., Pujol, S., Bauer, C., Jennings, D., Fennessy, F., Sonka, M., et al.: 3d slicer as an image computing platform for the quantitative imaging network. Magn. Reson. Imaging **30**(9), 1323–1341 (2012)
4. Gjesteby, L., Shan, H., Yang, Q., Xi, Y., Jin, Y., Giantsoudi, D., Paganetti, H., De Man, B., Wang, G.: A dual-stream deep convolutional network for reducing metal streak artifacts in ct images. Physics in Medicine & Biology **64**(23), 235003 (2019)
5. Hounsfield, G.N.: Computerized transverse axial scanning (tomography): Part 1. Description of system. British Journal of Radiology **46**(552), 1016–1022 (01 2014)
6. Huang, X., Wang, J., Tang, F., Zhong, T., Zhang, Y.: Metal artifact reduction on cervical ct images by deep residual learning. Biomed. Eng. Online **17**, 1–15 (2018)
7. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5967–5976 (2017)
8. Jiang, L., Dai, B., Wu, W., Loy, C.C.: Focal frequency loss for image reconstruction and synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13919–13929 (2021)
9. Kaposi, P., Youn, T., Tóth, A., Frank, V., Shariati, S., Szendrõi, A., Magyar, P., Bérczi, V.: Orthopaedic metallic artefact reduction algorithm facilitates ct evaluation of the urinary tract after hip prosthesis. Clin. Radiol. **75**(1), 78.e17-78.e24 (2020)
10. Kim, H., Yoo, S.K., Kim, D.W., Lee, H., Hong, C.S., Han, M.C., Kim, J.S.: Metal artifact reduction in kv ct images throughout two-step sequential deep convolutional neural networks by combining multi-modal imaging (martian). Sci. Rep. **12**(1), 20823 (2022)
11. Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P.: Elastix: a toolbox for intensity-based medical image registration. IEEE Trans. Med. Imaging **29**(1), 196–205 (2009)
12. Lee, D., Park, C., Lim, Y., Cho, H.: A metal artifact reduction method using a fully convolutional network in the sinogram and image domains for dental computed tomography. J. Digital Imaging **33**(2), 538–546 (2020)
13. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1833–1844 (2021)
14. Liang, K., Zhang, L., Yang, H., Yang, Y., Chen, Z., Xing, Y.: Metal artifact reduction for practical dental computed tomography by improving interpolation-based reconstruction with deep learning. Med. Phys. **46**(12), e823–e834 (2019)

15. Lin, W.A., Liao, H., Peng, C., Sun, X., Zhang, J., Luo, J., Chellappa, R., Zhou, S.K.: Dudonet: Dual domain network for ct metal artifact reduction. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10504–10513 (2019)
16. Liugang, G., Hongfei, S., Xinye, N., Mingming, F., Zheng, C., Tao, L.: Metal artifact reduction through mvcbct and kvct in radiotherapy. Sci. Rep. **6**(1), 37608 (2016)
17. Lossau (née Elss), T., Nickisch, H., Wissel, T., Morlock, M., Grass, M.: Learning metal artifact reduction in cardiac ct images with moving pacemakers. Medical Image Analysis **61**, 101655 (2020)
18. Lyu, Y., Lin, W.-A., Liao, H., Lu, J., Zhou, S.K.: Encoding Metal Mask Projection for Metal Artifact Reduction in Computed Tomography. In: Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racoceanu, D., Joskowicz, L. (eds.) MICCAI 2020. LNCS, vol. 12262, pp. 147–157. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59713-9_15
19. Ni, X., Shi, Z., Song, X., Tang, T., Li, S., Hou, Z., Zhang, W., Wang, W.F., Chen, F., Li, J., et al.: Metal artifacts reduction in kv-ct images with polymetallic dentures and complex metals based on mv-cbct images in radiotherapy. Sci. Rep. **13**(1), 8970 (2023)
20. Park, H.S., Lee, S.M., Kim, H.P., Seo, J.K., Chung, Y.E.: Ct sinogram-consistency learning for metal-induced beam hardening correction. Med. Phys. **45**(12), 5376–5384 (2018)
21. Paudel, M.R., Mackenzie, M., Fallone, B.G., Rathee, S.: Clinical evaluation of normalized metal artifact reduction in kvct using mvct prior images (mvct-nmar) for radiation therapy treatment planning. International Journal of Radiation Oncology* Biology* Physics **89**(3), 682–689 (2014)
22. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
23. Serrano-Antón, B., Otero-Cacho, A., López-Otero, D., Díaz-Fernández, B., Bastos-Fernández, M., Pérez-Muñuzuri, V., González-Juanatey, J.R., Muñuzuri, A.P.: Coronary artery segmentation based on transfer learning and unet architecture on computed tomography coronary angiography images. IEEE Access (2023)
24. Wang, H., Li, Y., He, N., Ma, K., Meng, D., Zheng, Y.: Dicdnet: deep interpretable convolutional dictionary network for metal artifact reduction in ct images. IEEE Trans. Med. Imaging **41**(4), 869–880 (2021)
25. Wang, J., Zhao, Y., Noble, J.H., Dawant, B.M.: Conditional Generative Adversarial Networks for Metal Artifact Reduction in CT Images of the Ear. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11070, pp. 3–11. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00928-1_1
26. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
27. Weng, W., Zhu, X.: Inet: convolutional networks for biomedical image segmentation. Ieee Access **9**, 16591–16603 (2021)
28. Yu, L., Zhang, Z., Li, X., Ren, H., Zhao, W., Xing, L.: Metal artifact reduction in 2d ct images with self-supervised cross-domain learning. Physics in Medicine & Biology **66**(17), 175003 (2021)

29. Zhang, Y., Yu, H.: Convolutional neural network based metal artifact reduction in x-ray computed tomography. IEEE Trans. Med. Imaging **37**(6), 1370–1381 (2018)
30. Zhu, L., Han, Y., Xi, X., Li, L., Yan, B.: Completion of metal-damaged traces based on deep learning in sinogram domain for metal artifacts reduction in ct images. Sensors **21**(24) (2021)

Self-prior Guided Mamba-UNet Networks for Medical Image Super-Resolution

Zexin Ji[1,2,4], Beiji Zou[1,2], Xiaoyan Kui[1,2(✉)], Pierre Vera[4], and Su Ruan[3]

[1] School of Computer Science and Engineering, Central South University, Changsha 410083, China
[2] Hunan Engineering Research Center of Machine Vision and Intelligent Medicine, Central South University, Changsha 410083, China
xykui@csu.edu.cn
[3] University of Rouen-Normandy, LITIS - QuantIF UR 4108, 76000 Rouen, France
[4] Department of Nuclear Medicine, Henri Becquerel Cancer Center, Rouen, France

Abstract. In this paper, we propose a self-prior guided Mamba-UNet network (SMamba-UNet) for medical image super-resolution. Existing methods are primarily based on convolutional neural networks (CNNs) or Transformers. CNNs-based methods fail to capture long-range dependencies, while Transformer-based approaches face heavy calculation challenges due to their quadratic computational complexity. Recently, State Space Models (SSMs) especially Mamba have emerged, capable of modeling long-range dependencies with linear computational complexity. Inspired by Mamba, our approach aims to learn the self-prior multi-scale contextual features under Mamba-UNet networks, which may help to super-resolve low-resolution medical images in an efficient way. Specifically, we obtain self-priors by perturbing the brightness inpainting of the input image during network training, which can learn detailed texture and brightness information that is beneficial for super-resolution. Furthermore, we combine Mamba with Unet network to mine global features at different levels. We also design an improved 2D-Selective-Scan (ISS2D) module to divide image features into different directional sequences to learn long-range dependencies in multiple directions, and adaptively fuse sequence information to enhance super-resolved feature representation. Both qualitative and quantitative experimental results demonstrate that our approach outperforms current state-of-the-art methods on two public medical datasets: the IXI and fastMRI.

Keywords: Medical imaging · Super-Resolution · State space models · Mamba · Unet · Deep learning.

1 Introduction

Medical imaging techniques play a vital role in supporting clinical diagnoses. However, obtaining high-quality medical images demands high-quality medical imaging equipment and a longer imaging time. With the advancement of deep

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15311, pp. 160–174, 2025.
https://doi.org/10.1007/978-3-031-78195-7_11

learning, super-resolution (SR) is considered as a promising direction for cost-saving improvements in image quality. Medical image super-resolution aims to restore the corresponding high-resolution images by adding missing details in low-resolution (LR) medical images.

To realize medical image super-resolution, initial research utilized methods of interpolation [16] and optimization [25]. These methods are simple to implement, but not precise enough in restoring image details. Subsequently, massive CNNs-based image super-resolution methods [2,18,21,22] have come up to learn the nonlinear mapping relationship and representative features. It demonstrates superior performance in the field of medical imaging. However, they are prone to local matching difficulties. This occurrence is credited to the inductive bias of CNNs, which limits their ability to capture long-range dependencies. Compared to CNNs, Vision Transformers [3] have emerged for modeling non-local dependencies. Leveraging the ability of self-attention of Transformer, researchers have designed tailored approaches to achieve super-resolution [5,10,12,33]. However, Transformer-based methods face excessive computational complexity and memory requirements due to the attention calculation of the pairwise affinity. Moreover, these methods are still insufficient to fully mine the precise features hidden within complex medical image distributions. Mamba [7] is now widely applied as an emerging sequence model, originating from tasks in natural image processing. It has been proposed for learning visual representations that can efficiently capture long-distance dependencies of images.

Motivated by these insights, we propose a self-prior guided Mamba-UNet network for medical image super-resolution, which taps into the potential of lightweight long-range modeling and fully mining super-resolved features. To achieve this, we design a self-prior learning in a super-resolution network. Specifically, the input is disturbed by a brightness inpainting in the feature distribution during training, leading to an incompleteness of the input. This can enhance the ability of the model to mine the own super-resolved feature information. We also develop a Mamba-based Unet to learn the multi-scale contextual dependencies of images from hierarchical levels. Mamba can calculate long-range dependencies using linear complexity. Compared with traditional 2D-Selective-Scan in vision Mamba, we further design the improved 2D-Selective-Scan (ISS2D) model to adaptively integrate relevant information between image sequences from different directions, thereby alleviating the lack of causal reasoning in the image field.

The contributions of our approach can be summarized as follows:

1) We first build a Mamba network for medical image super-resolution. With the designed Mamba-based Unet network, we can simultaneously learn the long-range dependencies of features at different scales.
2) We design self-prior learning to improve the local texture and brightness refinement by borrowing self-reasoning features, which can significantly strengthen the generation of semantically coherent features.
3) We devise an improved 2D-Selective-Scan (ISS2D) module to dynamically model the correlation between features of four direction sequences in images. It can better fuse four direction weighted features for image super-resolution.

4) Qualitative and qua1ntitative experiment results on IXI and fastMRI datasets demonstrate that our approach achieves superior performance comparing to the existing state-of-the-art super-resolution methods.

2 Related Work

2.1 Medical Image Super-Resolution

The basic CNNs have been proposed to deal with super-resolution. SRCNN [2] first introduced deep CNNs into super-resolution and achieved promising results. Then, many improved methods based on SRCNN have emerged, and significant progress has been made in medical image super-resolution tasks [13,14,18,21, 22,27]. Specifically, Qiu *et al.* proposed EMISR [22] based on the improvement of SRCNN for the knee MR image super-resolution. Liu *et al.* [18] developed an edge-enhanced super-resolution generative adversarial networks (EE-SRGAN) for medical image super-resolution. A dual U-Net residual network (DURN) [21] was designed to enhance cardiac MR image resolution. These CNNs-based approaches primarily capture local patterns of images due to the nature of convolution operations, lacking of ability to model long-range dependencies. Thanks to the long-range representation ability of self-attention in Transformer, it has shown superior performance compared to CNNs-based methods. Dosovitskity *et al.* [3] firstly proposed Vision Transformer (ViT) for computer vision task, such as medical image classification [24] and segmentation [11] tasks. The same holds in the area of medical image super-resolution [5,10,12,33]. SuperFormer [5] explored the effectiveness of a swim transformer in the 3D MR image super-resolution. The TransMRSR [10] further combined the local information of CNNs and the global information of Transformers for medical image super-resolution. Although these Transformer-based methods have achieved superior results, they require a large amount of GPU memory to train the model.

2.2 State Space Models

Recently, State Space Models (SSMs) [9], with roots in classical control theory, have made their way into deep learning, showcasing potential as an effective architecture for modeling sequences. These models blend the characteristics of recurrent neural networks (RNNs) and CNNs, establishing a novel approach to sequence analysis. Compared with Transformers with large parameter scales, SSMs have good characteristics of linear scaling of sequence length and modeling long-range dependencies. The Structured State Space Sequence Model named as S4 [8] is a pioneering work in deep state space models. It models in a content-agnostic static representation. Recently, a new SSM architecture named Mamba [7] was proposed based on S4. The authors designed a selective scan space state sequential model (S6) that is a data-related SSM with selective mechanisms and efficient hardware design. The breakthrough of Mamba in sequence modeling has aroused great interest in the field of computer vision [32].

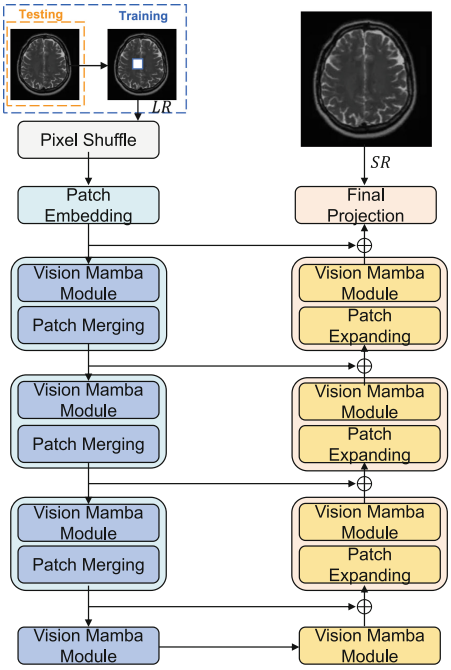


Fig. 1. Our SMamba-Unet framework primarily comprising a patch embedding layer, a Mamba-based encoder module, a Mamba-based decoder module, and a final projection layer.

Some works have applied Mamba as a sequence model backbone to visual tasks, such as medical image segmentation [23,30], video understanding [28], classification [19] and video understanding [28]. In this work, we first explore the potential of Mamba for medical image super-resolution.

3 Method

3.1 Overview

The core structure of our SMamba-Unet is depicted in Figure 1. Before being fed into the network, we perturb the image based on a central random brightness inpainting, which enables the network to mine its own prior information for image super-resolution. Then, we employ a pixel shuffle technique for learnable upscaling of low-resolution images, aiming to preserve the integrity of image information. Traditional upscaling methods like bicubic interpolation enlarge image pixels, often losing sharpness of organ contours and texture detail. In contrast, the pixel shuffle technique in neural networks rearranges multiple channel outputs into a higher-resolution image, shuffling pixels into the correct spatial order to better recover contours and textures. Our method primarily comprises patch embedding layer, a Mamba-based encoder module, a Mamba-based

decoder module, and a final projection layer. In detail, the input image is initially split into non-overlapping patches via patch embedding layer. The Mamba-based encoder mainly contains four layers. Each layer is equipped with the vision Mamba module and patch merging, which can extract contextual features and further increase the number of channels, respectively. Within the Mamba-based decoder, each stage progressively merges information from skip connections to upscale features while simultaneously decreasing the channel count by the vision Mamba module and patch expanding. Finally, the super-resolved medical image is generated by the final projection layer.

3.2 Self-prior Learning

From the error map shown in Figure 3 and Figure 4, it can be seen that the main difference between the super-resolved image and the high-quality image lies in the texture areas with significant brightness changes. It has been confirmed that the repetitive characteristics of an image serve as an effective prior for image super-resolution [26]. Therefore, we propose self-prior learning to mine valuable repetitive internal examples of super-resolved images to enhance the learning of texture and brightness feature representations in perturbed images. Unlike natural scenes, medical images contain a large amount of invalid noise background, while physicians mainly focus on the foreground areas with rich information in the middle of medical images where the organs of interests are presented. Therefore, we first define the width and height of the area of interest in the image, and the starting coordinates of the area in the image. Then, we randomly select a position within the region of interest to add a 5×5 brightness block to perturb the image, and enhance the learning ability of texture and brightness features in the image through self-reasoning inpainting. During our testing, we input the original low-resolution medical images to obtain high-quality images.

3.3 Vision Mamba Module

Before introducing Mamba [7], let's first revisit the Transformer. The Transformer views any text input as a sequence of tokens. It creates a self-attention matrix that compares each token with every other token. The values in the matrix encode the correlations between them. Generating a self-attention matrix for a sequence of length L requires approximately L^2 computations, which is quite computationally intensive. Currently, Mamba has emerged as a promising approach, which can be trained in parallel, while still performing inference that scales linearly with the length of the sequence. It is utilized to characterize the state representations and to forecast their subsequent states based on given inputs. It transforms a 1-D function or sequence $x(t) \in \mathbb{R}$ into the output $y(t) \in \mathbb{R}$ via a hidden state $h(t) \in \mathbb{R}^{\mathbb{N}}$, which usually realizes through linear ordinary differential equations (ODEs).

$$h'(t) = \mathbf{A}h(t) + \mathbf{B}x(t), y(t) = \mathbf{C}h(t), \tag{1}$$

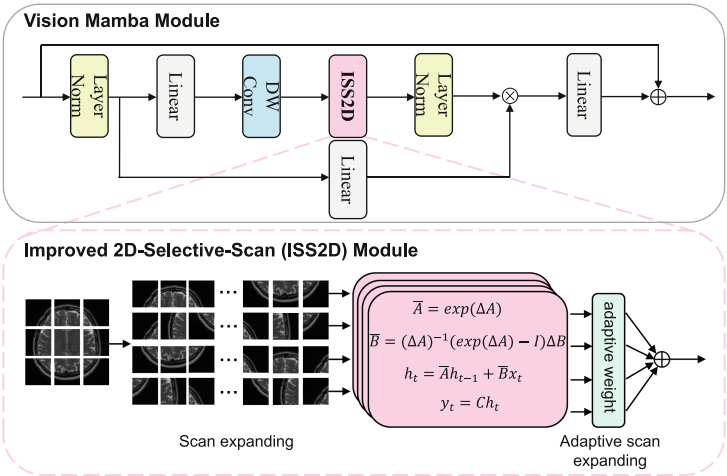


Fig. 2. The framework of the vision Mamba module and improved 2D selective scan (ISS2D) module.

where $\mathbf{A} \in \mathbb{R}^{\mathrm{N}\times\mathrm{N}}$ is the state matrix. $\mathbf{B} \in \mathbb{R}^{\mathbb{N}\times 1}$ and $\mathbf{C} \in \mathbb{R}^{1\times\mathbb{N}}$ are the projection parameters.

The zero-order hold (ZOH) technique [6] is used for discretizing ODEs into discrete functions, making it better adapted for deep learning contexts. By incorporating a timescale parameter Δ, it facilitates the conversion of the continuous-time system matrices $\mathbf{A}$ and $\mathbf{B}$ into their discrete equivalents, noted as $\overline{\mathbf{A}}$ and $\overline{\mathbf{B}}$. The specific steps involved in this conversion are meticulously designed to preserve the integrity of the original system's dynamics while making them compatible with the discrete computational environment of deep learning models. The discretization process is implemented as follows:

$$\overline{\mathbf{A}} = \exp(\Delta\mathbf{A}), \overline{\mathbf{B}} = (\Delta\mathbf{A})^{-1}(\exp(\Delta\mathbf{A}) - \mathbf{I}) \cdot \Delta\mathbf{B}. \tag{2}$$

Following discretization, Equation (1) adopts a form suitable for discrete-time processing, as follows:

$$h_t = \overline{\mathbf{A}}h_{t-1} + \overline{\mathbf{B}}x_t, y_t = \mathbf{C}h_t. \tag{3}$$

Mamba further presents a groundbreaking method within the realm of SSMs through its introduction of Selective Structured State Space Sequence Models (S6). This advancement permits dynamic parameterization of the SSMs, where the parameters $\overline{\mathbf{B}}$, $\mathbf{C}$ and Δ are directly determined by the input data, facilitating a model adaptation unique to each input. Defined by its linear complexity and further optimized for hardware efficiency, Mamba stands out for its exceptional capability in handling the modeling of lengthy sequences.

Similar to the Transformer, Mamba also processes its input by dividing the image into a series of image patches. Figure 2 shows the specific structure of the

vision Mamba module. Initially, the input patch feature map undergoes layer normalization, after which it splits into two distinct paths. The initial path processes the input via a linear layer. Concurrently, in the alternate path, the input is subjected to a linear layer, then it proceeds through a depthwise separable convolution (DW Conv), then advancing into the improved 2D-Selective-Scan (ISS2D) and layer normalization. Integration of these divergent pathways is achieved through a multiplication process and a linear layer, combining the processed outputs with the initial input to further enhance feature representation.

3.4 Improved 2D-Selective-Scan (ISS2D)

The detailed implementation of improved 2D-Selective-Scan (ISS2D) module is illustrated in Figure 2. Unlike sentences in natural language processing, the patches of the image lack a direct inferential relationship. Therefore, we use scan expanding directions corresponding to the horizontal and vertical sequences of medical images, that is, scanning from left to right and from right to left, coupled with scanning from top to bottom and from bottom to top. Different sequences are further processed by equations we mentioned before. The existing 2D-Selective-Scan Mamba models simply add the scanning results from all directions with equal weights to obtain output features, overlooking the importance of different sequential directions for medical image super-resolution. Therefore, we propose the ISS2D module to blend the four sequences with the weighting coefficients which are learned automatically, making the fusion of the four directions more relevant. Each directional sequence has a coefficient corresponding to its contribution. Summarizing features with the weight coefficients from four different directions can better enhance the global spatial feature connections of medical images across different directions.

3.5 Loss Function

We use the $\mathcal{L}_1$ loss to compute the absolute value of pixel differences between ground truth HR and super-resolved image SR, expressed as follows:

$$\mathcal{L}_1(SR, HR) = \frac{1}{n} \sum_{i=1}^{n} |SR_i - HR_i| , \tag{4}$$

where n is the number of images. $\mathcal{L}_1$ loss as pixel-level loss, leading to overly smooth generated images that miss out on semantic details. Therefore, we also use perceptual loss [15] to measure the discrepancy between images at feature representation levels, thereby improving the visual appeal of super-resolved medical images. Its calculation is as follows:

$$\mathcal{L}^{\phi,j}_{\text{perceptual}}(SR, HR) = \frac{1}{n} \sqrt{\sum_{i=1}^{n} (\phi_j(SR_i) - \phi_j(HR_i))^2}, \tag{5}$$

where ϕ represents the VGG19 network pre-trained on ImageNet. Through training on millions of labeled images, the VGG19 network has learned the ability to recognize a wide range of visual patterns. Therefore, we use it to help our feature extraction. The $\phi_j(\cdot)$ corresponds to the feature map produced by the j^{th} layer of the network ϕ. j is 36.

The final loss function is depicted as follows:

$$Loss = \mathcal{L}_1 + \beta \mathcal{L}^{\phi,j}_{\text{perceptual}}, \tag{6}$$

where β is the weighting factor.

4 Experiments

4.1 Datasets and Metrics

We assessed our method on T2-weighted MRI brain and knee images from the IXI[1] and fastMRI[2] dataset, respectively. The slices in the IXI dataset have a fixed size of 256 × 256 pixels with a resolution of 1mm, while the fastMRI dataset consists of slices having size of 320×320 pixels with a resolution 0.5mm. In our experiment, we used 368 subjects from the IXI dataset for the training and reserved 92 subjects for testing. Regarding the fastMRI dataset, we utilized the data for training on 227 subjects and for conducting tests on 45 subjects. We utilized the peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) [29] commonly used in super-resolution to quantitatively evaluate the medical image quality.

4.2 Implementation Details

To obtain low-resolution medical images for 2× and 4× super-resolution, we initially apply the degradation model in the frequency domain [20] to generate low-resolution images that better match the distribution of real scenes. Our proposed approach was implemented using the PyTorch framework on a NVIDIA RTX A6000 GPU. We used Adam optimizer to train the network with the initial learning rate of 1×10^{-4}. We incorporated 4 vision Mamba blocks in each level, with the channel count [96,128,384,768] for each level, respectively. The dropout rate used within the vision Mamba is 0.3. The dimension of state vectors is 16. The weighting factor β is 0.01.

4.3 Comparison with the State-of-the-arts

Quantitative Comparison. We compared our approach with the SRCNN [2], VDSR [17], FMISR [31], T^{2}Net [4], and DiVANet [1] on the IXI and fastMRI

[1] http://brain-development.org/ixi-dataset/.
[2] https://fastmri.med.nyu.edu/.

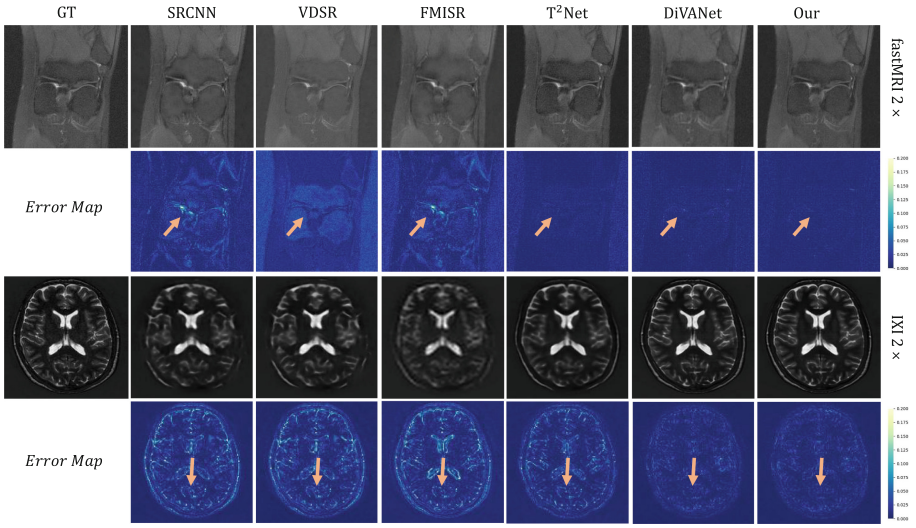


Fig. 3. Qualitative results on fastMRI and IXI dataset under 2× upsampling factor. The significant differences between different methods are shown by the yellow arrow.

dataset under 2× and 4× upsampling factors. Table 1 and Table 2 show the quantitative comparison results. It can be seen from the table that our approach achieved the highest PSNR and SSIM scores compared with other methods for all scaling factors. The primary reasons are: i) the proposed self-prior learning can let the model mine self-exemplar in the original image, fully investigating its own texture and brightness information and enhancing feature representation capability; and ii) the designed Mamba-based Unet module not only can exhaustively exploit local features from different levels but also can fully explore the long-range dependencies of features at different scales.

Table 1. Quantitative results with different methods on fastMRI and IXI dataset under 2× upsampling factor.

Method	fastMRI 2×		IXI 2×	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑
SRCNN	25.82	0.5602	29.23	0.8649
VDSR	27.42	0.6263	29.79	0.8772
FMISR	26.19	0.5583	29.50	0.8685
T^2Net	32.00	0.7158	31.31	0.9035
DiVANet	<u>31.98</u>	<u>0.7169</u>	<u>33.15</u>	<u>0.9320</u>
SMamba-Unet(ours)	**32.06**	**0.7180**	**33.36**	**0.9355**

Table 2. Quantitative results with different methods on fastMRI and IXI dataset under 4× upsampling factor.

Method	fastMRI 4×		IXI 4×	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑
SRCNN	19.74	0.3653	28.12	0.8357
VDSR	20.31	0.3839	28.34	0.8392
FMISR	24.35	0.5207	28.27	0.8349
T^2Net	30.56	0.6244	29.73	0.8773
DiVANet	<u>30.62</u>	<u>0.6352</u>	<u>30.46</u>	<u>0.8946</u>
SMamba-Unet(ours)	**30.70**	**0.6361**	**31.13**	**0.9081**

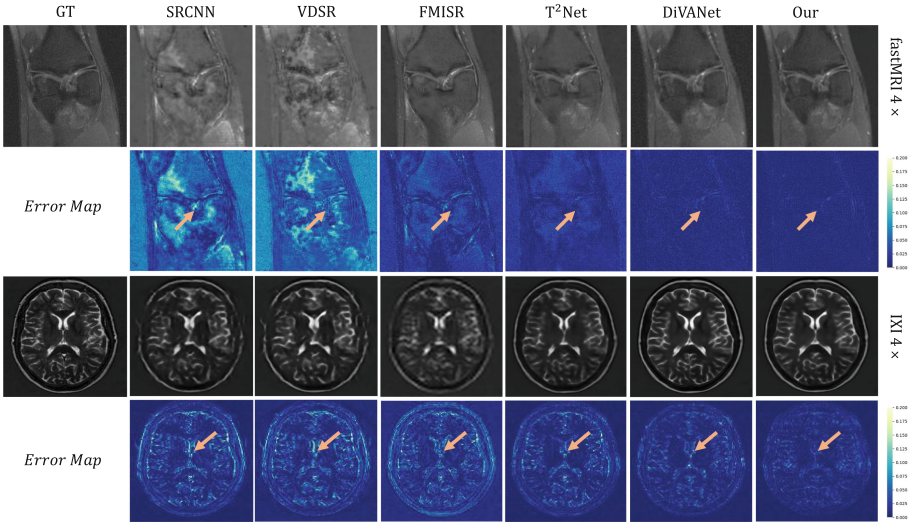


Fig. 4. Qualitative results on fastMRI and IXI dataset under 4× upsampling factor. The significant differences between different methods are shown by the yellow arrow.

Qualitative Comparison. The qualitative experimental results are shown in Figure 3 and Figure 4. In addition to displaying super-resolved images, we also show the corresponding error map to better illustrate the differences with ground truth in detail. The darker the error map, the better the super-resolved image. The specific difference between medical images generated by different methods is indicated by yellow arrows. It is evident that the image generated by SRCNN is very blurred. T^2Net and DiVANet can generate relatively clear soft tissue structures, although there are still some unpleasant blurry edges. From the overall error map, it can be seen that errors mainly exist in areas with significant changes in brightness. The designed self-prior learning method randomly uses brightness inpainting to complete damaged images in the foreground area. It can enhance the ability of the network to explore lost features and changes in

Table 3. Ablation study with different components under 2× upsampling factor.

Method	Mamba	ISS2D	SPL	PSNR↑	SSIM↑
Base	✓	✗	✗	33.18	0.9245
Base+ISS2D	✓	✓	✗	33.26	0.9284
Base+ISS2D+SPL (SMamba-UNet)	✓	✓	✓	**33.36**	**0.9355**

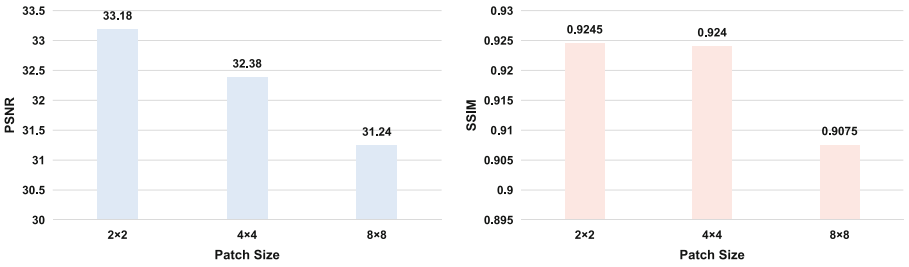


Fig. 5. Ablation study with different patch sizes in vision Mamba module.

brightness through its own information, thereby improving the learning capability of super-resolved features. Therefore, our approach has better overall clarity than other methods and can generate more delicate image details. This aligns with the findings from the quantitative analysis.

4.4 Ablation Analysis

Effectiveness analysis of the key component. To explore the effectiveness of the key component in our SMamba-Unet, we conducted the ablation experiment using the IXI dataset under 2× upsampling factor. We use the pure Mamba-based Unet as baseline, then gradually add improved 2D-Selective-Scan (ISS2D) and self-prior learning (SPL) to it. From Table 3, it can be seen that the performance of the super-resolution network gradually improves as different components participate. The PSNR value increased from 33.18 to 33.36. The value of SSIM increased from 0.9245 to 0.9355. This fully demonstrates that focusing on the importance of different sequence features of images can make the features more effective and accurate. In addition, self-prior learning makes the network pay more attention to its own texture and brightness feature information, which helps in the generation of super-resolved images.

Effectiveness analysis of patch size. In medical image super-resolution tasks using Mamba, selecting the appropriate patch size is an important design decision. To demonstrate the impact of different patch sizes on the performance of Mamba model, we set it to 2, 4, and 8, respectively, to observe the differences in performance. Figure 5 shows that smaller patch sizes achieve better performance. A smaller patch size means that the Mamba will extract finer-grained features from the medical image, which can capture more detailed information.

Table 4. Ablation study with MambaUnet and TransUnet on model complexity.

Method	Params[M]	FLOPS[G]	GPU Memory[Mib]	PSNR↑	SSIM↑
MambaUnet	27.57	18.47	4483	33.18	0.9245
TransUnet	41.05	32.36	30579	32.14	0.9203

Table 5. Ablation study on the weight β in the loss function.

β	PSNR↑	SSIM↑
0.1	33.28	0.9347
0.01	**33.36**	**0.9355**
0.001	33.34	0.9350

To balance performance and computational cost, we choose a patch size of 2 in our approach.

Effectiveness analysis of the baseline selection. We conducted baseline selection experiments to validate the effectiveness of methods based on Vision Mamba and Vision Transformer. We combine the Mamba and Unet as MambaUnet and combine the Transformer and Unet as TransUnet. Vision Transformers have high computational complexity when processing images because the computational complexity of the self-attention mechanism increases quadratically with the number of image patches. While, the hardware-aware algorithm of Mamba processes data with a linear relationship to the sequence length Table 4 shows the Params, FLOPs, and GPU Memory of TransUnet and MambaUnet for 128 $\times$ 128 inputs on the IXI dataset. It can be seen from the table that MambaUnet can achieve better performance with less computational complexity. This indicates that exploring Vision Mamba is an excellent approach to achieving efficient super-resolution.

Effectiveness analysis of weight β in the loss function: We evaluated the impact of weight β in the loss function. Specifically, we conducted ablation studies by setting β to $\{0.1, 0.01, 0.001\}$. Table 5 presents the quantitative evaluation results. The results indicate that $\beta = 0.01$ achieves the best performance than other settings.

5 Conclusion and Future Work

In this paper, we have developed a self-prior guided Mamba-UNet network (SMamba-UNet) for medical image super-resolution. Specifically, our method learns to exploit self-prior multi-scale contextual features within Mamba-UNet networks, potentially facilitating efficient medical image super-resolution. The quantitative and qualitative performance of our SMamba-UNet on the IXI and fastMRI datasets confirms the effectiveness of our approach. How to apply vision Mamba for the multi-model medical image super-resolution, which is an intriguing future path of our approach.

Acknowledgements. The work was supported by the National Key R&D Program of China (No.2018AAA0102100); the National Natural Science Foundation of China (Nos.U22A2034, 62177047); the Key Research and Development Program of Hunan Province (No.2022SK2054); Major Program from Xiangjiang Laboratory under Grant 23XJ02005; Central South University Research Programme of Advanced Interdisciplinary Studies (No.2023QYJC020); the Natural Science Foundation of Hunan Province (No.2024JJ6338); the Natural Science Foundation of Changsha (No.kq2402173); the Fundamental Research Funds for the Central Universities of Central South University (No.2024ZZTS0486); the China Scholarship Council (No.202306370195).

References

1. Behjati, P., Rodríguez, P., Fernández, C., Hupont, I., Mehri, A., Gonzàlez, J.: Single image super-resolution based on directional variance attention network. Pattern Recognit. **133**, 108997 (2023)
2. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell. **38**(2), 295–307 (2016)
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: 9th International Conference on Learning Representations, ICLR (2021)
4. Feng, C.-M., Yan, Y., Fu, H., Chen, L., Xu, Y.: Task Transformer Network for Joint MRI Reconstruction and Super-Resolution. In: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (eds.) MICCAI 2021. LNCS, vol. 12906, pp. 307–317. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87231-1_30
5. Forigua, C., Escobar, M., Arbelaez, P.: Superformer: Volumetric transformer architectures for MRI super-resolution. In: Simulation and Synthesis in Medical Imaging - 7th International Workshop, SASHIMI 2022. vol. 13570, pp. 132–141 (2022)
6. Galias, Z., Yu, X.: Analysis of zero-order holder discretization of two-dimensional sliding-mode control systems. IEEE Trans. Circuits Syst. II Express Briefs **55-II**(12), 1269–1273 (2008)
7. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
8. Gu, A., Goel, K., Ré, C.: Efficiently modeling long sequences with structured state spaces. In: The Tenth International Conference on Learning Representations, ICLR (2022)
9. Gu, A., Johnson, I., Goel, K., Saab, K., Dao, T., Rudra, A., Ré, C.: Combining recurrent, convolutional, and continuous-time models with linear state space layers. In: Advances in Neural Information Processing Systems, NeurIPS. pp. 572–585 (2021)
10. Huang, S., Liu, X., Tan, T., Hu, M., Wei, X., Chen, T., Sheng, B.: Transmrsr: transformer-based self-distilled generative prior for brain MRI super-resolution. Vis. Comput. **39**(8), 3647–3659 (2023). https://doi.org/10.1007/s00371-023-02938-3
11. Ji, W., Chung, A.C.S.: Unsupervised domain adaptation for medical image segmentation using transformer with meta attention. IEEE Trans. Medical Imaging **43**(2), 820–831 (2024)

12. Ji, Z., Kui, X., Liao, S., Zhu, C., Li, Y., Dai, Y., Zou, B.: Wavelet-aware transformer network for multi-contrast knee MRI super-resolution. In: IEEE International Conference on Bioinformatics and Biomedicine, BIBM. pp. 3037–3044 (2023)
13. Ji, Z., Zou, B., Kui, X., Li, Y., Liu, J., Zhao, W., Zhu, C., Dai, Y.: Reconstruction-guided multi-stage network for MRI super-resolution. In: Proceedings of the International Conference on Computer Vision and Deep Learning, CVDL. pp. 6:1–6:5 (2024)
14. Ji, Z., Zou, B., Kui, X., Liu, J., Zhao, W., Zhu, C., Dai, P., Dai, Y.: Deep learning-based magnetic resonance image super-resolution: a survey. Neural Comput. Appl. **36**(21), 12725–12752 (2024)
15. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 694–711. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_43
16. Keys, R.: Cubic convolution interpolation for digital image processing. IEEE Trans. Acoust. Speech Signal Process. **29**(6), 1153–1160 (1981)
17. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR. pp. 1646–1654 (2016)
18. Liu, J., Chen, F., Wang, X., Liao, H.: An edge enhanced SRGAN for MRI super resolution in slice-selection direction. In: Multimodal Brain Image Analysis and Mathematical Foundations of Computational Anatomy - 4th International Workshop, MBIA. vol. 11846, pp. 12–20 (2019). https://doi.org/10.1007/978-3-030-33226-6_2
19. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Liu, Y.: Vmamba: Visual state space model. CoRR **abs/2401.10166** (2024)
20. Lyu, Q., Shan, H., Steber, C., Helis, C., Whitlow, C., Chan, M., Wang, G.: Multi-contrast super-resolution MRI through a progressive network. IEEE Trans. Medical Imaging **39**(9), 2738–2749 (2020)
21. Qiu, D., Cheng, Y., Wang, X.: Dual u-net residual networks for cardiac magnetic resonance images super-resolution. Comput. Methods Programs Biomed. **218**, 106707 (2022). https://doi.org/10.1016/j.cmpb.2022.106707
22. Qiu, D., Zhang, S., Liu, Y., Zhu, J., Zheng, L.: Super-resolution reconstruction of knee magnetic resonance imaging based on deep learning. Comput. Methods Programs Biomed. **187**, 105059 (2020). https://doi.org/10.1016/j.cmpb.2019.105059
23. Ruan, J., Xiang, S.: Vm-unet: Vision mamba unet for medical image segmentation. CoRR **abs/2402.02491** (2024)
24. Shi, J., Tang, L., Gao, Z., Li, Y., Wang, C., Gong, T., Li, C., Fu, H.: Mg-trans: Multi-scale graph transformer with information bottleneck for whole slide image classification. IEEE Trans. Medical Imaging **42**(12), 3871–3883 (2023)
25. Stark, H., Oskoui, P.: High-resolution image recovery from image-plane arrays, using convex projections. J. Opt. Soc. Am. A **6**(11), 1715–1726 (1989)
26. Su, J., Gan, M., Chen, G., Yin, J., Chen, C.L.P.: Global learnable attention for single image super-resolution. IEEE Trans. Pattern Anal. Mach. Intell. **45**(7), 8453–8465 (2023)
27. Umirzakova, S., Ahmad, S., Khan, L.U., Whangbo, T.K.: Medical image super-resolution for smart healthcare applications: A comprehensive survey. Inf. Fusion **103**, 102075 (2024)
28. Wang, J., Zhu, W., Wang, P., Yu, X., Liu, L., Omar, M., Hamid, R.: Selective structured state-spaces for long-form video understanding. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR. pp. 6387–6397 (2023)

29. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
30. Xing, Z., Ye, T., Yang, Y., Liu, G., Zhu, L.: Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. arXiv preprint arXiv:2401.13560 (2024)
31. Zhang, S., Liang, G., Pan, S., Zheng, L.: A fast medical image super resolution method based on deep learning network. IEEE Access **7**, 12319–12327 (2019)
32. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: Efficient visual representation learning with bidirectional state space model. arXiv preprint arXiv:2401.09417 (2024)
33. Zou, B., Ji, Z., Zhu, C., Dai, Y., Zhang, W., Kui, X.: Multi-scale deformable transformer for multi-contrast knee MRI super-resolution. Biomed. Signal Process. Control. **79**(Part), 104154 (2023)

MU-Net: Modified U-Net for Precise Localization and Segmentation of Lumber-Spine Regions from Sagittal Views

Md. Kaisar Ahmed(✉), Felix Havugimana, Kazi Ashraf Moinudin, and Mohammed Yeasin

Department of Electrical and Computer Engineering, University of Memphis, Memphis, TN, USA
mahmed3@memphis.edu

Abstract. Sagittal views provide detailed and critical information about the anatomy and pathology in diagnosing and managing lumbar spine diseases. Radiologists comprehensively evaluate the spinal alignment, structural integrity, and health of bone and soft tissue elements in the sagittal view in diagnosing various spinal conditions and treatment planning. The first step in the diagnosis is localizing the region of interest, typically the lumbar spine segments. To accomplish this, we present a modified U-Net model (MU-Net) for the segmentation and localization of the lumbar spine from sagittal views of Magnetic Resonance Imaging (MRI) images. We employ different techniques to augment data, address the issue of limited training samples, and improve the generalization of deep models. We use the two YOLOv series for localization and MU-Net for segmentation. The MU-Net model achieves an accuracy of 98.93%, a Mean Intersection Over Union (IoU) of 84.29%, and a Dice Coefficient of 98.43%. For the localization, YOLOv8 yields a Precision, Recall, and Mean Average Precision (mAP) of 99.6%, 99.2%, and 99.4%, respectively.

Keywords: Lumbar spine · Localization · Segmentation. Deep Learning · Data Augmentation · MU-Net · YOLOv (series) · MRI · mAP

1 Introduction

Segmentation and localization of medical images, particularly in the mid-sagittal view of the lumbar vertebrae, play a vital role in the effective analysis of spinal deformities, significantly impacting diagnostic applications and treatment planning [1]. Furthermore, these processes are crucial for monitoring the progression of diseases related to the lumbar spine, aiding in the adjustment of therapeutic strategies [2]. Additionally, precise segmentation and localization are fundamental for accurately diagnosing various diseases associated with the lumbar vertebrae [3]. The most common lumbar vertebrae-related diseases include lumbar disc herniation (central, right, left), stenosis, compression of the dural sac and

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15311, pp. 175–190, 2025.
https://doi.org/10.1007/978-3-031-78195-7_12

thecal sac, foraminal stenosis, annular tear, facet hypertrophy, facet arthropathy, ligamentum flavum hypertrophy [4]. Automating the segmentation and localization of medical images can help medical professionals (e.g., radiologists) make more accurate diagnoses and treatment plans for these diseases.

Manual segmentation is widely used to delineate anatomical structures and identify disease-affected areas in lumbar vertebrae. However, this method is time-consuming, requires significant expertise, and is labor-intensive. Automatic segmentation and localization techniques have been developed to address these challenges, allowing efficient and consistent analysis of proper annotated and augmented datasets. Recently, deep learning-based approaches have shown significant promise in medical image segmentation and localization. These models can extract complex information from images, producing accurate and precise segmentation and localization outcomes.

Most reported literature uses a limited amount of training data to train deep-learning models for anatomical structure segmentation and lumbar spine vertebrae localization. However, it is essential to note that deep learning models trained on small datasets are susceptible to overfitting. Mabarki et al. [5] utilized convolutional neural networks based on the Visual Geometry Group 19 (VGG19) architecture to detect herniations in the lumbar disc. They tested the system with only 200 patients. Ala et al. [6] used the centroid distance function to detect disc herniation. In another study [7], authors used two segmentation techniques on mid-sagittal view MRI images to classify the spondylolisthesis and lumbar lordosis. Ghosh et al. [8] introduced a deep learning method to detect and segment tissues in lumbar sagittal MRI. Gang et al. [9] proposed a method to enhance the 'You Only Look Once' (YOLO)-tiny model by integrating three additional Convolutional Neural Network (CNN) layers. While their system can detect spinal fractures with an accuracy of 85.63%, these results suggest that further refinements are necessary to achieve more precise localization to lumbar vertebrae. In [10] used YOLOv5 to accurately locate the lumbar spine with a mAP of 97.5% and diagnose lumbar lordosis with 74.5% accuracy. They further cropped images from YOLOv5 bounding boxes passed through HED U-Net to obtain detailed information about vertebrae and edges. However, edge-based segmentation raises issues in precisely segmenting and localizing the lumbar vertebrae. Our understanding of edge-based segmentation is that it heavily relies on detecting changes in intensity or color, making it highly dependent on image contrast. Consequently, low-contrast images may not yield effective results.

The literature highlights various methods that leverage limited and diversely proportioned training data; however, these approaches often struggle with robustness. Data augmentation can improve adaptability and enhance performance. Rigorous model evaluation is essential. Our methodology, depicted in Figure 1, refines the training process and employs MU-Net for multi-class segmentation across broader and more diverse datasets. We mitigate issues such as high computational costs and overfitting through strategic data augmentation. Additionally, we incorporate the latest YOLOv8 for rapid and accurate local-

ization of lumbar spines. In the context of automatically sagittal view lumbar vertebrae segmentation and localization, our efforts and contributions aim to:

1. Enhance the sagittal view lumbar spine dataset through various augmentation techniques.
2. Utilize modified U-Net; the study categorized lumbar spine regions in sagittal view images, facilitating deformity segmentation.
3. Employ YOLOv8, which played a pivotal role in the precise location of vertebrae and significantly contributed to the advancement of automated spinal issue analysis.
4. Improve the localization process, for which YOLOv5 was employed to compare results, ensuring the method's effectiveness.
5. Utilize a collaborative approach where MU-Net, YOLOv8, and YOLOv5 work together to enhance both segmentation and localization, thereby improving the diagnosis of lumbar deformities.

2 Materials and Methods

This section discusses the materials, including datasets and deep-learning methodologies for segmenting and locating lumbar spine vertebrae. We will also cover the training setup, augmentation techniques, experimental frameworks, and validation pipelines.

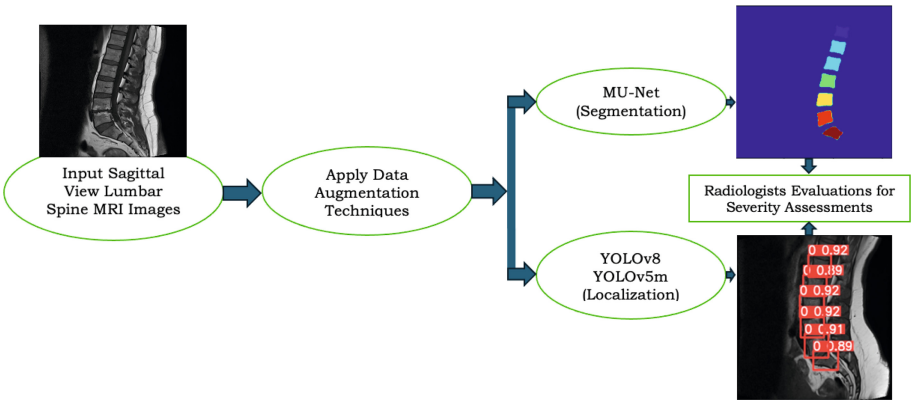
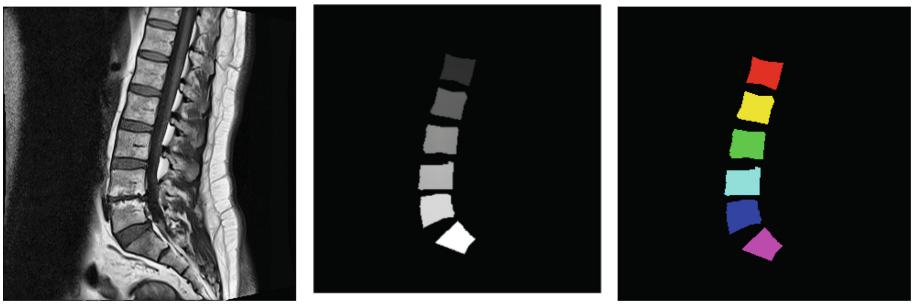


Fig. 1. Proposed Methodology: The input sagittal view lumbar spine MRI images undergo various augmentation techniques applied to both images and labels, such as salt-and-pepper noise, horizontal flipping, 90-degree and 180-degree rotations, random cropping, random rotation, and shearing. These augmented images and labels are then processed by the modified U-Net (MU-Net) for segmentation and YOLOv5m and YOLOv8 for localization. Radiologists will use segmented and localized images for severity assessments.

2.1 Dataset Description and Augmentation

A handful of publicly available datasets have been developed with the collaboration of hospitals and many challenges, aiming to improve the diagnosis of spinal deformities. These datasets are useful resources for understanding and addressing issues related to the lumbar spine. This work explores a dataset focusing on mid-sagittal views of the lumbar spine and its augmentation techniques to automatically segment and localize the lumbar vertebrae precisely.



(a) (b) (c)

Fig. 2. Represents the mid-sagittal view of the human lumber spine dataset. (a) the mid-sagittal view of the lumbar spine, (b) the marked pixel-wise label of the human lumber spine, and (c) the pseudo-colored label of the human lumber spine.

Mid-Sagittal View Dataset for Segmentation and Localization: Our study shown in Fig. 1 used the Lumbar Spine Composite Dataset [11], available on Mendeley Data, which includes sagittal MRI images of 514 subjects. This dataset was collected from people experiencing back pain at the Irbid Speciality Hospital in Jordan from September 2015 to July 2016. We used a labeled dataset [7] for mid-sagittal views to segment the lumber vertebrae. Initially, the dataset contained data for 515 subjects. However, one subject had to be removed due to a noisy image [11], resulting in data for 514 subjects. This dataset has ground truth labels, including marked pixel-wise and pseudo-colored labels for segmentation shown in Fig. 2 [3]. However, training the deep learning models with only these small samples and their corresponding labels may not yield effective results. This is where data augmentation becomes crucial in enhancing the training process. We applied augmentation techniques randomly to both images and their corresponding labels to the initial dataset consisting of 514 subjects; the dataset size significantly expanded to 2,513 samples, as visually depicted in Fig. 3 and Fig. 4.

We used the same dataset for localization and segmentation, sourced from [3]. The data was annotated in YOLOv format with LabelMe and Roboflow Annotate [12] for lumbar vertebrae localization. YOLO format was used to label

the images with six instances of a singular class, "V" for vertebrae, and six labels encompassing five lumbar vertebrae (L1 to L5) and the sacrum (S1) in the sagittal view, as shown in Fig. 5 [10].

Data Augmentation Approaches: Our work explores the efficacy of data augmentation [13] as an important step to enhance our model's performance and generalization capabilities in lumbar spine segmentation and localization. We enhanced the sagittal view lumbar spine dataset through various augmentation techniques. These techniques include adding salt-and-pepper noise (`5%` affected pixels), horizontal flipping, 90-degree rotations, upside-down rotation, random zooming with adjustable zoom levels (`min_zoom=0`, `max_zoom=20`), random rotation angles (`min_angle=-50`, `max_angle=50`), and shearing (`horizontal shear=0.12`,`vertical shear=0.12`) that discussed as in Fig. 3. These thoughtfully adjusted augmentation methods introduce diverse variations, enriching the dataset to enhance model robustness and generalization for precise segmentation and localization of lumbar vertebrae.

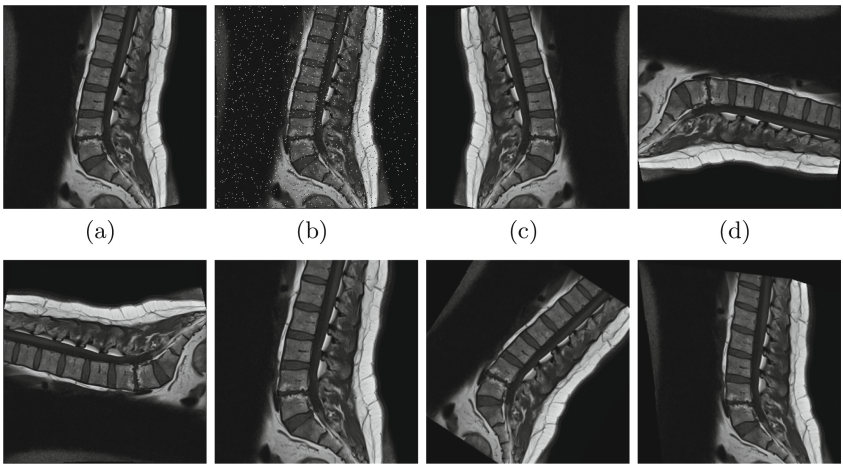


Fig. 3. Images illustrating various augmentation methods. (a) Original sagittal view lumbar spine image. (b) Salt-and-pepper noise (5% affected pixels) for increased variability. (c) Horizontal flipping for an alternative perspective. (d) 90-degree clockwise rotation for diverse orientations. (e) 180-degree rotation (upside-down) for varied viewpoints. (f) Random cropping with adjustable zoom levels (`min_zoom=0`, `max_zoom=20`) for focused regions. (g) Random rotation angle (`min_angle=-50, max_angle=50`) for varied orientation. (h) Shearing (horizontal shear=0.12, vertical shear=0.12) for simulating distortions.

The same careful augmentation techniques were applied to the labels to ensure alignment between the augmented images and their corresponding labels.

These transformations, including horizontal flipping, 90-degree rotations, upside-down rotation, random cropping with adjustable zoom levels, and random rotation angles, were replicated from the images to the labels shown in Fig. 4.

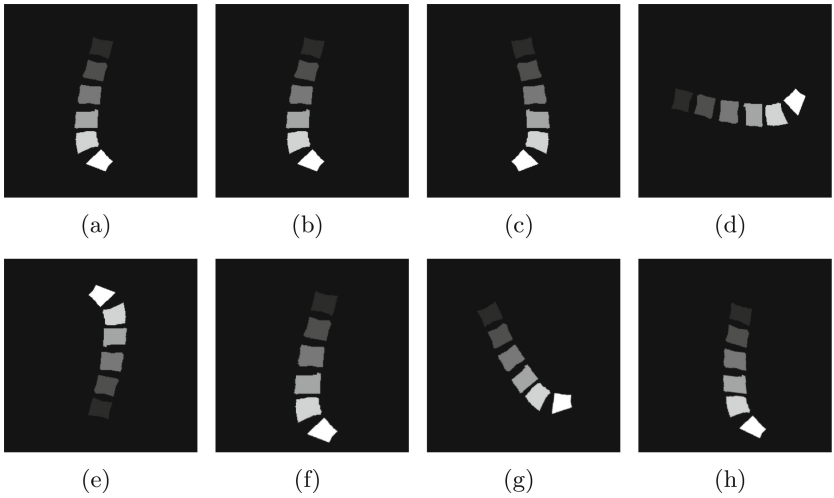


Fig. 4. Corresponding labels illustrating various augmentation methods aligned with augmented images. (a) Original Image. (b) Instead of Salt-and-pepper noise, replace it with the original image. (c) Horizontal Flip. (d) 90-Degree Rotation. (e) Upside-Down Rotation. (f) Random Cropping (Zoom: 0-20%). (g) Random Rotation Angle (-50 to 50 degrees). (h) Shearing (Horizontal: 0.12, Vertical: 0.12).

However, it's crucial to note that adding any type of noise to labels is unsuitable. That's why we replaced the salt and pepper noise with the original images for the labels to maintain consistency, as shown in Fig. 4 (b). Labels serve as ground truth annotations, and introducing noise can compromise the model's accuracy, hindering the model's ability to learn correct associations. That's why we did not add the noise with the labels.

2.2 Segmentation and Localization Methods

For our work, we used a Multi-Class U-Net to segment the sagittal view of the lumbar spine. This approach differs from Binary U-Net Segmentation. However, performing Multi-Class segmentation is quite straightforward if we understand the original concept of U-Net [14]; we need to make a few adjustments to the model. In our study, we modify the U-Net [15] that was initially designed for Sandstone Semantic Segmentation only for four classes. Our study utilized this model to segment the mid-sagittal view of the lumbar spine vertebrae and corresponding labels for seven classes: background, lumbar (L1 to L5), and sacrum (S1). We adjusted several filters within the layers and the final layer of the U-Net to perform lumbar spine segmentation.

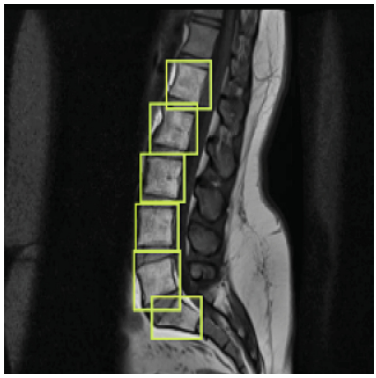


Fig. 5. The figure depicts annotations in a mid-sagittal view. The image has a corresponding label file with the same name, created after annotation, providing details about the class, coordinates, and dimensions of the lumbar vertebrae and sacrum bone.

To address the complexity of the task and ensure the robustness of our approach, we made several critical modifications to the original U-Net architecture. We utilized 3x3 filters consistently across all convolutional layers to maintain spatial resolution, which is essential for accurately capturing the fine details in lumbar spine vertebrae images. The number of kernels was incrementally increased from 16 to 256 in the contracting path, allowing the model to learn progressively more abstract features at deeper layers and correspondingly decreased in the expansive path to reconstruct the image with fine-grained details. Dropout layers with varying dropout rates (0.1 to 0.3) were added to prevent overfitting by randomly deactivating a fraction of neurons during training, thus enhancing the model's generalization capability. Additionally, Conv2DTranspose layers were used for up-sampling in the expansive path, followed by concatenation with corresponding feature maps from the contracting path. This approach preserves high-resolution features from earlier layers and combines them with the up-sampled features, thereby improving the segmentation accuracy. These modifications collectively enhance the model's ability to perform precise and reliable multi-class segmentation of the lumbar spine vertebrae.

We chose the U-Net architecture that we modified, which is useful for biomedical image segmentation, particularly in sagittal view lumber vertebrae segmentation. Its feature integration capability helps accurately identify complex biological structures [16]. The U-Net's simple structure makes it well-suited for processing medical images with well-organized experimental data.

In the context of lumbar spine localization, we utilized YOLOv8 [17] to identify and locate the lumbar vertebrae. We opted for YOLOv8 due to its advanced features within the YOLOv series. The primary advantage of using YOLOv8 is its developer-friendly characteristics, enhanced user experience, faster processing speed, and customization. Unlike previous models in the YOLOv series, which have slow processing and require executing multiple Python files for various

tasks like data preparation, training, or inference, YOLOv8 simplifies the overall workflow with a command-line interface. YOLOv5 [18] offers four variants: YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x, each designed for different computational and accuracy needs. For our study, we employed the YOLOv5m model for comparison, specifically optimized to improve the localization of lumbar vertebrae.

2.3 Training Details for Segmentation and Localization

The applied models: **MU-Net, YOLOv8, and YOLOv5m**, were trained on an NVIDIA GeForce RTX SUPER with 7996MiB of memory, adding the GPU's computational capabilities for efficient model training. We used augmented Mid-Sagittal view MRI images and their corresponding ground truth masks to train the MU-Net model. The employed dataset was divided into three subsets: 70% for training, 15% for validation, and 15% for testing. The test set was not used during training or validation, ensuring an unbiased evaluation of the model's performance.

The validation set was used solely to tune hyperparameters and monitor the training process. Early stopping was implemented to halt training when the validation loss did not improve for 10 consecutive epochs, and learning rate decay was applied using the ReduceLROnPlateau callback, which reduced the learning rate by a factor of 0.2 if the validation loss did not improve for 5 consecutive epochs, with a minimum learning rate of 0.001. The model was trained for a maximum of 90 epochs.

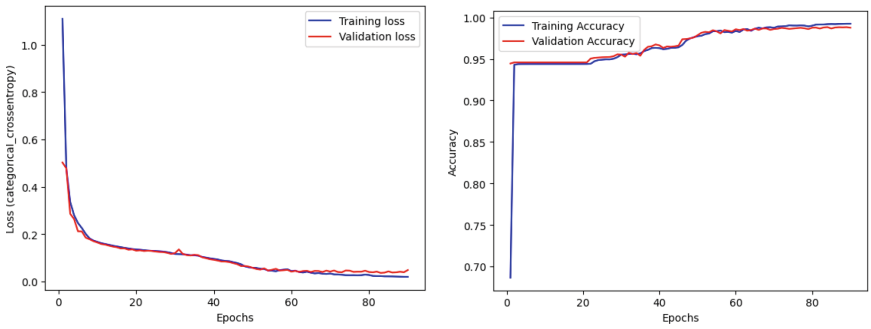
The model was compiled using the Adam optimizer and the categorical cross-entropy loss function, which is suitable for multi-class segmentation tasks. The training was conducted with a batch size of 32 images per iteration.

The evaluation metrics, including accuracy, Mean Intersection over Union (IoU), Dice Coefficient, Mean Average Precision (mAP), and Jaccard Index, were calculated using the test set, which was separate from the training and validation sets. More details will be discussed in Section 3.

Figure 6 (a) and (b) illustrates the training and validation loss and accuracy graphs over 60 epochs of the MU-Net model. The gap between the training and validation lines is minimal, indicating no significant overfitting. This suggests that the model generalizes well to new data. The model performance is evaluated on a separate set of sagittal images not seen during training, ensuring its generalization to unseen data.

The training and validation loss curve in Figure 6 (a) shows a steep decline in loss during the initial epochs, which then stabilizes, indicating that the model has effectively learned the training data and generalizes well to the validation data. Similarly, the accuracy curve in Figure 6 (b) demonstrates rapid improvement in accuracy during the initial training phase, followed by a stable convergence, reflecting the model's robustness and reliability.

For vertebrae localization, we randomly augmented 514 subject images by different augmentation techniques described in Fig. 3, 4 to increase the training



(a) (b)

Fig. 6. The figure represents the training and validation loss, along with accuracy, observed over 60 epochs in the MU-Net Model. (a) Optimal Training Performance: The training loss (yellow) compared to validation loss (red), (b) Accuracy Convergence: Both training (yellow) and validation (red) accuracy steadily approach optimal values. (Color figure online)

samples to 2,162 images. After applying annotation approaches in YOLOv format [10], the images and corresponding YOLOv formatted labels go through the training using YOLOv8. The training employs a batch size of 32 and completed 90 epochs. The dataset is partitioned into 85% for training, 10% for validation, and 5% for testing. This allocation was made to maximize the amount of data available for training and validation, which is crucial for optimizing the performance of YOLO models, which are known for their need for extensive training data. We used the same training setup for YOLOv5m.

Despite the smaller test set, YOLOv5m and YOLOv8 achieved high confidence scores in detecting the lumbar vertebrae and sacrum, indicating robust performance that will be described in Section 3. The primary reason for this split was to ensure that the models received enough varied examples during training to learn effectively, with the validation set used to tune hyperparameters and prevent overfitting.

3 Evaluation, Results and Discussion

In this section, we compare and discuss the results of our evaluation of three models: MU-Net, YOLOv8, and YOLOv5m. As illustrated in Figure 1, these models were employed in our proposed approach, utilizing an augmented dataset alongside state-of-the-art (SOTA) methods. The models were evaluated based on several key metrics: Accuracy, Mean Intersection over Union (IoU), Dice Coefficient, Jaccard Index for precise segmentation, and Mean Average Precision (mAP) for the localization of lumbar vertebrae. Our analysis of these metrics provides valuable insights into the effectiveness of these models in our methods of diagnosing lumbar spine deformities.

We used Intersection Over Union (IoU) to evaluate the accuracy of each class in sagittal view lumbar spine segmentations. IoU measures the overlap between predicted and ground truth regions for seven distinct classes: background, lumbar (L1 to L5), and sacrum (S1). A higher IoU score, nearing 1, indicates greater accuracy.

Accuracy, another key metric, quantifies the overall correctness of the segmentation by comparing the number of correctly predicted pixels to the total number of pixels. A higher accuracy indicates a better overall performance.

The Dice Coefficient measures the overlap between the predicted and ground truth regions but emphasizes the similarity between the sets. It is calculated as twice the area of overlap divided by the total number of pixels in both the predicted and ground truth regions. A Dice Coefficient closer to 1 indicates a more accurate segmentation.

The Jaccard Index assesses the similarity between the predicted and ground truth regions. It is calculated as the intersection divided by the union of the two sets. A higher Jaccard Index indicates better performance in segmentation tasks.

Additionally, we use Mean Average Precision (mAP) to assess the performance of lumbar vertebrae localization. mAP evaluates the precision of the predicted bounding boxes against the ground truth boxes, with a higher mAP indicating more precise localization of vertebrae.

Table 1. Intersection Over Union and Jaccard Index Values for Seven Classes: background, lumbar (L1 to L5), and sacrum (S1).

Class	Intersection Over Union	Jaccard Index
Background	0.99147016	0.99147016
L1	0.806324	0.80516865
L2	0.74521583	0.74507784
L3	0.7884944	0.73730023
L4	0.87640816	0.80159759
L5	0.8690623	0.83781334
S1	0.8235583	0.80476218

The **MU-Net** model in our proposed method in Fig. 1 achieved an accuracy of 98.93% with a Mean Intersection Over Union (IoU) of 84.29%. The IoU values obtained for distinctive classes in Table 1 are notably impressive. This range of IoU for medical images is expected to be obtained due to various artifacts and the inherent noise in the images of anatomical structures. The Jaccard Index for each class, also shown in Table 1, further validates the high accuracy of the model, with values close to the IoU, indicating robust segmentation performance. Additionally, the MU-Net model achieved a Dice Coefficient of 98.27%, which outperforms this work [7]. This high Dice Coefficient reflects the model's excellent overlap between the predicted and ground truth regions.

The MU-Net model performed well on test images, with the IoU value for each class depicted in Table 1. The predicted images were plotted in a 'jet' color map using the MU-Net model, which showed almost identical results to the testing label in Fig. 7. Based on the observed outcomes, the MU-Net model exhibits superior performance and mitigates the issue of overfitting by utilizing augmented training data. The diversity of the training data significantly enhanced the model's generalization capabilities, enabling the MU-Net model to segment lumbar vertebrae with high precision.

However, the MU-Net model had incorrect predictions on some completely unseen images in Fig. 8. The main challenge is to increase the Mean IoU value in medical imaging while reducing these wrong predictions. In semantic segmentation, accuracy doesn't tell us much. So, we must prepare the images and their corresponding masks precisely to get more accurate predictions and improve the Mean IoU. To prevent the model from overfitting, a larger dataset with accurate annotations, specifically for medical image segmentation, must be constructed.

Table 2. Performance Metrics Comparison between YOLOv8 and YOLOv5m Object Detection Models for localizing Sagittal view lumbar vertebrae

Performance Metric	YOLOv8	YOLOv5m
Precision(%)	99.1	99.2
Recall(%)	99.5	99.5
mAP(%)	99.4	99.3

To precisely localize each vertebra (L1 to L5, and S1), Table 2 compares the performance metrics for YOLOv8 and YOLOv5m in localizing lumbar vertebrae from sagittal view images. YOLOv8 achieves a mAP of 99.4%, indicating a slight improvement in precision for detecting and localizing abnormalities in the lumbar spine, compared to YOLOv5m's mAP of 99.3%. Both models exhibit an identical recall rate of 99.5%, highlighting their effectiveness in consistently identifying lumbar vertebrae as a key factor in diagnosing spinal deformities. Interestingly, YOLOv8 has a slight advantage in precision over YOLOv5m's by 0.1%, suggesting a better ability to locate the vertebrae precisely. The table reveals advancements with YOLOv8, suggesting its enhanced capability in accurately detecting spinal conditions. These enhancements are crucial for accurately evaluating and devising treatment plans for lumbar spine issues, potentially improving patient outcomes in orthopedic and neurosurgical fields.

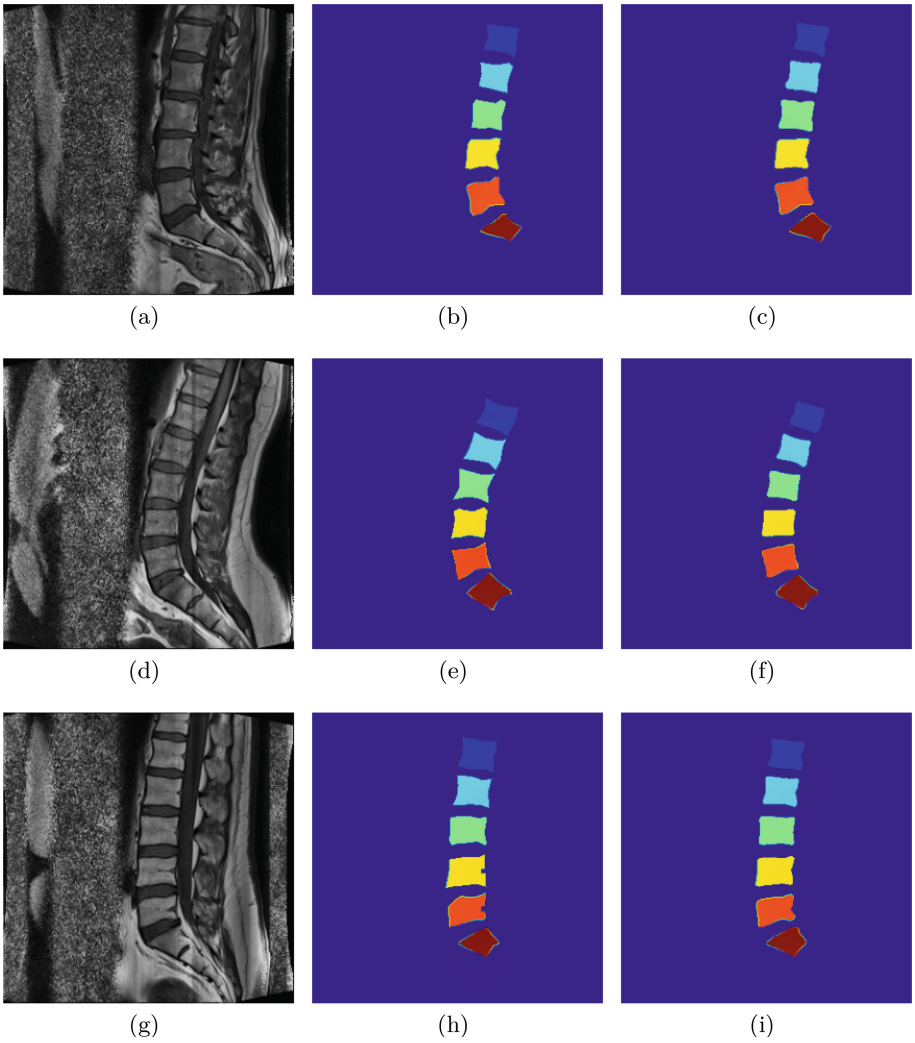


Fig. 7. Evaluation of MU-Net Model Performance on Sagittal View Lumbar Spine Images: Detailed Analysis of Testing Images (a, d, g), Ground Truth Labels (b, e, h), and Model Predictions (c, f, i)

Table 3. Comparision of Speed Metrics for YOLOv8 vs. YOLOv5m in Sagittal View Lumbar Spine Vertebrae Localization

Speed Metric	YOLOv8	YOLOv5m
Preprocess	1.9 ms per image	0.5 ms per image
Inference	5.7 ms per image	6.0 ms per image
Postprocess	6.3 ms per image	4.0 ms per image

Table 4. Comparative Performance of Our Methods and State-of-the-Art Methods for Lumbar Vertebrae Segmentation and Localization

Method	Segmentation IoU	Localization mAP	Sample Size	Reference
YOLOv2	77.3%	75.30%	Only 40 Patients	Gang et al [9]
SegNet	58.0%	N/A	200 Patients	W. Mbarki, et al. [5]
UNet	74.0%	N/A	514	R.F. Masood et al. [7]
YOLOv3	N/A	91.7%	514	Mushtaq et al. [10]
YOLOv5s	N/A	97.5%	514	Mushtaq et al [10]
YOLOv5m	N/A	99.3%	514 (augmented to 2,162)	Ours
YOLOv8	N/A	99.4%	514 (augmented to 2,162)	Ours
MU-Net	84.29%	N/A	514 (augmented to 2,513)	Ours

We also compared the speed metrics of YOLOv models shown in Table 3. Our analysis revealed that YOLOv8 has longer pre-processing and post-processing times than YOLOv5m. However, YOLOv8 boasts a faster inference time, which is crucial for real-time diagnostics. Despite YOLOv8's marginally quicker analysis phase, its overall processing time may affect throughput in large datasets.

Figure 9 demonstrates YOLOv8's consistently high confidence levels across lumbar vertebrae, while YOLOv5m shows slightly higher confidence levels in some instances. Our augmented dataset for lumbar vertebrae localization yielded superior outcomes compared to state-of-the-art (SOTA) methods.

In Table 4, we compare the results achieved by our adopted methods in Fig. 1, including the MU-Net architecture, YOLOv8, and YOLOv5m, with those obtained by various researchers using deep learning models on the same and different datasets. The comparison clearly shows the improved performance of our methods in both segmentation and localization tasks, thereby validating the effectiveness of the modifications made to the U-Net architecture and the use of YOLOv8 for localization. Our MU-Net model achieved the notable segmentation IoU of 84.29%, outperforming other models like UNet (74.0%), and SegNet (58.0%). This demonstrates the effectiveness of our modifications to the U-Net architecture for the task of lumbar vertebrae segmentation. Localization Performance: Our applied YOLOv8 model achieved the highest localization mAP of 99.4%, followed closely by YOLOv5m (99.3%). These results surpass the performance of YOLOv5s (97.5%), YOLOv3 (91.7%), and YOLOv2 (75.30%), highlighting the effectiveness of our approach in accurately localizing lumbar vertebrae.

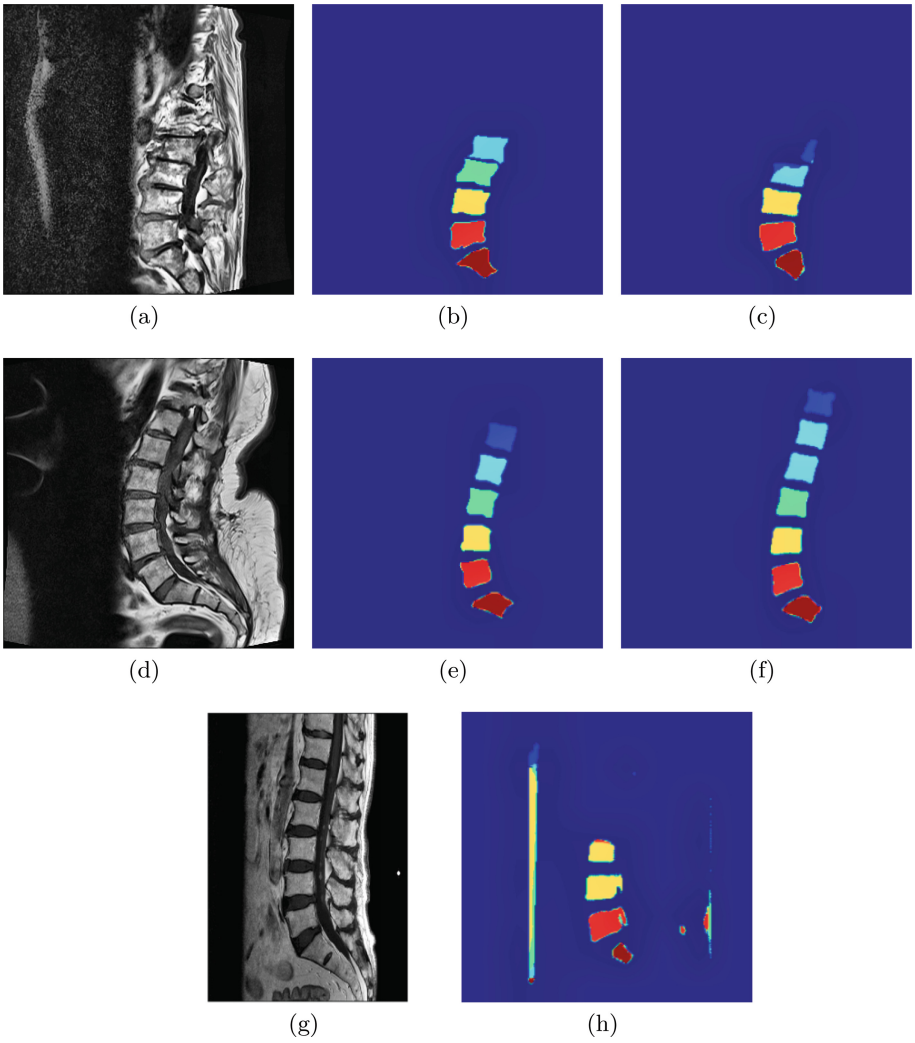
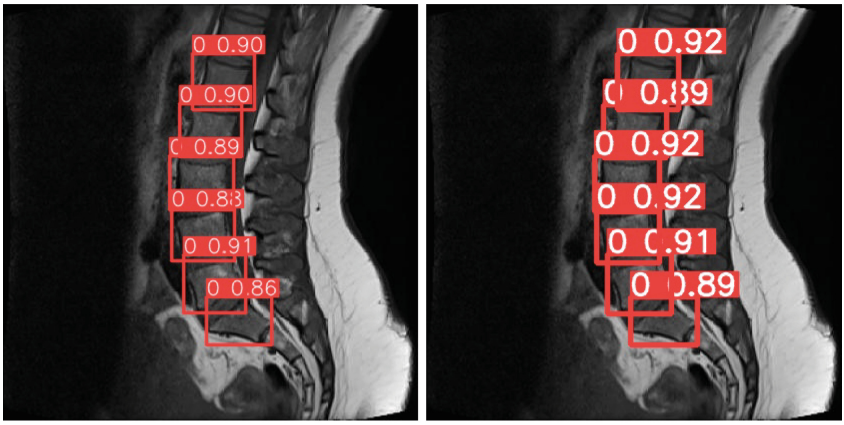


Fig. 8. Instances of Unsuccessful Predictions by the MU-Net Model on External Test Images: Analysis of Unseen Testing Images (a, d, g), Ground Truth Labels (b, e), and Model Predictions (c, f, h)

MU-Net segments the vertebrae, while YOLOv8 and YOLOv5m localize them. This process aids radiologists in determining disease severity and automates the workflow, facilitating the creation of a practical system. YOLOv8 and YOLOv5m perform similarly, with only a marginal difference in inference time, making either suitable for localization. The improved performance of our methods in both segmentation and localization tasks validates the effectiveness.



(a) (b)

Fig. 9. (a) YOLOv8 Vertebrae Detection, (b) YOLOv5 Vertebrae Detection

4 Conclusion

Our study refined the Modified-UNet model for segmentation and utilized YOLOv8 and YOLOv5m for precise lumbar vertebrae localization. The outcomes revealed superior IoU scores for vertebrae L1-L5 and the S1 sacrum with the MU-Net application and YOLOv models, achieving mAP scores of 99.3% and 99.4% for YOLOv5m and YOLOv8 respectively, highlighted their enhanced accuracy in lumbar deformity analysis. These methods on our augmented dataset surpass previous techniques, providing crucial insights for choosing between YOLOv8 and YOLOv5m based on confidence, efficiency, and application needs. Moving forward, we aim to further explore sagittal lumbar spine measurements to advance the diagnosis and understanding of related conditions, contributing to improved public health outcomes.

References

1. Chengwen Chu et al. “Fully automatic localization and segmentation of 3D vertebral bodies from CT/MR images via a learning-based method”. In: PloS one 10.11 (2015), e0143327
2. Jeffrey De Fauw et al. “Clinically applicable deep learning for diagnosis and referral in retinal disease”. In: Nature medicine 24.9 (2018), pp. 1342-1350
3. Rao Farhat Masood et al. Composite Dataset of Lumbar Spine Mid-Sagittal Images with Annotations and Clinically Relevant Spinal Measurements. Mendeley Data. Version V2. 2021. https://doi.org/10.17632/k3b363f3vz.2.
4. Damian Hoy et al. “The global burden of low back pain: estimates from the Global Burden of Disease 2010 study”. In: Annals of the rheumatic diseases 73.6 (2014), pp. 968-974

5. Wafa Mbarki et al. “Lumbar spine discs classification based on deep convolutional neural networks using axial view MRI”. In: Interdisciplinary Neurosurgery 22 (2020), p. 100837
6. Ala S Al Kafri et al. “Detecting the disc herniation in segmented lumbar spine mr image using centroid distance function”. In: 2017 10th International Conference on Developments in eSystems Engineering (DeSE). IEEE. 2017, pp. 9-13
7. Rao Farhat Masood et al. “Deep learning based vertebral body segmentation with extraction of spinal measurements and disorder disease classification”. In: Biomedical Signal Processing and Control 71 (2022), p. 103230
8. Subarna Ghosh and Vipin Chaudhary. “Supervised methods for detection and segmentation of tissues in clinical lumbar MRI”. In: Computerized medical imaging and graphics 38.7 (2014), pp. 639-649
9. Gang Sha, Junsheng Wu, and Bin Yu. “Detection of spinal fracture lesions based on improved Yolov2”. In: 2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA). IEEE. 2020, pp. 235-238
10. Malaika Mushtaq et al. “Localization and edge-based segmentation of lumbar spine vertebrae to identify the deformities using deep learning models”. In: Sensors 22.4 (2022), p. 1547
11. Rao Farhat Masood et al. “A Composite Dataset of Lumbar Spine Images with Mid-Sagittal View Annotations and Clinically Significant Spinal Measurements”. In: 2022 2nd International Conference on Digital Futures and Transformative Technologies (ICoDT2). IEEE. 2022, pp. 1-5
12. Roboflow. Labeling Your Data: The Essential Step for Machine Learning. Accessed: Date of access. 2023. url: https : / / blog . roboflow . com / labeling/
13. Amey Gondhalekar. Data Augmentation: Is It Really Necessary? Accessed: December 17, 2023. 2023. url: https://medium.com/analytics-vidhya/data-augmentation-is-it-really-necessary-b3cb12ab3c3f
14. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-net: Convolutional networks for biomedical image segmentation”. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings,Part III 18. Springer. 2015, pp. 234–241
15. Bnsreenu. Multiclass U-Net for Sandstone Segmentation. https://github.com/bnsreenu/python_for_microscopists/blob/master/208_multiclass_Unet_sandstone.py. Accessed: December 17, 2023. 2023
16. Kaiming He et al. “Deep residual learning for image recognition”. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, pp. 770-778
17. Ultralytics. YOLOv8 Documentation. https://docs.ultralytics.com/. 2023
18. Ultralytics. YOLOv5 on PyTorch Hub. https://pytorch.org/hub/ultralytics_yolov5/. 2024

Harnessing Vision Transformers for Precise and Explainable Breast Cancer Diagnosis

Hossam Magdy Balaha[1(✉)], Khadiga M. Ali[2], Dibson Gondim[3], Mohammed Ghazal[4], and Ayman El-Baz[1]

[1] Bioengineering Department, J.B. Speed School of Engineering, University of Louisville, Louisville, KY, USA
hmbala01@louisville.edu

[2] Pathology Department, Faculty of Medicine, Mansoura University, Mansoura, Egypt

[3] Department of Pathology and Laboratory Medicine, University of Louisville, Louisville, KY, USA

[4] Electrical, Computer, and Biomedical Engineering Department, Abu Dhabi University, Abu Dhabi, UAE

Abstract. Breast cancer (BC) remains a significant global health challenge, impacting millions of lives annually. Traditional histopathological analysis, while essential, can be subjective and time-consuming, potentially leading to diagnostic inaccuracies. This study proposes a novel Computer-Aided Diagnosis (CAD) framework utilizing Vision Transformers (ViTs) for BC diagnosis from histopathology slides. ViTs excel in capturing global dependencies within images, offering enhanced diagnostic accuracy compared to conventional methods. The framework integrates ViTs with advanced decision-making techniques like 2-tier majority fusion and SHapley Additive exPlanations (SHAP) for improved interpretability. Experimental results on a dataset of post-neoadjuvant therapy breast cancer samples demonstrate the efficacy of the proposed approach, achieving high performance metrics and providing insights into model predictions. The proposed approach achieves state-of-the-art performance with an accuracy exceeding 97% surpassing existing methods both on the utilized dataset and an external benchmark, specifically the Breast Cancer Histopathological Database (BreakHis). Time complexity analysis suggests that the proposed framework offers computational efficiency, with the dominant factors influencing overall complexity being the number of patches, sequence length, and number of layers in the ViT model. This study contributes a robust methodology towards enhancing BC diagnostic precision and efficiency through cutting-edge AI technologies.

Keywords: Breast Cancer (BC) · Computer-Aided Design (CAD) · Histopathology Diagnosis · Residual Cancer Burden Index (RCBi) · SHapley Additive exPlanations (SHAP)

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15311, pp. 191–206, 2025.
https://doi.org/10.1007/978-3-031-78195-7_13

1 Introduction

Breast cancer (BC) is a prevalent form of cancer that develops in the cells of the breasts, typically beginning in the milk-producing ducts (ductal carcinoma) or the lobules (lobular carcinoma) [1,14]. It is the second most common type of cancer diagnosed in women worldwide, and although it predominantly affects women, it can also occur in men [12]. Statistics reveal the widespread impact of BC on individuals and societies globally. According to the World Health Organization (WHO), an estimated 2.3 million new cases of BC were diagnosed in 2020 alone, making it a significant public health concern [14,32]. Additionally, BC is a leading cause of cancer-related deaths among women, underscoring the urgency for early detection and effective treatment strategies [21].

Histopathological analysis has an important role in diagnosing BC and guiding the different treatment decisions. This process involves examining tissue samples obtained through biopsy (or surgery) under a microscope to identify abnormalities indicative of cancerous growth [24]. Pathologists assess various histological features, such as cell morphology, nuclear characteristics, and tissue architecture, to classify tumors and determine their aggressiveness and potential for metastasis [2].

Traditional histopathological analysis, although crucial, can be time-consuming and subjective, depending heavily on the pathologist's expertise. Determining the tumor grade from histopathology often requires a council of doctors to reach a final decision, and time is a critical factor in this process. In some cases, diagnoses are made by a single expert, which can result in false positives and false negatives, further complicating patient outcomes.

Consequently, we are motivated to enhance the diagnostic process and support the experts in their decisions. The suggested approach treat images as sequences of patches from different magnification levels and utilize self-attention mechanisms, enabling them to capture global dependencies more effectively than conventional Convolutional Neural Networks (CNNs). This can lead to improved classification of histopathology slides, potentially resulting in earlier and more accurate BC diagnoses.

Additionally, utilizing a dataset based on the Residual Cancer Burden Index (RCBi) is pivotal. The RCBi is a clinically validated tool used to assess the response to neoadjuvant therapy (NAT) and is associated with prognosis. Moreover, integrating advanced techniques such as SHapley Additive exPlanations (SHAP) is expected to boost the interpretability and reliability of the results, ensuring that the system not only performs well but also provides insights into its decision-making process.

Therefore, the current study proposes a novel approach for BC diagnosis from histopathology slides using a CAD framework. This framework utilizes Vision Transformers (ViTs) to perform the diagnosis. The proposed CAD framework aims to improve the accuracy and efficiency of BC diagnosis by integrating ViTs with advanced decision-making techniques. The hypotheses of the current study are: (1) ViTs can effectively classify histopathology slides for BC diagnosis, (2) incorporating 2-tier majority fusion (2-Tier MF) and SHapley Additive exPlana-

tions (SHAP) techniques can enhance the interpretability and reliability of the BC diagnostic process, and (3) the system can decrease the percentage of false positives and negatives by helping experts ensure their decisions.

The contributions of the current study are: (1) it introduces a novel CAD framework for BC diagnosis that integrates state-of-the-art ViTs with 2-Tier MF and SHAP techniques. (2) the study provides empirical evidence of the effectiveness of ViTs in classifying histopathology slides, demonstrating state-of-the-art performance on the utilized dataset. (3) by employing 2-Tier MF, the study enhances the decision-making process by aggregating multiple ViTs models. (4) the incorporation of SHAP allows for a comprehensive interpretation of model predictions showing the features and regions of histopathology slides that contribute most significantly to the diagnosis.

2 Related Studies

Recent advancements in technology, particularly in the field of artificial intelligence (AI), have revolutionized the analysis of histopathological slides in BC diagnosis [7,11,26].

For example, Sharma and Mehra [26] explored automatic multi-classification of BC histopathological images, a vital field in biomedical informatics. They compared two machine learning approaches using the BreakHis dataset. The first involved handcrafted feature extraction, while the second utilized transfer learning with pre-trained networks (VGG16, VGG19, and ResNet50). Results showed that using pre-trained networks as feature extractors surpassed baseline and handcrafted methods across all magnifications, with augmentation notably enhancing classification accuracy. Specifically, the VGG16 network with linear SVM achieved the highest accuracy, reaching 93.97% for patch-based classification at 40× magnification. *However, this study has limitations. It does not utilize state-of-the-art techniques such as ViTs and You Only Look Once (YOLO), which are known to capture global context more effectively than traditional CNNs.*

Furthermore, Elmannai et al. [13] aimed to automate analysis and diagnosis of BC using histopathological images. They employed transfer learning with two deep CNNs, Inception and Xception, to extract features from BC tissue images. These features were merged, reduced with dropout, and fed into fully connected layers for classification. The study achieved remarkable results, with sub-image classification reaching 97.29% accuracy and 99.58% sensitivity for carcinoma cases. Whole image classification attained 100% accuracy with majority vote and 95% with maximum probability fusion decision, outperforming previous methods in accuracy and sensitivity. *This study has several limitations: (1) similar to the previous related study, they did not utilize SOTA approaches, (2) They did not specify whether the reported results were from training, testing, or validation datasets, which can affect the interpretation of the findings and reproducibility, and (3) the study did not incorporate explainable AI techniques to provide insights into the decision-making process.*

Additionally, Wang et al. [31] aimed to predict gBRCA mutation risk in BC patients using whole-slide pathology features. They employed a deep CNN

based on ResNet, trained on whole-slide images (WSIs) divided into smaller tiles for classification. Models were trained on tiles cropped at different magnifications to assess information levels. External validation showed AUCs ranging from 0.551 to 0.774 for magnification tiles and 0.635 to 0.828 for magnification slides, with histological grade impacting prediction accuracy. *However, this study did not incorporate explainable AI techniques to provide insights into the decision-making process.*

Despite the notable advancements discussed in recent studies utilizing AI for BC diagnosis from histopathological images, there remains a significant research gap. Current literature predominantly focuses on traditional CNNs and transfer learning techniques, such as VGG16, ResNet, Inception, and Xception. However, there is a limited exploration of state-of-the-art approaches such as ViTs and YOLO, which have demonstrated superior capability in capturing global context and enhancing classification accuracy in other domains. Moreover, these studies do not incorporate explainable AI techniques to provide insights into the decision-making process.

3 Materials

In this study, we utilized a dataset named Post-NAT-BRCA. It was developed by the Department of Anatomic Pathology at Sunnybrook Health Sciences Centre (SBHSC) in Toronto, Canada. It consists of 96 WSIs stained with Hematoxylin and Eosin (H&E). The slides were taken from 54 patients who underwent neoadjuvant therapy (NAT) [23]. Manual annotations of tumor cellularity and cell labels, provided as Sedeen annotation files. The dataset we used in our study is accessible to the public, ensuring transparency and availability for further research [18]. The dataset classes are low grade tumor cellularity, medium grade tumor cellularity, high grade tumor cellularity, and normal.

We organized annotated patches into four separate folders based on their assigned categories in a systematic manner. These categories were extracted from three magnification levels (x16, x4, and x2). Patches are extracted from the 96 WSIs, with each class having approximately 12,000 patches, resulting in a total of over 48,000 patches. Each patch measures 1,024 by 1,024 pixels. This extraction process is performed at three magnification levels, with a 10% overlap between patches. Samples from the base level (i.e., highest magnification level) are presented in Figure 1.

4 Methods

The current study proposed a CAD framework (See Figure 2) for diagnosing the BC utilizing histopathology slides. It utilizes three major phases: (1) data acquisition, annotations and patches extraction, (2) classification, tuning, and evaluation, and (3) majority fusion and AI explainability using SHAP. The data acquisition, annotations and patches extraction process is illustrated in the Materials (See Section 3).

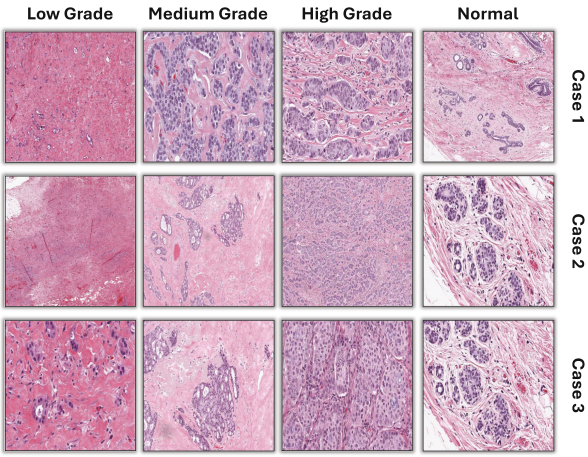


Fig. 1. Samples from the base level (i.e., highest magnification level) for different categories: Low Grade, Medium Grade, High Grade, and Normal.

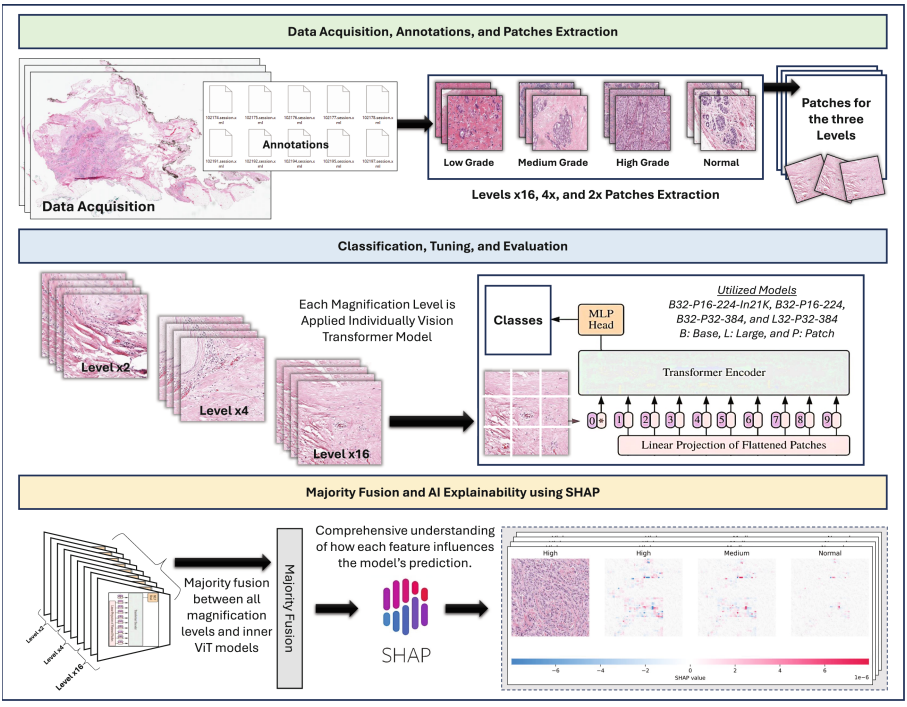


Fig. 2. The suggested Computer-aided design (CAD) framework in the current study for breast cancer (BC) diagnosis from hitsopathology slides.

4.1 Classification and Evaluation via Vision Transformers

Vision Transformers, or ViTs, represent a recent breakthrough in computer vision tasks, including image classification. Unlike traditional CNNs, which rely on hierarchical feature extraction through convolutional layers, ViTs utilize the Transformer architecture, originally proposed for natural language processing tasks [29].

The Transformer architecture, introduced by Vaswani et al. [29] in the context of sequence-to-sequence learning, consists of self-attention mechanisms that allow the model to weigh the importance of different input elements when making predictions [22]. This self-attention mechanism has proven to be highly effective in capturing long-range dependencies in sequences, making it well-suited for various tasks beyond natural language processing, including image classification [28]. ViTs handle the images like rows of patches. Each patch goes through a series of changes. The changes involve understanding the whole image and then using that understanding to sort the patches [25].

Let $X = \{x_1, x_2, \ldots, x_N\}$ represent the sequence of patches extracted from the input image where N is the number of patches and x_i denotes the patch of the i-th patch. The embedding and position encoding process, represented as $z_i = x_i + p_i$, will obtain the embedded vector (z_i) where p_i is the position of the patch.

The self-attention mechanism in ViTs computes attention scores between all pairs of patches and generates weighted representations of each patch based on these scores. The output of the self-attention layer can be expressed as in Equation 1 where W_Q, W_K, and W_V are learnable parameters of the self-attention layer, and d_k represents the dimensionality of the key $(Z \cdot W_K)$ vectors. The SoftMax operation ensures that the attention scores sum up to 1 across all patches. By stacking multiple layers of multi-head attention encoders, Z will contain the captured complex patterns and dependencies in the patches.

$$\text{Attention}(Z) = \text{SoftMax}\left(\frac{(Z \cdot W_Q) \cdot (Z \cdot W_K)^T}{\sqrt{d_k}}\right) \times (Z \cdot W_V) \tag{1}$$

After computing the attended representations, a feedforward neural network (FFN) is applied to each patch independently to capture the non-linear relationships. The output of the FFN is represented in Equation 2 where the weights and biases (W_1, b_1, W_2, and b_2) are learnable matrices.

$$\text{FFN}(Z) = \text{ReLU}(Z \times W_1 + b_1) \times W_2 + b_2 \tag{2}$$

4.2 Majority Fusion and AI Explainability using SHAP

The 2-tier majority fusion process (See Figure 2) involves an approach to decision-making, integrating the outputs from the selected models and the three magnification levels. In the first tier, majority fusion is performed at each magnification level (i.e., 2x, 4x, and 16x) for each patch based on the top-M models.

Then, in the second tier, these results undergo another majority fusion procedure to produce a final decision.

After obtaining the 2-tier majority fused patches, SHAP is utilized to clarify the model's predictions by attributing the contribution of each feature to the overall outcome. This approach provides a clear understanding of the model's predictions. For diagnosing histopathology slides using ViTs, SHAP offers insights into how different parts of a WSI influence the model's final diagnosis [4,20].

The Shapley value is the major component of the SHAP. It originates from cooperative game theory [20]. The Shapley value assigns a worth to each aspect based on its individual impact on the final prediction [16]. Mathematically, the Shapley value for a specific aspect i is defined by Equation 3 where $\phi_i(v)$ represents the Shapley value for aspect i, v denotes the aspect's values, N signifies the set of all aspects, S is a subset of aspects excluding i, and f represents the model's prediction function [4,17]. In the context of current study of histopathology diagnosis, f is the prediction function of the 2-tier majority fusion process, and N would represent the set of all regions or features extracted from the histopathology image.

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{(|S|!(|N| - |S| - 1)!)}{|N|!} \times [f(S \cup \{i\}) - f(S)] \tag{3}$$

5 Experiments and Discussion

This study is conducted using a software environment based on Python. Windows 11 is used as the operating system, and Anaconda is selected as the distribution platform. The hardware includes an NVIDIA GPU with 6GB of memory, 256GB of RAM, and an Intel Core i7 processor.

To evaluate the performance, metrics such as its accuracy, precision, and recall [6]. Accuracy measures the proportion of correctly classified instances out of the total instances using $(TP + TN)/(TP + TN + FP + FN)$ where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives [8,30].

Precision measures the proportion of true positive predictions among all positive predictions made by the model using $TP/(TP + FP)$. Recall (sensitivity or true positive rate), measures the proportion of actual positive instances that were correctly identified by the model using $TP/(TP + FN)$ [10]. Specificity measures the proportion of actual negative instances that were correctly identified by the model using $TN/(TN + FP)$ [9]. Balanced accuracy (BAC) is the arithmetic mean of sensitivity and specificity. The F1 score is the harmonic mean of precision and recall, providing a balance between the two metrics sing $(2 \times \text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall})$ [5].

The utilized ViT Google pretrained models are B32-P16-224-In21K, B16-P16-224-In21K,, B32-P16-224, B16-P16-224, B32-P32-384, B16-P32-384, L32-P32-384, and L16-P32-384 where (B) is Base, (L) is Large, and (P) is Patch.

For instance, B32-P16-224-In21K refers to a base model that uses a batch size of 32, sequence of fixed-size patches (resolution 16x16), and pretrained on images sized at 224x224 pixels. Each experiment is performed 10 times, and the mean along with the standard deviation is reported for each metric. Table 1 reports the performance metrics for the ViT models and the best combinations for the three magnification levels utilizing the Post-NAT-BRCA dataset. Each row

At the 16x magnification level, the ViT model B32-P16-224-In21K with augmentation enabled performed the best overall. It achieved an impressive accuracy of 96.82% and a BAC of 96.25%, reflecting its strong overall performance in correctly identifying both positive and negative cases. Precision and recall were both high at around 94.5%, indicating a good balance between identifying true positives and minimizing false negatives. The F1 score and IoU were also robust, suggesting effective performance in capturing relevant features in the patches. On the other hand, models without augmentation showed a notable drop in performance, particularly in recall and F1 score, highlighting the importance of augmentation in enhancing the model's ability to generalize from the training data.

For the 4x magnification level, the B32-P16-224-In21K model with augmentation again stood out, achieving an accuracy of 98.11% and a BAC of 97.63%. This model demonstrated exceptional precision (96.66%) and recall (96.61%), resulting in a high F1 score and IoU, which were crucial for the model's overall effectiveness. Additionally, combinations of models, such as B32-P16-224-In21K with augmentation both enabled and disabled, provided slightly higher performance metrics, particularly in BAC and IoU. This suggests that combining modelscan enhance performance by utilizing the strengths of each individual approach.

At the 2x magnification level, the results mirrored those at the 4x level, with the B32-P16-224-In21K model with augmentation achieving the highest accuracy (98.06%) and BAC (97.61%). The precision, recall, and F1 score were consistently high, indicating reliable performance in detecting relevant features at this magnification. The model combination approach also showed effectiveness, with the best-performing combination achieving an accuracy of 98.47% and a BAC of 98.08%. This level of magnification benefits from the ability to capture more contextual information in the patches, contributing to the high performance across all metrics.

Across all magnification levels, models with augmentation consistently outperformed those without, and combinations of models provided marginal gains in performance, suggesting that diverse training strategies can be beneficial. The B32-P16-224-In21K model with augmentation was particularly effective across different magnifications, demonstrating its robustness and adaptability to varying magnification levels of patch detail.

Figures 3 explores the SHAP explainability across an example image four the three magnification levels. It shows how the fusion accumulates to get a final decision and probabilities.

Table 1. The reported performance metrics for the ViT models and the best combinations for the 3 magnification levels utilizing the Post-NAT-BRCA dataset.

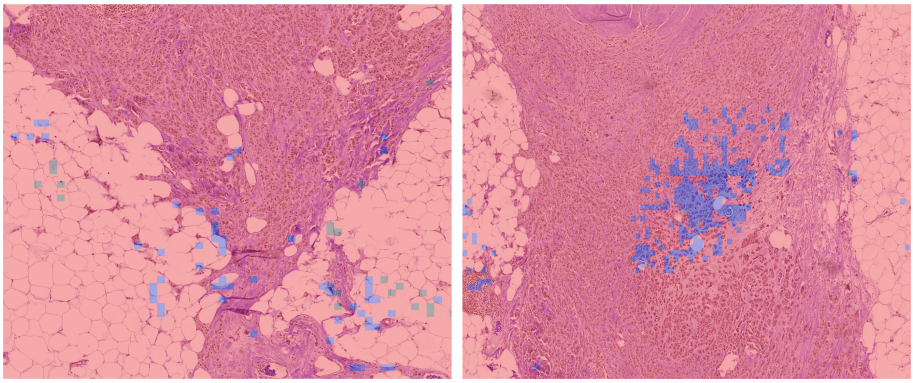
Model	AUG	ACC	PRC	Recall	SPC	F1	IoU	BAC
Magnification Level 16x Performance Metrics								
B32-P16-224-In21K	✓	96.82% ± 0.65%	94.53% ± 0.21%	94.61% ± 0.29%	97.89% ± 0.82%	94.58% ± 0.12%	89.69% ± 0.22%	96.25% ± 0.18%
B32-P16-224	✓	95.83% ± 0.48%	92.70% ± 0.55%	92.65% ± 0.32%	96.89% ± 0.79%	92.65% ± 0.29%	86.33% ± 0.57%	94.77% ± 0.41%
B32-P32-384	✓	96.74% ± 0.35%	94.28% ± 0.17%	94.21% ± 0.58%	97.47% ± 0.33%	94.22% ± 0.54%	89.06% ± 0.26%	95.82% ± 0.13%
B32-P16-224-In21K	✗	94.10% ± 0.89%	88.93% ± 0.23%	88.42% ± 0.17%	95.83% ± 0.21%	88.32% ± 0.41%	79.48% ± 0.45%	92.17% ± 0.28%
B32-P16-224	✗	95.37% ± 0.74%	91.53% ± 0.31%	91.42% ± 0.19%	95.94% ± 0.22%	91.31% ± 0.63%	84.21% ± 0.88%	93.69% ± 0.49%
B32-P32-384	✗	91.06% ± 0.52%	84.07% ± 0.13%	83.39% ± 0.34%	92.64% ± 0.29%	83.15% ± 0.16%	71.28% ± 0.39%	87.99% ± 0.44%
Magnification Level 4x Performance Metrics								
B32-P16-224-In21K	✓	98.11% ± 0.58%	96.66% ± 0.28%	96.61% ± 0.24%	98.65% ± 0.32%	96.59% ± 0.14%	93.31% ± 0.23%	97.63% ± 0.27%
B32-P16-224	✓	98.05% ± 0.32%	96.45% ± 0.25%	96.36% ± 0.31%	98.36% ± 0.39%	96.42% ± 0.26%	93.02% ± 0.34%	97.42% ± 0.16%
B32-P32-384	✓	97.92% ± 0.65%	96.25% ± 0.27%	96.35% ± 0.18%	98.42% ± 0.44%	96.34% ± 0.24%	92.88% ± 0.21%	97.29% ± 0.29%
L32-P32-384	✓	98.00% ± 0.72%	96.53% ± 0.19%	96.52% ± 0.27%	98.44% ± 0.33%	96.49% ± 0.16%	93.14% ± 0.39%	97.49% ± 0.18%
B32-P16-224-In21K	✗	98.03% ± 0.48%	96.30% ± 0.14%	96.42% ± 0.22%	98.57% ± 0.27%	96.32% ± 0.12%	92.94% ± 0.34%	97.48% ± 0.15%
B32-P16-224-In21K + B32-P16-224-In21K	✓/✗	98.45% ± 0.53%	97.22% ± 0.33%	97.26% ± 0.29%	98.85% ± 0.41%	97.32% ± 0.21%	94.79% ± 0.28%	98.05% ± 0.26%
Magnification Level 2x Performance Metrics								
B32-P16-224-In21K	✓	98.06% ± 0.51%	96.62% ± 0.22%	96.55% ± 0.33%	98.64% ± 0.42%	96.54% ± 0.27%	93.34% ± 0.12%	97.61% ± 0.19%
B32-P16-224	✓	97.99% ± 0.38%	96.38% ± 0.27%	96.38% ± 0.21%	98.36% ± 0.34%	96.42% ± 0.25%	93.10% ± 0.18%	97.39% ± 0.28%
B32-P32-384	✓	97.99% ± 0.29%	96.27% ± 0.31%	96.33% ± 0.17%	98.42% ± 0.26%	96.31% ± 0.24%	92.90% ± 0.29%	97.32% ± 0.34%
L32-P32-384	✓	98.01% ± 0.42%	96.45% ± 0.19%	96.43% ± 0.14%	98.45% ± 0.35%	96.42% ± 0.16%	93.17% ± 0.27%	97.47% ± 0.12%
B32-P16-224-In21K	✗	98.04% ± 0.37%	96.31% ± 0.22%	96.39% ± 0.12%	98.56% ± 0.28%	96.33% ± 0.21%	92.97% ± 0.26%	97.42% ± 0.15%
B32-P16-224-In21K + B32-P16-224-In21K	✓/✗	98.47% ± 0.32%	97.24% ± 0.26%	97.26% ± 0.31%	98.86% ± 0.24%	97.28% ± 0.22%	94.68% ± 0.29%	98.08% ± 0.18%

AUG: Augmentation, ACC: Accuracy, PRC: Precision, SPC: Specificity, IoU: Intersection over Union, and BAC: Balanced Accuracy. Models Naming: (B) is Base, (L) is Large, and (P) is Patch. For instance, B32-P16-224-In21K refers to a base model that uses a batch size of 32, sequence of fixed-size patches (resolution 16x16), and pretrained on images sized at 224x224 pixels. Each experiment is performed 10 times, and the mean along with the standard deviation is reported for each metric. Bold cells in the rows highlight the best models and combinations.

AUG: Augmentation, ACC: Accuracy, PRC: Precision, SPC: Specificity, IoU: Intersection over Union, and BAC: Balanced Accuracy. Models Naming: (B) is Base, (L) is Large, and (P) is Patch. For instance, B32-P16-224-In21K refers to a base model that uses a batch size of 32, sequence of fixed-size patches (resolution 16x16), and pretrained on images sized at 224x224 pixels. Each experiment is performed 10 times, and the mean along with the standard deviation is reported for each metric. Bold cells in the rows highlight the best models and combinations.

Fig. 3. This study explores the SHAP explainability across an example image four the three magnification levels. It shows how the fusion accumulates to get a final decision and probabilities.

Figure 4 illustrates the testing process of the proposed approach on multiple regions of a WSI. Majority voting between the models is utilized to determine the prediction of each patch. The patches are extracted from the 16x magnification level, with dimensions $(1024, 1024)$ and an overlap of 64 between adjacent patches. The results indicate that a significant portion of the patches are accurately diagnosed, underscoring the effectiveness of the proposed approach. The color overlays provide visual cues: green signifies healthy areas, blue represents low-risk regions, yellow indicates medium-risk areas, and red highlights high-risk regions. The first region, on the left, reflects a specific category, and the small overlays inside it indicate some misclassified areas/regions. These misclassifications can be corrected by applying post-processing using the surrounding ROIs. If the surrounding ROIs at a distance d are related to a specific category, then this region should be updated to reflect that category. This approach is considered successful as it corrects small regions within the very large WSI.



(a) First Region (b) Second Region

Fig. 4. Graphical presentation after testing regions using the suggested approach. The color overlays serve as visual indicators: green represents healthy regions, blue signifies low-risk areas, yellow denotes medium-risk regions, and red highlights high-risk regions.

5.1 Comparison with Pretrained CNNs

Compared with ViTs, pretrained CNN models like MobileNet, DenseNet, and VGG16/19 exhibit varying performance on a BC histopathology dataset. While these CNNs have been widely used and established in computer vision tasks, ViTs demonstrate superior performance in several aspects.

Table 2 showcases the comparison, where ViTs, especially the higher-resolution variants, outperform pretrained CNNs across multiple metrics. ViTs show exceptional accuracy, precision, recall, and specificity, with scores consistently surpassing 90%. In contrast, pretrained CNNs such as VGG16 and VGG19 exhibit slightly lower performance, indicating that traditional CNN architectures may not capture the intricate patterns present in histopathological images as effectively as ViTs. Additionally, ViTs demonstrate balanced scores across various evaluation metrics, highlighting their robustness in classification tasks.

Moreover, when considering different levels of magnifications as shown in Table 1, ViTs consistently outperform CNNs across different model configurations and augmentation settings. Augmentation generally improves performance across all models, but ViTs consistently achieve higher accuracy and other metrics compared to CNNs, even without augmentation.

5.2 External Benchmark Validation

The suggested approach is applied on an external dataset named Breast Cancer Histopathological Database (BreakHis) [27] to study its validations and generalization to other benchmarks. Table 3 compares the results of this study with those of other related studies on the BreakHis dataset. The results of the current study show a significant improvement over previous related studies on the

Table 2. Table showing various pretrained CNN models using the utilized breast cancer (BC) histopathology dataset.

Model	ACC	PRC	Recall	SPC	F1	IoU	BAC
MobileNet	90.75%	82.79%	82.85%	92.55%	82.69%	70.83%	87.70%
MobileNetV2	90.24%	82.44%	82.23%	91.68%	82.03%	69.89%	86.96%
DenseNet121	90.54%	82.56%	82.57%	92.49%	82.44%	70.44%	87.53%
DenseNet169	90.86%	83.07%	83.19%	92.84%	83.05%	71.33%	88.01%
DenseNet201	90.84%	83.34%	83.39%	92.81%	83.31%	71.65%	88.10%
VGG16	90.11%	82.13%	82.03%	91.72%	81.81%	69.43%	86.87%
VGG19	83.96%	72.61%	71.68%	86.00%	71.11%	55.42%	78.84%

ACC: Accuracy, PRC: Precision, SPC: Specificity, IoU: Intersection over Union, and BAC: Balanced Accuracy.

BreakHis dataset. Specifically, this study achieved an accuracy of 96.95%, surpassing the highest accuracy of 95.05%. Additionally, the precision of 97.46% and recall of 92.99% in this study indicate a robust performance in correctly identifying positive cases. The specificity of 98.84% further highlights the model's ability to accurately detect negative cases, reducing false positives. The F1 score of 95.17% and IoU of 90.79% also demonstrate an overall balanced and effective performance. This comprehensive set of metrics underscores the advancements and efficacy of the methods employed in the current study compared to prior research efforts.

Table 3. Comparison of this study's results with those of other related studies on the BreakHis dataset.

Study	Year	ACC	PRC	Recall	SPC	F1	IoU	BAC
Nahid et al. [19]	2018	92.19%	98.00%	94.94%	98.00%	-	-	-
Sharma and Mehra [26]	2020	93.97%	-	-	-	-	-	-
Agarwal et al. [3]	2022	94.67%	92.60%	80.52%	-	85.2%	-	-
Han et al. [15]	2017	94.90%	-	-	-	-	-	-
Zhang et al. [33]	2020	95.05%	-	-	-	-	-	-
Current Study	2024	96.95%	97.46%	95.99%	98.84%	95.17%	93.79%	95.92%

ACC: Accuracy, PRC: Precision, SPC: Specificity, IoU: Intersection over Union, and BAC: Balanced Accuracy.

5.3 Time Complexity Analysis

Time complexity analysis is crucial for assessing the computational efficiency of the proposed CAD framework based on ViTs for BC diagnosis from histopathology slides. In this section, we analyze the time complexity of key components of

the framework, including image preprocessing, ViT model inference, 2-Tier MF, and SHAP computation. Let N denote the number of patches extracted from a given WSI and F denote the number of features (e.g., pixels) per patch. The time complexity of image preprocessing is typically linear with respect to the number of patches, i.e., $O(N)$.

The time complexity of ViT model inference depends on the number of patches (N), the sequence length (L), and the number of layers (M) in the ViT architecture. For each patch, the self-attention mechanism computes attention scores with all other patches, resulting in a time complexity of $O(N^2 \times L \times M)$ for a single layer. With M layers, the overall time complexity becomes $O(N^2 \times L \times M^2)$. 2-Tier MF involves aggregating sub-decisions made at different magnification levels and sources. Let K denote the number of sub-decisions to be aggregated. The time complexity of 2-Tier MF is typically linear with respect to the number of sub-decisions, i.e., $O(K)$.

The time complexity of SHAP computation depends on the number of permutations of features, resulting in a complexity of $O(N! \times F)$ in the worst case. However, approximation methods such as sampling or kernel approximation can reduce the computational complexity to a more manageable level, often linear or logarithmic in N.

The overall time complexity of the proposed system can be approximated by summing the complexities of its constituent components. Assuming image preprocessing, ViT model inference, 2-Tier MF, and SHAP computation are performed sequentially, the overall complexity can be approximated as the sum of the individual complexities as: $O(N) + O(N^2 \times L \times M^2) + O(K) + O(N! \times F)$. The dominant factors influencing the overall time complexity are typically N, L, and M.

5.4 Clinical Relevance in Enhancing BC Diagnosis

The suggested approach holds significant clinical relevance in enhancing BC diagnosis. By utilizing advanced ViTs and self-attention mechanisms, this method can more accurately classify histopathology slides, potentially leading to earlier and more precise BC diagnoses. Traditional histopathological analysis is often time-consuming and subjective, depending heavily on the expertise of pathologists. The proposed approach addresses these limitations by treating images as sequences of patches from different magnification levels and capturing global dependencies more effectively than conventional CNNs. This results in improved diagnostic accuracy and efficiency. Moreover, integrating the RCBi, a clinically validated tool for assessing responses to neoadjuvant therapy, enhances the method's clinical applicability and prognostic value. Additionally, the use of SHAP ensures the interpretability and reliability of the diagnostic process, providing insights into model decision-making. This transparency is crucial for clinical adoption, as it helps build trust in AI-driven diagnostics. *The suggested approach represents a significant advancement in the field of digital pathology, offering a robust, interpretable, and efficient tool for BC diagnosis.*

6 Conclusions and Future Directions

Breast cancer (BC) is a significant public health concern globally, demanding accurate and timely diagnosis for effective treatment and management. The study introduces a CAD framework for BC diagnosis from histopathology slides. Based on the findings and experiments conducted in this study, several key conclusions can be drawn regarding the proposed CAD framework for BC diagnosis using histopathology slides. The integration of ViTs with advanced decision-making techniques like 2-Tier MF and SHAP has demonstrated promising results in enhancing both the accuracy and interpretability of BC diagnosis. Firstly, ViTs proved to be highly effective in classifying histopathology slides across different magnification levels. The experimental results consistently showed that ViT models, especially when augmented, outperformed traditional CNN architectures like VGG16 and DenseNet in terms of accuracy, precision, recall, and specificity. This superiority can be attributed to ViTs' ability to capture global dependencies and contextual information within histopathological images, which are crucial for accurate diagnosis.

Secondly, the 2-tier majority fusion approach significantly improved decision-making by aggregating predictions from multiple ViT models at each magnification level. This ensemble method not only enhanced classification accuracy but also provided robustness against potential biases or errors that may arise from individual models. The hierarchical fusion process, combining outputs at both patch and slide levels, effectively utilized the strengths of different models, leading to more reliable diagnostic outcomes. Thirdly, the incorporation of SHAP for explainable AI added a layer of transparency to the diagnostic process. By attributing the contribution of each feature and region within histopathology slides to the final diagnosis, SHAP enhanced the interpretability of ViT predictions. This interpretability is crucial for gaining insights into the decision-making process of AI models, ensuring that clinicians and researchers can trust and validate the model's decisions.

Moreover, the study's validation on external datasets, such as the BreakHis, demonstrated the generalizability and robustness of the proposed framework. Achieving superior performance compared to existing methods on these benchmarks underscores the efficacy of ViTs coupled with advanced decision fusion techniques in automated BC diagnosis.

Future directions in BC diagnosis should prioritize collecting diverse multimodal data and expanding datasets. Additionally, Integrating various imaging modalities such as MRI with histopathological data can enhance diagnostic accuracy. Finally, increasing the volume and diversity of annotated WSIs will enable the development of more robust models, especially for rare BC subtypes.

References

1. Aboudessouki, A., Ali, K.M., Elsharkawy, M., Alksas, A., Mahmoud, A., Khalifa, F., Ghazal, M., Yousaf, J., Khalifeh, H.A., El-Baz, A.: Automated diagnosis of

breast cancer using deep learning-based whole slide image analysis of molecular biomarkers. In: 2023 IEEE International Conference on Image Processing (ICIP). pp. 2965–2969 (2023). https://doi.org/10.1109/ICIP49359.2023.10222479
2. Aebi, S., Davidson, T., Gruber, G., Cardoso, F.: Primary breast cancer: Esmo clinical practice guidelines for diagnosis, treatment and follow-up. Annals of oncology **22**, vi12–vi24 (2011)
3. Agarwal, P., Yadav, A., Mathur, P.: Breast cancer prediction on breakhis dataset using deep cnn and transfer learning model. In: Data Engineering for Smart Systems: Proceedings of SSIC 2021. pp. 77–88. Springer (2022)
4. Aljadani, A., Alharthi, B., Farsi, M.A., Balaha, H.M., Badawy, M., Elhosseini, M.A.: Mathematical modeling and analysis of credit scoring using the lime explainer: A comprehensive approach. Mathematics **11**(19), 4055 (2023)
5. Badawy, M., Balaha, H.M., Maklad, A.S., Almars, A.M., Elhosseini, M.A.: Revolutionizing oral cancer detection: An approach using aquila and gorilla algorithms optimized transfer learning-based cnns. Biomimetics **8**(6), 499 (2023)
6. Baghdadi, N.A., Malki, A., Balaha, H.M., AbdulAzeem, Y., Badawy, M., Elhosseini, M.: Classification of breast cancer using a manta-ray foraging optimized transfer learning framework. PeerJ Computer Science **8**, e1054 (2022)
7. Baghdadi, N.A., Malki, A., Balaha, H.M., Badawy, M., Elhosseini, M.: A3c-tl-gto: Alzheimer automatic accurate classification using transfer learning and artificial gorilla troops optimizer. Sensors **22**(11), 4250 (2022)
8. Balaha, H.M., Hassan, A.E.S.: Comprehensive machine and deep learning analysis of sensor-based human activity recognition. Neural Comput. Appl. **35**(17), 12793–12831 (2023)
9. Balaha, H.M., Saif, M., Tamer, A., Abdelhay, E.H.: Hybrid deep learning and genetic algorithms approach (hmb-dlgaha) for the early ultrasound diagnoses of breast cancer. Neural Comput. Appl. **34**(11), 8671–8695 (2022)
10. Boursalie, O., Samavi, R., Doyle, T.E.: Evaluation metrics for deep learning imputation models. In: International Workshop on Health Intelligence. pp. 309–322. Springer (2021)
11. Chan, R.C., To, C.K.C., Cheng, K.C.T., Yoshikazu, T., Yan, L.L.A., Tse, G.M.: Artificial intelligence in breast cancer histopathology. Histopathology **82**(1), 198–210 (2023)
12. Das, A.K., Biswas, S.K., Bhattacharya, A., Alam, E.: Introduction to breast cancer and awareness. In: 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS). vol. 1, pp. 227–232. IEEE (2021)
13. Elmannai, H., Hamdi, M., AlGarni, A.: Deep learning models combining for breast cancer histopathology image classification. International Journal of Computational Intelligence Systems **14**(1), 1003 (2021)
14. Gamal, A., Sharafeldeen, A., Alnaghy, E., Alghandour, R., Saleh Alghamdi, N., Ali, K.M., Shamaa, S., Aboueleneen, A., Elsaid Tolba, A., Elmougy, S., Ghazal, M., Contractor, S., El-Baz, A.: A novel machine learning approach for predicting neoadjuvant chemotherapy response in breast cancer: Integration of multimodal radiomics with clinical and molecular subtype markers. IEEE Access **12**, 104983–105003 (2024). https://doi.org/10.1109/access.2024.3432459
15. Han, Z., Wei, B., Zheng, Y., Yin, Y., Li, K., Li, S.: Breast cancer multi-classification from histopathological images with structured deep learning model. Sci. Rep. **7**(1), 4172 (2017)
16. Huang, Y., Wang, X., Cao, Y., Li, M., Li, L., Chen, H., Tang, S., Lan, X., Jiang, F., Zhang, J.: Multiparametric mri model to predict molecular subtypes of breast

cancer using shapley additive explanations interpretability analysis. Diagnostic and Interventional Imaging (2024)
17. Kim, Y., Kim, Y.: Explainable heat-related mortality with random forest and shapley additive explanations (shap) models. Sustain. Urban Areas **79**, 103677 (2022)
18. Martel, A., Nofech-Mozes, S., Salama, S., Akbar, S., Peikari, M.: Assessment of residual breast cancer cellularity after neoadjuvant chemotherapy using digital pathology. The Cancer Imaging Archive (2019)
19. Nahid, A.A., Kong, Y.: Histopathological breast-image classification using local and frequency domains by convolutional neural network. Information **9**(1), 19 (2018)
20. Nohara, Y., Matsumoto, K., Soejima, H., Nakashima, N.: Explanation of machine learning models using shapley additive explanation and application for real data in hospital. Comput. Methods Programs Biomed. **214**, 106584 (2022)
21. Parkin, D.M., Fernández, L.M.: Use of statistics to assess the global burden of breast cancer. Breast J. **12**, S70–S80 (2006)
22. Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., Tran, D.: Image transformer. In: International conference on machine learning. pp. 4055–4064. PMLR (2018)
23. Peikari, M., Salama, S., Nofech-Mozes, S., Martel, A.L.: Automatic cellularity assessment from post-treated breast surgical specimens. Cytometry A **91**(11), 1078–1087 (2017)
24. Rakha, E.A., Reis-Filho, J.S., Baehner, F., Dabbs, D.J., Decker, T., Eusebi, V., Fox, S.B., Ichihara, S., Jacquemier, J., Lakhani, S.R., et al.: Breast cancer prognostic classification in the molecular era: the role of histological grade. Breast Cancer Res. **12**, 1–12 (2010)
25. Shamshad, F., Khan, S., Zamir, S.W., Khan, M.H., Hayat, M., Khan, F.S., Fu, H.: Transformers in medical imaging: A survey. Medical Image Analysis p. 102802 (2023)
26. Sharma, S., Mehra, R.: Conventional machine learning and deep learning approach for multi-classification of breast cancer histopathology images–a comparative insight. J. Digit. Imaging **33**(3), 632–654 (2020)
27. Spanhol, F., Oliveira, L., Petitjean, C., Heutte, L.: Breast cancer histopathological database (breakhis) (2021)
28. Vaswani, A., Parmar, N., Uszkoreit, J., Shazeer, N., Kaiser, L.: Image transformer (2018)
29. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
30. Vujović, Ž, et al.: Classification model evaluation metrics. Int. J. Adv. Comput. Sci. Appl. **12**(6), 599–606 (2021)
31. Wang, X., Zou, C., Zhang, Y., Xie, L., Zhang, Y.: Prediction of brca gene mutation in breast cancer based on deep learning and histopathology images. Front. Genet. **12**, 661109 (2021)
32. Wilkinson, L., Gathani, T.: Understanding breast cancer as a global health concern. Br. J. Radiol. **95**(1130), 20211033 (2022)
33. Zhang, J., Wei, X., Dong, J., Liu, B.: Aggregated deep global feature representation for breast cancer histopathology image classification. Journal of Medical Imaging and Health Informatics **10**(11), 2778–2783 (2020)

Directed Brain Network Transformer for Psychiatric Diagnosis

Xu Zhu[1,2], Zhiwei Qi[2,3(✉)], Kun Yue[1,2], Yunshan Su[4], and Liang Duan[1,2]

[1] School of Information Science and Engineering, Yunnan University, Kunming, China
xuzhustu@mail.ynu.edu.cn, {kyue,duanl}@ynu.edu.cn
[2] Yunnan Key Laboratory of Intelligent Systems and Computing, Yunnan University, Kunming, China
[3] School of Education, Yunnan University, Kunming, China
maryqizhiwei@ynu.edu.cn
[4] Department of Radiology, Affiliated Hospital of Yunnan University, Kunming, China

Abstract. Human brain is a complex organ that consists of billions of neurons and trillions of connections among the neurons. To describe the correlations among the time series of the brain regions, we model the brain as a functional brain network to diagnose psychosis. However, traditional methods of functional brain network construction are usually noisy and do not consider the causal relationships among brain regions. To obtain the causal relationships and improve diagnosis interpretability, we propose a directed brain network Transformer (DBNT) for psychiatric diagnosis. First, the causal relationships in the blood-oxygen-level-dependent time series of brain regions are extracted to generate a directed brain network. Then, the feature encoding method is proposed to obtain local and global features of the brain networks by using the DBNT. Experimental results demonstrate that the accuracy of DBNT increases by 8.1% and 6.4% compared to state-of-the-art methods on two large-scale brain network datasets. DBNT also highlights the brain regions associated with psychosis and provides interpretation for diagnosis.

Keywords: Brain Network · Psychiatric Diagnosis · Causal Relationship · Transformer · Feature Encoding

1 Introduction

The behavioral, emotional and cognitive deficiencies of patients with psychosis have a major effect on their lives [27]. Millions of people suffer from psychosis,

Supported by the Program of Yunnan Key Laboratory of Intelligent Systems and Computing (202405AV340009) and Yunnan Fundamental Research Project (202401AT070462).

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15311, pp. 207–221, 2025.
https://doi.org/10.1007/978-3-031-78195-7_14

and the yearly expense of medical care exceeds trillions of dollars [18]. The current diagnosis of psychiatric disorders relies on the subjective judgment of clinicians regarding patients' cognitive, behavioral and emotional functioning [15]. Most treatments are ineffective due to the lack of understanding of pathology of psychosis, although there are many treatments available that address subjective symptoms and psychology [21]. The brain network-based psychiatric diagnosis is becoming increasingly important in neuroimage studies to understand the organization of the brain of psychiatric patients and healthy controls [11].

Functional magnetic resonance imaging (fMRI) is developed for brain network construction [26]. There are two key steps by using fMRI to diagnose psychosis: functional brain network construction and psychiatric prediction [9]. The first step is to construct functional brain network from an individual's fMRI data. A set of regions of interest (ROI) on the brain atlases is selected as the nodes of the functional brain networks, and pairwise connectivities of blood-oxygen-level-dependent (BOLD) signal time series from each ROI are extracted as the edges [7]. For the constructed functional brain networks, some brain regions (also named nodes in the functional brain network) are co-activated or co-deactivated simultaneously when performing action, language and vision [28]. By using this pattern, the nodes of functional brain networks could be classified into diverse functional modules to analyze psychosis for their diagnosis. Thus, the second step is to feed the functional brain network into a classification model to predict whether a person has psychosis or not. However, the brain network-based psychiatric diagnosis methods are noisy and inaccurate, since these methods focus on capturing the statistical associations of ROIs rather than that of causal relationships [26].

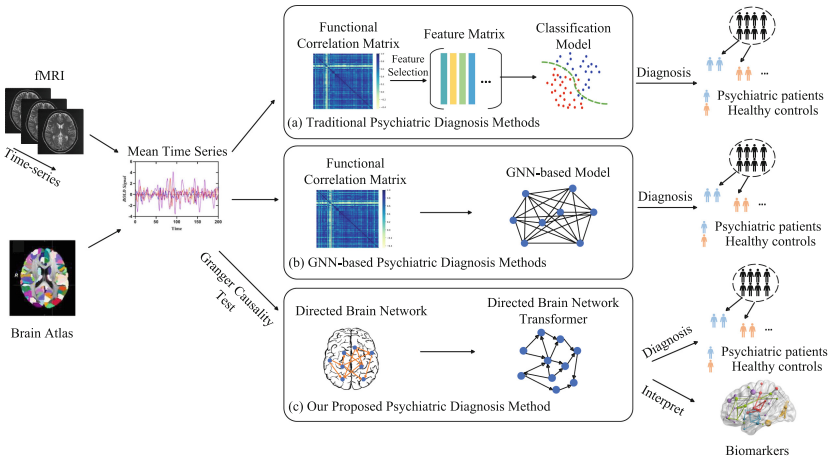


Fig. 1. The overall pipeline for (a) Traditional psychiatric diagnosis methods. (b) GNN-based psychiatric diagnosis methods. (c) Our proposed psychiatric diagnosis methods. For our proposed diagnostic method, a directed brain network is constructed by performing GC test on the mean time series, and inputted into the DBNT.

As shown in Fig. 1 (a), the traditional psychiatric diagnosis methods are based on shallow models or linear classifiers, which cannot capture the topological and nonlinear relationships of complex brain networks [29]. Recently, as shown in Fig. 1 (b), graph neural network (GNN) has gained significant attention for their exceptional ability to analyze graph-structured data [6]. The GNN-based psychiatric diagnosis methods capture the structural features of the brain network, but only consider the local structure of neighboring nodes [17]. This localized propagation limits the ability of the GNN to handle global information and long-range dependencies. To capture both local and global structures, graph Transformer-based methods inject edge features into the attention mechanism and leverage the eigenvectors of each node as positional embeddings [10,22]. However, the above methods suffer from the following challenges:

Ignorance of causality. Current psychiatric diagnosis methods are based on the nodes' correlations of brain networks, and ignore the causal relationships among brain regions and may lead to inaccurate results.

Unaffordable calculation cost. The number of edges in traditional brain networks can reach tens of thousands, and the computational cost of the generation of all edge features in graph Transformer-based methods is prohibitive.

Lack of brain network properties. Brain network is a special network with unique properties, such as node and topology features, which have not been considered in the traditional methods and lead to poor interpretability for psychiatric diagnosis.

To address the first challenge, we propose the method for directed brain network construction to test whether the causal relationships exist among brain regions. To obtain the causal relationships, we adopt the univariate and multivariate Granger causality (GC) test on the BOLD time series of the brain regions, since the current BOLD time series correlate with previous BOLD time series from its own and other brain regions. The existing causal relationships will be represented by the direction of the edges, and thus the directed brain network can be generated.

To address the second challenge, we adopt the significance test to exclude spurious causal relationships in the directed brain network and reduce the computational cost of graph Transformer. Specifically, we adopt the significance test to exclude these spurious causal relationships if and only if the level of significance is not larger than the given threshold. Accordingly, the number of edges in the directed brain networks is reduced from tens of thousands to a few thousand.

To address the third challenge, we design three encoding methods in our directed brain network Transformer (DBNT) to capture the node centrality, global feature, and path feature of the brain networks, respectively. On the basis of the unique node features such as clustering coefficient, node efficiency and betweenness centrality [24], we first propose a node centrality encoding method to identify the important nodes associated with psychosis. A global feature encoding method is designed to generate the global structure features of the brain networks. To utilize the weight and direction of the edges, we design a path feature encoding method of the shortest path. With the three encoding meth-

ods, DBNT can capture the unique properties of brain networks and generate network embeddings for psychiatric diagnosis.

To summarize, our contributions are as follows:

- We propose the method for constructing the directed brain network by performing GC tests on the BOLD time series of different brain regions to generate the directed network and eliminate spurious causal relationships.
- We propose a DBNT with three types of encoding methods to enhance the extraction capability of node centrality, global features and path features in brain networks by encoding their unique properties.
- We conduct experiments on two large-scale brain network datasets. Experimental results demonstrate that DBNT is optimal on all metrics and can identify the brain regions associated with psychosis.

The rest of this paper is organized as follows: Section 2 introduces related work. Section 3 gives our methodology. Section 4 presents experimental results and brain network visualization of DBNT. Section 5 concludes and discusses future work.

2 Related Work

GNN based Brain Network Analysis. The brain networks can be seen as complex graph structures that have recently attracted widespread interest due to the power of GNN in different downstream tasks. Li et al. [17] proposed BrainGNN to analyze fMRI and discover neurological biomarkers which has novel ROI-aware graph convolutional layers to leverage the topological and functional information of fMRI. Cui et al. [6] proposed BrainGB, a benchmark for brain network analysis with GNN to summarize brain network construction pipelines for both functional and structural neuroimaging modalities. FBNETGEN [11] can generate the GNN-compatible functional brain networks from fMRI data and dynamically optimize them for downstream tasks. Zheng et al. [29] proposed a Granger causality-inspired GNN to identify the most influential subgraph that is causally related to the prediction. The GNN-based method calculates the pairwise connectivities of each ROI through Pearson correlation and generates a brain network. However, Pearson correlation cannot capture the causal relationships among the nodes. Therefore, brain networks generated based on correlations are noisy and inaccurate, which impede the understanding of the brain network structure.

Graph Transformer based Brain Network Analysis. Transformer has achieved great success in NLP and CV tasks due to the self-attention mechanism [16]. Graph Transformer is proposed by leveraging the self-attention in graph representation learning, which has recently been applied to brain network analysis. Kan et al. [14] proposed a brain network Transformer (BNT) that combines the self-supervised soft clustering and orthogonal clustering readout operations based on orthogonal projection. By leveraging this cluster read, BNT can generate distinguishable cluster-aware node embeddings and infographic

embeddings. Wei et al. [8] proposed a Transformer-based hierarchical clustering model for brain network analysis by learning a globally shared cluster allocation via a Transformer encoder and continuously adapted cluster allocation to downstream tasks. Anushree et al. [1] proposed Com-BrainTF, a hierarchical local-global transformer architecture by learning node embeddings within and between communities for autism spectrum disorder prediction tasks. These graph Transformer-based methods are performed on undirected brain networks to obtain their representations. However, undirected brain networks are complete graphs that hinder the differentiation of node features and the computational cost of generating edge features is unaffordable in graph Transformer.

3 Methodology

3.1 Directed Brain Network Construction

As shown in Fig. 1 (c), our proposed diagnostic method aims to predict psychosis on the given BOLD time series dataset $\mathcal{D} = \{D_q\}_{q=1}^{n} \in \mathbb{R}^{n \times V \times T}$, where n is the number of patients and healthy controls, V is the number of ROI, T is the length of time series, D_q is the data of the q-th patient or healthy individual in $\mathcal{D}$. Our method contains two components: directed brain network construction and DBNT. The directed brain networks $G_D \in \mathbb{R}^{n \times V \times V}$ are generated by conducting the GC test on $\mathcal{D}$ and fed into the DBNT to obtain network embeddings $\mathbf{Z} = \{\mathbf{z}_q\}_{q=1}^{n}$. Then, $\mathbf{z}_q$ is flattened and passed to a classifier for prediction $\hat{Y} \in \mathbb{R}^{n \times |\mathfrak{c}|}$, where $\mathfrak{c}$ is the set of classes $\hat{Y}$.

According to the GC theory [20], for two time series $\mathcal{X}$ and $\mathcal{Y}$, $\mathcal{X}$ is deemed to be causal of $\mathcal{Y}$ if leveraging the history of $\mathcal{X}$ does reduce the variance of the prediction of $\mathcal{Y}$. To explore the causal relationships among brain regions, the GC test is adopted to generate directed brain networks. Specifically, upon corresponding time series $D_{q,t}^{v_i}$ and $D_{q,t}^{v_j}$ of two brain regions v_i and v_j in D_q, if the variance σ of the residual ε of multivariate model is lower than that of univariate model, then there exists a causal relationship. Therefore, the residuals ε_i and ε_j should be first calculated by the univariate model of $D_{q,t}^{v_i}$ and $D_{q,t}^{v_j}$, where the univariate model is formulated as:

$$D_{q,t}^{v_i} = \sum_{m=1}^{\tau_i} \alpha D_{q,t-m}^{v_i} + \varepsilon_i \tag{1}$$

$$D_{q,t}^{v_j} = \sum_{m=1}^{\tau_j} \beta D_{q,t-m}^{v_j} + \varepsilon_j \tag{2}$$

where α and β are constant coefficients, τ_i and τ_j are the lag time step of the univariate model for $D_{q,t}^{v_i}$ and $D_{q,t}^{v_j}$. τ should be adjusted according to the specific situation to obtain the appropriate value.

Then, we calculate ε'_{ji} and ε'_{ij} of the multivariate model, defined as:

$$D_{q,t}^{v_i} = \sum_{m=1}^{\tau_i} \alpha D_{q,t-m}^{v_i} + \sum_{m=1}^{\tau_j} \gamma D_{q,t-m}^{v_j} + \varepsilon'_{ji} \tag{3}$$

$$D_{q,t}^{v_j} = \sum_{m=1}^{\tau_j} \beta D_{q,t-m}^{v_j} + \sum_{m=1}^{\tau_i} \delta D_{q,t-m}^{v_i} + \varepsilon'_{ij} \tag{4}$$

where γ and δ are constant coefficients.

If the variance σ_{ε_i} is higher than $\sigma_{\varepsilon'_{ji}}$ of the multivariate model's residual, the causal relationship $\mathcal{F}_{v_j \to v_i}$ of $D_{q,t}^{v_i}$ and $D_{q,t}^{v_j}$ will exist. Thus, the different causal relationships $\mathcal{F}_{v_j \to v_i}$ of $D_{q,t}^{v_j} \to D_{q,t}^{v_i}$ and $\mathcal{F}_{v_i \to v_j}$ of $D_{q,t}^{v_i} \to D_{q,t}^{v_j}$ are denoted as:

$$\mathcal{F}_{v_j \to v_i} = \ln \frac{\sigma_{\varepsilon_i}}{\sigma_{\varepsilon'_{ji}}} \tag{5}$$

$$\mathcal{F}_{v_i \to v_j} = \ln \frac{\sigma_{\varepsilon_j}}{\sigma_{\varepsilon'_{ij}}} \tag{6}$$

If $\mathcal{F}_{v_j \to v_i} > 0$, a causal relationship $\mathcal{F}_{v_j \to v_i}$ exists between $D_{q,t}^{v_j}$ and $D_{q,t}^{v_i}$, which means that there is a directed edge $v_j \to v_i$ in the brain network G_{D_q}. Similarly, if $\mathcal{F}_{v_i \to v_j} > 0$, a causal relationship exists between $D_{q,t}^{v_i}$ and $D_{q,t}^{v_j}$, which means that there is a directed edge v_i to v_j in G_{D_q}. To eliminate spurious causal relationships, we conduct the F-test to evaluate the causal relationships of the BOLD time series. Specifically, we first make the assumption and select a level of significance θ, then calculate the F-value. If $F_{ji} > F_\theta$, the original assumption will be rejected. The F-value F_{ji} is defined as :

$$F_{ji} = \frac{(MSE_i - MSE_{ji})(T - 2\tau_j)}{MSE_{ji}\tau_j} \tag{7}$$

where MSE_i and MSE_{ji} represent the mean squared residuals of ε_i and ε'_{ji}.

We first assume that no causal relationship $\mathcal{F}$ exists between $D_{q,t}^{v_i}$ and $D_{q,t}^{v_j}$ and select the level of significance $\theta \leq 0.01$. If $F_{ji} \geq F_{0.01}$, the assumption is rejected and the causal relationship $\mathcal{F}_{v_j \to v_i}$ is real rather than spurious.

3.2 Directed Brain Network Transformer

In this section, we propose the node centrality encoding, global feature encoding, and path feature encoding shown in Fig. 2 to capture the unique properties and global features and path features of the brain networks.

Node Centrality Encoding In a general graph, the importance of a node is primarily measured by its degree centrality, which is the sum of its in-degree and out-degree. In order to measure the node importance, the unique node's

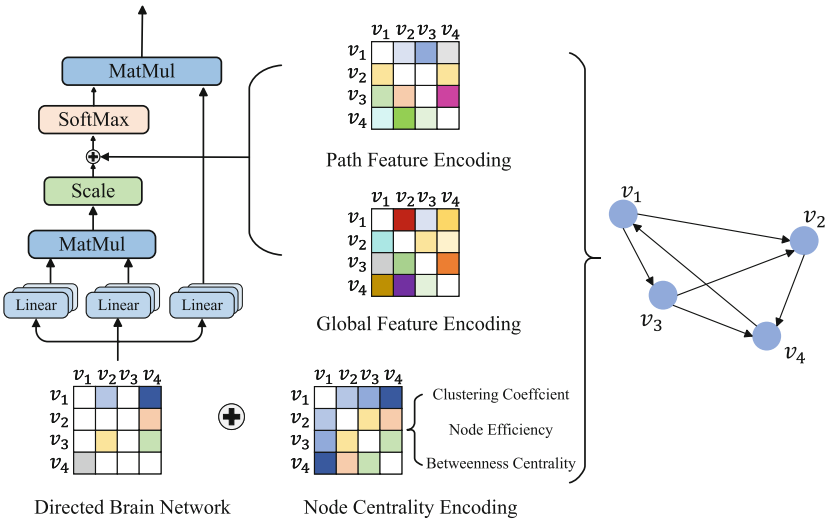


Fig. 2. Illustration of our proposed node centrality encoding, global feature encoding, and path feature encoding.

properties of the brain networks, such as clustering coefficient, node efficiency and betweenness centrality, could be adopted [24]. The clustering coefficient C_i of v_i is defined as the ratio of the number of edges between adjacent nodes to the maximum number of edges between these adjacent nodes. C_i measures the degree of clustering among nodes and reflects the efficiency of information processing in local brain regions, defined as:

$$C_i = \frac{2e_i}{k_i(k_i - 1)} \tag{8}$$

where k_i is the number of neighbours of v_i, e_i represents the number of the edges.

The node efficiency E_i of v_i is the most reliable measure of node importance in brain networks [24]. E_i reflects the efficiency of information exchange within the sub-network and network fault tolerance, defined as:

$$E_i = \frac{1}{N_{G_i}(N_{G_i} - 1)} \sum_{j \neq k \in G_i} \frac{1}{\ell_{j,k}} \tag{9}$$

where G_i represents the sub-network of neighboring nodes connected to v_i. N_{G_i} represents the number of nodes in G_i. $\ell_{j,k}$ represents the length of the shortest path from v_j to v_k.

The betweenness centrality B is a measure of a node's influence on the overall flow of information in the network. A node v_i with large B_i indicates most of the shortest paths among other nodes must pass through v_i.

$$B_i = \sum_{i \neq j \neq k \in G} \frac{\varsigma_{j,k}(i)}{\varsigma_{j,k}} \tag{10}$$

where $\varsigma_{j,k}$ is the number of all shortest paths from v_j to v_k, and $\varsigma_{j,k}(i)$ is the number of the shortest paths through v_i.

Thus, the initial node feature $h_i^{(0)}$ of v_i can be expressed as the sum of the node feature vector $\mathbf{x}_i$ and the node centrality, defined as:

$$h_i^{(0)} = \mathbf{x}_i + \mathbf{z}^-_{deg^-(v_i)} + \mathbf{z}^+_{deg^+(v_i)} + C_i + E_i + B_i \tag{11}$$

where $\mathbf{z}^-$ and $\mathbf{z}^+$ are learnable embedding vectors specified by the indegree $deg^-_{v_i}$ and outdegree $deg^+_{v_i}$ respectively.

Global Feature Encoding The Transformer is originally designed for modeling the sequence data instead of brain network data. To obtain the global structural features of brain networks by using Transformer, we propose the global feature encoding function $\phi(v_i, v_j)$ to choose the distance of shortest path between v_i and v_j, where $\phi(v_i, v_j)$ denotes the positional relationship between v_i and v_j. If there is not a path from v_i to v_j, then $\phi(v_i, v_j) = -1$.

Furthermore, the locations of brain regions associated with psychosis are typically adjacent or in close proximity to each other [3]. We set a learnable scalar b of the function b_ϕ as a bias term in the self-attention module of Transformer so that each node in brain network can adaptively focus on other nodes. When b is a decreasing function w.r.t. $\phi(v_i, v_j)$, each node will pay more attention to its neighbors, which is consistent with the brain region location relationships of psychosis. A_{ij} is the (i, j)-element of the Query-Key product matrix A.

$$A_{ij} = \frac{(h_i W_Q)(h_j W_K)^T}{\sqrt{d}} + b_{\phi(v_i, v_j)} \tag{12}$$

where h_i and h_j are the node features of v_i and v_j. W_Q and W_K are two learnable weight matrices in Transformer. d is the dimension of h_i.

Path Feature Encoding The signal transmission in the brain is directional and often show coactivation among the brain regions associated with psychosis, which means that brain regions on the signal transmission path are correlated. For this, we propose path feature encoding. The path feature $l_{i,j}$ is defined as the dot product of node features h_i and edge features e_i on the shortest path $v_i \rightarrow v_j$ as follows:

$$l_{i,j} = \sum_{i=1}^{j} h_i (e_i \mathbf{w}_i)^T \tag{13}$$

where e_i represents the edge feature of the i-th edge in the shortest path $v_i \rightarrow v_j$, and $\mathbf{w}_e$ is the weight embedding of the e_i.

To enhance the feature extraction capability of Transformer for the shortest path, the path feature $l_{i,j}$ is merged into the attention module as a bias term that aims to modify A_{ij}.

$$A_{ij} = \frac{(h_i W_Q)(h_j W_K)^T}{\sqrt{d}} + b_{\phi(v_i,v_j)} + \sum_{i=1}^{j} h_i (e_i \mathbf{w}_i)^T \tag{14}$$

4 Experiments

In this section, we conduct extensive experiments on two large-scale brain network datasets to evaluate the effectiveness and interpretability of DBNT.

4.1 Datasets

Our experiments are conducted on two large-scale brain network datasets:

(1) **Autism Brain Imaging Data Exchange[1](ABIDE).** This dataset includes resting-state functional magnetic resonance imaging (rs-fMRI) data from 17 international sites, and all data are anonymous [4], and contains brain networks from 1009 subjects, with 516 being Autism spectrum disorder (ASD) patients. The definition of brain region is based on Craddock 200 atlas [5].

(2) **Attention Deficit Hyperactivity Disorder[2](ADHD).** The ADHD dataset contains 947 rs-fMRI and anatomical datasets clustered in 8 separate imaging sites, of which 585 rs-fMRI were from normally developing individuals and 362 rs-fMRI from children and adolescents with ADHD [2]. The regional definitions of ADHD are based on anatomical automatic labeling 164 ROI maps.

Task: We choose psychiatric diagnosis as an assessment task on ABIDE and ADHD datasets, which are split into the train/valid/test datasets in a ratio of 7:1:2 for evaluation.

Evaluation Metrics: As the label distributions of both ABIDE and ADHD datasets are balanced, we use the area under the receiver operating characteristic curve (AUROC), accuracy, and F1-Score as the performance metrics. For accuracy, we use 0.5 as the threshold after obtaining the diagnosis result.

4.2 Performance Comparison

The proposed DBNT is compared to the following three types of baselines:

– **CNN-based Model for Brain Network Analysis.** BrainNetCNN [13] utilizes graph convolution operations to capture structure features.

[1] https://fcon_1000.projects.nitrc.org/indi/abide/
[2] http://fcon_1000.projects.nitrc.org/indi/adhd200/

– **GNN-based Models for Brain Network Analysis.** Graph Attention Network (GAT) [23] and FBNETGEN [11] utilize GNN models to learn and analyze brain networks.
– **Transformer-based Models for Brain Network Analysis.** Graph Transformer [14] utilizes Transformer to learn node embeddings, and BNT [12] design a global pooling operator to generate graph-level embeddings.

Table 1. Performance comparison with five baselines.

Method	Dataset: ABIDE			Dataset: ADHD		
	AUROC	Accuracy	F1-Score	AUROC	Accuracy	F1-Score
GAT	60.2±3.4	56.4±2.6	57.7±1.8	68.2±3.5	63.7±3.1	62.1±3.6
BrainNetCNN	63.1±2.0	58.6±2.2	59.2±2.6	68.7±2.2	61.6±3.1	60.7±2.6
FBNETGEN	62.4±1.2	57.8±1.8	58.0±1.9	71.2±1.7	64.8±1.6	64.3±1.4
BNT	61.3±2.3	55.7±2.3	57.2±2.2	67.9±2.6	61.3±2.5	60.4±2.2
Graph Transformer	62.2±1.0	57.2±1.3	58.4±1.8	69.3±1.4	62.2±1.7	61.6±1.5
DBNT	**70.7±2.6**	**66.7±2.5**	**66.4±2.3**	**75.8±3.1**	**71.2±2.6**	**70.4±2.3**

To ensure fairness, a grid search is applied to all baselines, and the best one is reported in Table 1. The experimental results demonstrate that our proposed DBNT consistently outperforms the baselines with higher AUROC, accuracy, and F1-Score. DBNT outperforms other baselines by 7.6% and 4.6% in AUROC, 8.1% and 6.4% in accuracy, and 7.2% and 6.1% in F1-Score on the two datasets. These results show that brain networks generated by Pearson correlations are inaccurate, since they fail to capture the complex causal relationships and these baselines do not consider the unique properties of brain networks.

4.3 Ablation Studies and Hyperparameter Experiments

For ablation studies in Table 2, we modify the original DBNT by using all encoding methods to observe the performance of the DBNT variant by removing one of the components. From the final results, the performance of DBNT with all design encoding methods outperforms others. In particular, the performance of the model with node centrality encoding and global feature encoding is close to the optimal performance, and the effect of global feature encoding is the best among the three encoding methods.

For hyperparameter experiments, the lag time step τ has an impact on the performance of DBNT. Table 3 reveals that DBNT achieves the best performance on both datasets when the τ is 2 and 3. This suggests that these specific lag lengths are most effective in capturing and incorporating relevant temporal dependencies in the brain networks. It is worth noting that the complexity of the directed brain network construction by the ABIDE dataset poses a challenge.

Table 2. Ablation Studies. The impact of different encoding methods on DBNT. (node means Node Centrality Encoding, GF means Global Feature Encoding, PF means Path Feature Encoding.)

Method	Dataset: ABIDE			Dataset: ADHD		
	AUROC	Accuracy	F1-Score	AUROC	Accuracy	F1-Score
DBNT	70.7±2.6	66.7±2.5	66.4±2.3	75.8±3.1	71.2±2.6	70.4±2.3
DBNT_node	63.1±1.6	58.3±3.7	59.4±2.1	70.1±3.3	62.3±3.6	61.4±3.3
DBNT_GF	65.5±1.3	59.1±1.6	60.3±1.4	71.8±3.0	63.5±3.3	62.3±3.6
DBNT_PF	63.7±1.8	59.3±2.1	59.9±2.5	71.2±2.9	63.1±3.4	62.9±3.4
DBNT_node+GF	66.8±1.6	62.4±2.0	62.3±2.7	73.7±2.4	66.7±2.6	65.6±2.8
DBNT_node+PF	65.4±1.5	61.8±2.9	62.0±2.4	72.4±2.7	65.8±2.5	64.9±2.4
DBNT_GF+PF	66.1±1.3	62.1±2.8	61.3±2.7	73.5±2.5	66.2±2.8	65.8±2.2

Table 3. Impact of the lag time step on DBNT.

Lag time step (τ)	Dataset: ABIDE		Dataset: ADHD	
	AUROC	Accuracy	AUROC	Accuracy
τ =1	63.9±1.9	60.3±2.4	71.2±4.0	65.2±3.5
τ =2	**70.7±2.6**	**66.7±2.5**	73.7±2.7	68.1±3.3
τ =3	67.4±2.0	64.1±2.2	**75.8±3.1**	**71.2±2.6**
τ =4	–	–	72.7±3.5	67.3±3.1
τ =5	–	–	70.2±3.6	65.9±3.8

Considering the trade-off between performance and computation complexity, when the length of lag time steps exceeds 3 ($\tau > 3$), τ is set to 2 or 3 that becomes a good choice for DBNT on the ABIDE dataset.

4.4 Brain Network Visualization

To demonstrate our DBNT can identify the brain regions associated with psychosis, we visualize the brain network using BrainNet Viewer [25], and the results are shown in Fig. 3. The figure visually demonstrates the brain network connections learned by DBNT on two psychiatric datasets. Specifically, DBNT divides the learned graphs according to their class labels and calculates the average network by averaging the weights of each edge in the same class. Fig. 3 (a) and Fig. 3 (b) show the connections between Autistic patients and healthy controls. Fig. 3 (c) and Fig. 3 (d) show the connections between ADHD patients and healthy controls. The results are summarized as follows:

– Fig. 3 (a) and Fig. 3 (b) demonstrate that notable variations are observed in the contact regions responsible for memory and emotion in the brains of ASD patients, and the frontal lobe regions influence attention and judgment.

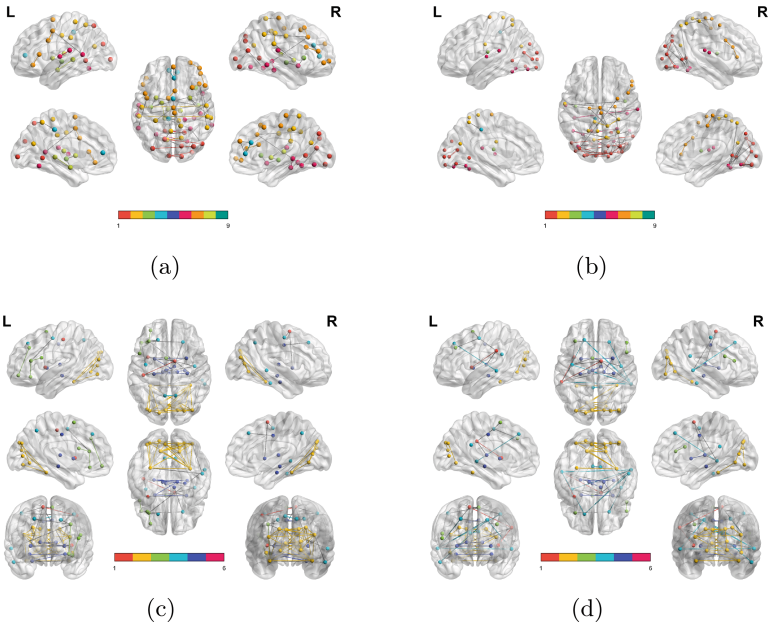


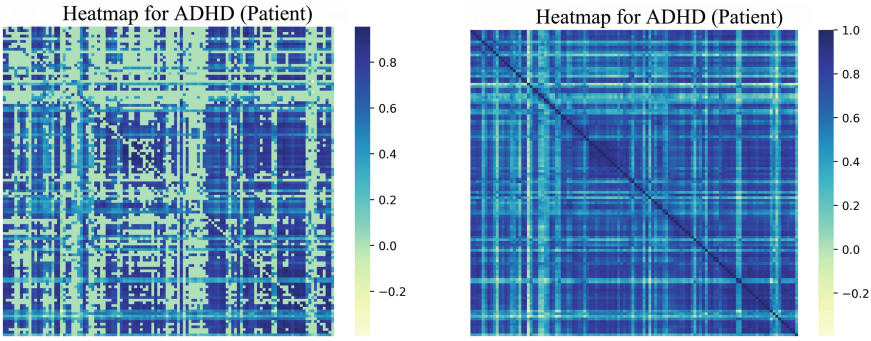
Fig. 3. Visualizations of predominant connectivities generated via DBNT on different biological sexes. Edges spanning multiple neural systems are colored gray, whereas those linking nodes within the same neural system are colored appropriately.

As compared to previous studies [19], these changes are primarily evident in the connections between brain regions, such as the prefrontal cortex, parietal lobe and inferior parietal lobe.

– Fig. 3 (c) and Fig. 3 (d) demonstrate that patients with ADHD predominantly exhibit changes in the motor center of the brain. Specifically, the changes are observed in the connections between brain regions, such as the prefrontal cortex, amygdala, and attention network, when compared to the healthy controls. These visual representations provide valuable insights into the specific regions and connections that undergo significant changes in individuals with different psychosis, shedding light on the underlying neural mechanisms associated with these conditions.

The heat map of a brain network can visualize whether the network is sparse or dense. Fig. 4 shows the heat map of the directed brain network and the baselines. For the ADHD dataset, the edge number in the generated directed brain network is reduced by approximately 40%. The heat map generated by the baselines exhibits a high degree of density, with a multitude of negative edges present. This high density suggests strong connectivities among brain regions, potentially indicating redundant or spurious connections. The baselines demonstrate a fully connected network, while conversely DBNT produces a heat map with sparse connections between different brain regions. This sparsity indicates a selective

and focused representation of brain connections, which highlights the meaningful and relevant signals. The sparse nature of the connections implies a more efficient and discriminative representation of the underlying brain networks.



(a) DBNT (b) Baseline

Fig. 4. Generated brain networks of DBNT and baseline.

4.5 Summary

The above experimental results are summarized as follows:

- Our proposed DBNT achieves the best performance improving by at least 4% over the suboptimal FBNETGEN in all metrics, demonstrating the effectiveness of DBNT for psychiatric diagnosis.
- DBNT containing three encoding methods performs the best, which implies DBNT can capture the node, global and path features of brain networks.
- Brain connection changes in ASD patients are centered on the prefrontal, amygdala, and hippocampus regions, and in ADHD patients are centered on the regions of the prefrontal lobe, cerebellum and the default mode network.

5 Conclusion

In this paper, we propose directed brain network Transformer for psychiatric diagnosis including the directed brain network construction and DBNT implementation. To construct directed brain network, the GC test is used to explore causal relationships among brain regions. The generated network overcomes the problem of edge redundancy and improves diagnostic accuracy. To capture the unique properties and global features of brain networks, we design three encoding methods that enhance the feature extraction capability of DBNT. Experiments on two large-scale brain network datasets demonstrate DBNT's superior performance. Brain network visualization reveals the ability of DBNT to identify brain

regions related to psychosis and display the connection changes consistent with clinical observations.

In the future, we will consider exploring a method for determining the optimal lag time step to enhance the efficiency of generating directed brain networks. To improve the generalization ability of DBNT, an research issue is to utilize pre-training and meta-learning techniques for few-shot learning, which can somewhat address the limitations imposed by small and difficult-to-acquire datasets.

References

1. Bannadabhavi, A., Lee, S., Deng, W., Ying, R., Li, X.: Community-aware transformer for autism prediction in fmri connectome. In: Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 287–297 (2023)
2. Bellec, P., Chu, C., Chouinard-Decorte, F., Benhajali, Y., Margulies, D.S., Craddock, R.C.: The neuro bureau adhd-200 preprocessed repository. NeuroImage pp. 275–286 (2017)
3. Canario, E., Chen, D., Biswal, B.: A review of resting-state fmri and its use to examine psychiatric disorders. Psychoradiology **1**(1), 42–53 (2021)
4. Craddock, C., Benhajali, Y., Chu, C., Chouinard, F., Evans, A., Jakab, A., Khundrakpam, B.S., Lewis, J.D., Li, Q., Milham, M.: The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives. Front. Neuroinform. **7**(27), 5 (2013)
5. Craddock, R.C., James, G.A., Holtzheimer, P.E., III., Hu, X.P., Mayberg, H.S.: A whole brain fmri atlas generated via spatially constrained spectral clustering. Hum. Brain Mapp. **33**(8), 1914–1928 (2012)
6. Cui, H., Dai, W., Zhu, Y., Kan, X., Gu, A.A.C., Lukemire, J., Zhan, L., He, L., Guo, Y., Yang, C.: Braingb: A benchmark for brain network analysis with graph neural networks. IEEE Trans. Med. Imaging **42**(2), 493–506 (2022)
7. Cui, H., Dai, W., Zhu, Y., Yang, C.: Interpretable graph neural networks for connectome-based brain disorder analysis. In: Proceedings of Medical Image Computing and Computer Assisted Intervention. p. 375-385 (2022)
8. Dai, W., Cui, H., Kan, X., Guo, Y., van Rooij, S., Yang, C.: Transformer-based hierarchical clustering for brain network analysis. In: Proceedings of The IEEE International Symposium on Biomedical Imaging. pp. 1–5 (2023)
9. Delvigne, V., Wannous, H., Vandeborre, J.P., Ris, L.: Spatio-temporal analysis of transformer based architecture for attention estimation from eeg. In: Proceedings of International Conference on Pattern Recognition. pp. 1076–1082 (2022)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Proceedings of International Conference on Learning Representations (2021)
11. Kan, X., Cui, H., Lukemire, J., Yang, C.: Fbnetgen: Task-aware gnn-based fmri analysis via functional brain network generation. In: Proceedings of International Conference on Medical Imaging with Deep Learning. pp. 618–637 (2022)
12. Kan, X., Dai, W., Yang, C.J.: Brain network transformer. In: Proceedings of Advances in Neural Information Processing Systems (2022)

13. Kawahara, J., Brown, C.J., Miller, S.P., Booth, B.G., Chau, V., Grunau, R.E., Zwicker, J.G.: Brainnetcnn: Convolutional neural networks for brain networks; towards predicting neurodevelopment. Neuroimage **146**, 1038–1049 (2017)
14. Kreuzer, D., Beaini, D., Hamilton, W., Létourneau, V., Tossou, P.: Rethinking graph transformers with spectral attention. In: Proceedings of Advances in Neural Information Processing Systems. vol. 34, pp. 21618–21629 (2021)
15. Krystal, J.H.: Psychiatric disorders: diagnosis to therapy. Cell **157**(1), 201–214 (2014)
16. Li, L., Duan, L., Wang, J., He, C., Chen, Z., Xie, G., Deng, S., Luo, Z.: Memory-enhanced transformer for representation learning on temporal heterogeneous graphs. Data Science and Engineering **8**(2), 98–111 (2023)
17. Li, X., Zhou, Y., Dvornek, N., Zhang, M., Gao, S., Zhuang, J., Scheinost, D., Staib, L.H., Ventola, P., Duncan, J.S.: Braingnn: Interpretable brain graph neural network for fmri analysis. Med. Image Anal. **74**, 102233 (2021)
18. Scangos, K.W., State, M.W., Miller, A.H., Baker, J.T.: New and emerging approaches to treat psychiatric disorders. Nat. Med. **29**(2), 317–333 (2023)
19. Scheltens, P., De Strooper, B., Kivipelto, M., Holstege, H., Chételat, G., Teunissen, C.E., Cummings, J., van der Flier, W.M.: Alzheimer's disease. The Lancet **397**(10284), 1577–1590 (2021)
20. Shojaie, A., Fox, E.B.: Granger causality: A review and recent advances. Annual Review of Statistics and Its Application **9**, 289–319 (2022)
21. Tong, Y., Li, Z., Huang, H., Gao, L., Xu, M., Hu, Z.: Research of spatial context convolutional neural networks for early diagnosis of alzheimer's disease. The Journal of Supercomputing pp. 1–19 (2023)
22. Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A., Li, Y.: Maxvit: Multi-axis vision transformer. In: Proceedings of European conference on computer vision. pp. 459–479 (2022)
23. Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y.: Graph attention networks. In: Proceedings of The International Conference on Learning Representations (2018)
24. Wu, Q., Lei, H., Mao, T., Deng, Y., Zhang, X., Jiang, Y., Zhong, X.: Test-retest reliability of resting brain small-world network properties across different data processing and modeling strategies. Brain Sci. **13**(5), 825 (2023)
25. Xia, M., Wang, J., He, Y.: Brainnet viewer: a network visualization tool for human brain connectomics. PLoS ONE **8**(7), 689–699 (2013)
26. Yu, Y., Kan, X., Cui, H., Xu, R., Zheng, Y., Song, X., Zhu, Y., Zhang, K.: Deep dag learning of effective brain connectivity for fmri analysis. In: Proceedings of IEEE International Symposium on Biomedical Imaging. pp. 1–5 (2023)
27. Zhang, R., Volkow, N.D.: Seasonality of brain function: role in psychiatric disorders. Transl. Psychiatry **13**(1), 65 (2023)
28. Zhdanov, M., Steinmann, S., Hoffmann, N.: Investigating brain connectivity with graph neural networks and gnnexplainer. In: Proceedings of International Conference on Pattern Recognition. pp. 5155–5161 (2022)
29. Zheng, K., Yu, S., Chen, B.: Ci-gnn: A granger causality-inspired graph neural network for interpretable brain network-based psychiatric diagnosis. arXiv preprint arXiv:2301.01642 (2023)

Learning to Estimate Motion Between Non-adjacent Frames in Cardiac Cine MRI Data: A Fusion Approach

Nicolas Portal[1,2(✉)], Thomas Dietenbeck[2], Saud Khan[2], Vincent Nguyen[2], Mikael Prigent[3], Mohamed Zarai[3], Khaoula Bouazizi[2,3], Johanne Sylvain[4], Alban Redheuil[2,3,5], Gilles Montalescot[4], Nadjia Kachenoura[2,3], and Catherine Achard[1]

[1] Sorbonne Université, Institut des Systèmes Intelligents et de Robotique, ISIR, 75005 Paris, France
portal@isir.upmc.fr
[2] Sorbonne Université, Laboratoire d'Imagerie Biomédicale, LIB, 75006 Paris, France
[3] Institut de Cardiométabolisme et Nutrition (ICAN), 75013 Paris, France
[4] Sorbonne Université, ACTION Group, Pitié-Salpêtrière Hospital (AP-HP), 75013 Paris, France
[5] Imagerie Cardio-Thoracique (ICT), Sorbonne Université, AP-HP, Groupe Hospitalier Pitié-Salpêtrière, 75013 Paris, France

Abstract. Information about the motion of pixels between images is crucial for many computer vision tasks. When dealing with cardiac sequences, information about the heart's motion can help physicians diagnose pathologies. Most methods that try to estimate this motion rely on pair of frames. This can lead to suboptimal performance when the amount of motion between them is important as it is the case when considering distant frames in a video sequence. Moreover, performing registration image by image leads to the integration of registration errors and is also suboptimal. In this work, a new registration method that uses all the frames in a video sequence is presented and applied to cardiac cine-MRI in short-axis views. A first neural network is used to compute motion flows between adjacent frames. Then, a second one processes the output of the first network to merge motion flows according to the time dimension throughout the sequence. Estimated flows are used to propagate segmentation masks across the sequence. The method is tested on an in-house dataset containing 271 patients. Segmentation, similarity and motion flow regularization metrics are computed to assess the model performance. The proposed approach achieves an average registration Dice score and SSIM between the end-diastole and end-systole frame of 95.26 ± 0.01 and 86 ± 0.05 respectively against 93.24 ± 0.02 and 82.75 ± 0.06 for the best Voxelmorph version.

Keywords: Cardiac · Deep Learning · Image registration · Motion estimation

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15311, pp. 222–237, 2025.
https://doi.org/10.1007/978-3-031-78195-7_15

1 Introduction

In cardiac imaging, it is interesting to monitor the way parts of the heart behave during its contraction and expansion phase. To carry out this investigation, a common biomarker is the cardiac strain that represents, for each cardiac phase, the percent of deformation of a cardiac structure relative to its end-diastolic shape [9]. These strain curves are usually obtained using echocardiography sequences, but conventional or Deep-Learning feature-tracking algorithms [21,26,28] have recently been proposed in dynamic Magnetic Resonance Imaging (cine-MRI).

Approximating pixels apparent motion between two frames and registering these two frames are related tasks. Indeed, non-rigid image registration algorithms consist in finding a non-linear transformation to align a moving image with a fixed image. Motion flows, which represent the motion of pixels in the x and y direction are a specific kind of non-linear transformation that can be learned by a deep learning algorithm. The use of ground truth optical flow data is the preferred approach to estimate these motion flows as this allows the network to learn in a fully supervised manner. As a consequence, in the past few years, many datasets with ground truth optical flows were created [3,25,27]. However, images found in these datasets are very different from medical images and the process to generate this data is very time-consuming.

In the medical imaging domain, motion flow estimation studies have mainly focused on unsupervised image registration techniques. These methods generally try to optimize both a similarity error between the warped image and the fixed image, and a regularization error to obtain a smooth displacement field [1,31,43]. Similarity errors usually consist in minimizing the absolute or squared difference between the fixed image and the warped image. The cross correlation [23] and mutual information [32] loss functions are also regularly used as a similarity measure. Regularization error typically consist in minimizing the norm of the gradient of the velocity field [1,40]. As such, these methods do not require any ground truth deformation flow and are fully unsupervised.

However, according to [1,13,14] when segmentation labels are available, the addition of a segmentation error to the two previous loss components leads to more accurate results. However, algorithms are then limited in the number of images used in training since few frames have a ground truth segmentation annotation. For example, in the field of cardiac imaging, only the end-diastole and end-systole frames are typically labelled [2,4]. These frames correspond to the time of the cardiac cycle when the heart is the most relaxed and contracted respectively and exhibit larger motion than any other possible pair of frames in the video. As a consequence, learning to predict this motion flow is more challenging than for a pair of frames closer in the video and for which the motion is smaller. However, and as evidenced in this work, estimating the motion between adjacent frames and iteratively composing this motion leads to the integration of registration errors and is also suboptimal. To tackle this issue, we present a new registration method where flows are first computed between adjacent frames and then aggregated using a trained network to avoid multiple interpolations.

2 Related work

2.1 Optical flow methods

Under the assumption of brightness constancy, it is possible to compute the motion of pixels between two images. Before the advent of deep learning, algorithms were based on computer vision principles and mathematical models. [24] developed a technique which consisted in solving a system of linear equations under the assumption of constant displacement in the local neighborhood of pixels. Block matching algorithms, which try to minimize a similarity measure between blocks of pixels in the first and the second image are also regularly used [17,45]. Other methods are based on phase information, which is robust to illumination change [10,29]. These traditional methods for optical flow estimation relies on handcrafted features and assumptions. They may not generalize well to images with complex non-rigid motion, illumination variation and noise. Recently, deep learning based approaches have gained popularity. [7] presented two different architectures diverging in the format of the data they take as input and the presence of a correlation layer at the bottleneck. Several approaches opted to use a spatial pyramid to infer the flow in a coarse to fine manner [33,36]. This allows for a more progressive estimation, able to better handle large displacement. These approaches have in common the iterative nature of their flow estimation process and the use of a warping function at each resolution of the pyramid. [16] also tackled the optical flow task using an iterative approach where several networks are stacked to refine the prediction process. [37], designed a new architecture relying on a recurrent layer at the bottleneck of the network to iteratively refine the estimated optical flow by processing a multi-resolution cost volume. The cost volume is a fundamental concept in optical flow estimation which represents the similarity between patches. It is represented as a 4D tensor where for each pixel in an image, its correlation with every pixel of the second image is computed. Following this latter work, many architectures also relied upon an iterative mechanism to process the cost volume [11,20]. Recently, multiple methods used transformer layers to better take advantage of the cost volume [15,35].

2.2 Motion flow for medical image registration

In the medical domain, motion flows have been primarily used for image registration which has been a topic of active research for decades with important developments in the past few years. Before the widespread use of deep learning based techniques, some medical image registration algorithms computed the optical flow by minimizing an energy function using a variational model [12,30]. With the emergence of deep learning, first methods focused on learning a sparse motion flow, that is, the displacement of a set of control points using either a Thin Plate Spline (TPS) approach or B-spline approach for interpolation [5,8,34]. The introduction of the spatial transformer by [19] made it easier to directly compute dense motion flows by using a grid-based fully differentiable module at

the output of a Convolutional Neural Network (CNN). This is why it has been widely adopted for medical image registration. Since ground truth motion flow is difficult to obtain, most of this work focused on unsupervised image registration [22,39]. Recently, many studies showed that using segmentation annotations can lead to improved registration performance with more realistic motion flows. These approaches usually use a segmentation loss between the deformed annotation of the moving image and the annotation of the fixed image [13,14]. However, these methods greatly reduce the number of available images since ground truth segmentation labels are usually not available for many of them, especially in videos. Recently, some methods tried to integrate new regularization penalties that allow to generate more physically plausible motions [31,44]. Others applied the scaling and squaring method to iteratively integrate the learned displacement field in order to produce a diffeomorphic motion, greatly reducing the number of pixels with negative Jacobian determinant [6,42].

3 Method

Let $F_{t-1,t} \in \mathbb{R}^{w \times h \times 2}$ denote the motion flow representing the vertical and horizontal displacement of pixels from the frame $I_{t-1} \in \mathbb{R}^{w \times h}$ to the frame I_t. Given a video sequence of T frames $S = \{I_1, ..., I_T\}$, where only the end-diastole (ED) and end-systole (ES) frames I_{ED} and I_{ES} are provided with ground truth segmentation labels, we want to estimate the motion $F_{1,t}$ of every pixel between the first frame I_1 and every other frame $I_t \; \forall t \in [2; T]$. The underlying hypothesis of this work is that, the smaller the amount of motion between two frames, the easier it is to estimate the displacement of pixels. Based on this principle, an algorithm could be designed to estimate $F_{1,t}$ by relying on elementary flows $F_{t-1,t}$, since these latter should contain a small amount of error.
In this work, such an algorithm is designed around 2 neural networks f_1 and f_2 with the same architecture but parametrized by different weights. f_1 estimates the motion between adjacent frames $F_{t-1,t} \; \forall t \in [2; T]$. Then f_2 progressively fuses these flows in a successive manner in order to obtain the motion between non-adjacent frames $F_{1,t}$.

3.1 Iterative motion aggregation

To begin with, the first network f_1 takes as input a sequence S and outputs a set of $T-1$ motion flows $F_{t-1,t} \forall t \in [2; T]$ between adjacent frames. Then, starting with $F_{1,2}$, that is directly obtained as output of the first network f_1, the second network f_2 iteratively aggregates motion flows and builds $F1, t$ using $F_{1,t-1}$ and the elementary flow $F_{t-1,t}$. To be more specific, let us consider two consecutive times $t-1$ and t in the cardiac cycle. The flow $F_{t-1,t}$ is estimated between the corresponding images I_{t-1} and I_t using the network f_1: $F_{t-1,t} = f_1(I_{t-1}, I_t)$. The spatial transformer described in [19] allows to compute the registered image $R_{t,t-1}$ obtained after warping the frame I_t towards I_{t-1} with the flow $F_{t-1,t}$.

Then, the registration error $E_{t,t-1} = R_{t,t-1} - I_{t-1}$ between $R_{t,t-1}$ and I_{t-1} is computed and used to create: $X_{1,t-1}$ and $X_{t-1,t}$ as the concatenation:

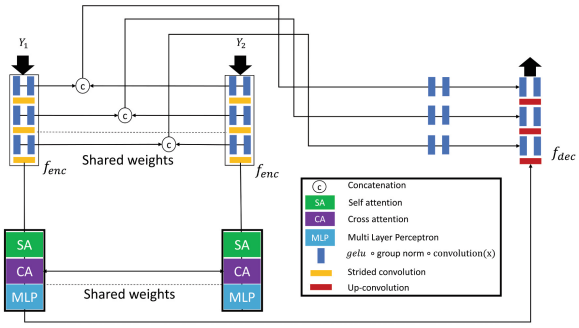
$$X_{1,t-1} = [F_{1,t-1}; I_1; I_{t-1}; R_{t-1,1}; E_{t-1,1}] \quad X_{t-1,t} = [F_{t-1,t}; I_{t-1}; I_t; R_{t,t-1}; E_{t,t-1}] \quad (1)$$

Finally, our iterative motion aggregation process, that computes the flow between the first frame and the frame t is obtained using the second network f_2 and the equation:

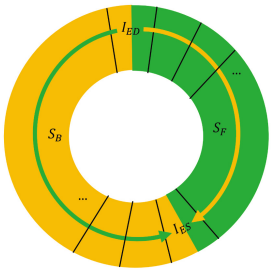
$$F_{1,t} = F_{1,t-1} + f_2(X_{1,t-1}, X_{t-1,t}) \quad \forall t \in [3; T] \quad (2)$$

3.2 Network architecture

Since both f_1 and f_2 process a pair of inputs, for the sake of simplicity, it was decided to use the same architecture for both network. This architecture is depicted in Figure 1a. f_1 and f_2 both use a siamese-like encoder with shared weights. Given a pair of inputs Y_1 and Y_2 (Y is I for f_1 and X for f_2), the encoder f_{enc} is used twice to extract features $F_{Y_1} = f_{enc}(Y_1)$ and $F_{Y_2} = f_{enc}(Y_2)$ separately. Then, the transformer layer found in the decoding part of the network proposed by [38] is used so that each feature map F_{Y_1} and F_{Y_2} pay attention to each other. Finally, after this cross-attention block, only the feature map F_{Y_1} is passed to the decoder f_{dec} in order to get the final output motion flow. Low-level feature maps of same resolution estimated by the encoder for both inputs Y_1 and Y_2, are concatenated and go through 2 convolution layers to halve the number of features. Then, they are concatenated a second time with corresponding same-resolution feature maps from the decoder (U-net like architecture).



(a) Architecture of f_1 and f_2.



(b) S_F and S_B in the cardiac cycle.

Fig. 1. Network architecture and training process

3.3 Training and inference procedures

A similar process, which takes advantage of the cyclic nature of the cardiac cycle, is followed for training and inference. The cine-MRI sequence, covering

the whole cardiac cycle, is split in two halves S_F or S_B which correspond to the forward and backward motion starting with the ED frame and ending with the ES frame as described in Figure 1. For training, due to GPU memory constraints, we randomly select S_F or S_B. Then $N-2$ frames are uniformly sampled in the sequence while keeping I_{ED} as the first frame and I_{ES} as the last one in order to obtain a sequence of length N. Thus, only the first and last frame of the sequence have a ground truth segmentation label. For inference, S_F and S_B are processed as a whole by the model without using any sampling mechanism.

3.4 Loss functions

The loss function $\mathcal{L}$ is composed of several components. The one which assesses the similarity between the registered image $R_{t,t-1}$ and I_{t-1} is defined as:

$$\mathcal{L}_{sim}(I_{t-1}, R_{t,t-1}) = 1 - \text{NCC}(I_{t-1}, R_{t,t-1}) \tag{3}$$

Where NCC refers to the Normalized Cross Correlation function. A segmentation loss $\mathcal{L}_{seg}$ is also employed between the ground truth ED segmentation label Y_{ED} and the ED segmentation obtained by warping the ground truth end-systole segmentation label Y_{ES} using the flow $F_{1,N}$.

$$\mathcal{L}_{seg} = f_{seg}(Y_{ED}, Y_{ES} \circ F_{1,N}) \tag{4}$$

f_{seg} is a combination of Dice loss and cross-entropy as proposed in [18]. Finally, to ensure that smooth motion flows are generated, the function $\mathcal{L}_{smooth}$ minimizes the gradient of flows in the x and y directions:

$$\mathcal{L}_{smooth}(F) = \frac{1}{|\Omega|} \sum_{p \in \Omega} \|\nabla F(p)\|_2^2 \tag{5}$$

This smoothness cost function is applied to all outputs of f_1 and f_2 $F_{t-1,t}$ and $F_{1,t}$ respectively $\forall t \in [2; N]$. Thus, the final loss function $\mathcal{L}$ is defined as:

$$\begin{aligned} \mathcal{L} = & \frac{1}{N-1} \sum_{t=2}^{N} [\lambda_1 \mathcal{L}_{sim}(I_{t-1}, R_{t,t-1}) + \lambda_2 \mathcal{L}_{smooth}(F_{t-1,t})] \\ & + \frac{1}{N-1} \sum_{t=2}^{N} [\lambda_1 \mathcal{L}_{sim}(I_1, R_{t,1}) + \lambda_2 \mathcal{L}_{smooth}(F_{1,t})] + \lambda_3 \mathcal{L}_{seg} \end{aligned} \tag{6}$$

In our experiments, the best results were achieved with $\lambda_1 = 0.5, \lambda_2 = 1.0, \lambda_3 = 1.0$.

4 Experiments

4.1 Dataset and Pre-processing

A clinical dataset of short-axis cardiac cine MRI data is used to measure the performance of the approach. The dataset contains 271 patients. Data for one

patient is made up of several cardiac sequences stacked to form a sequence of 3D volumes and covering one cardiac cycle. Each sequence contains between 20 and 80 volumes and each volume is a stack of 3 to 7 slices. In total, the dataset contains 912 video sequences of 2D slices. Each sequence covers one cardiac cycle. The average pixel size and slice thickness is $0.97mm^2$ and $7.47mm$ respectively. The dataset has been processed by a local software which provides segmentation labels of the left ventricular cavity (LV), right ventricular cavity (RV) and myocardium (MYO) for all frames [21]. Note that only the end-diastole and end-systole segmentation labels are used for training while all labels are used to compute metrics on the test set. 80% of the dataset was used for training and the rest for testing. 20% of the training set was used for validation. The separation is carried out randomly and patients in each of the 3 sets are different. A pre-trained binary segmentation network is used to crop each frame to a size of 192x192 around the cardiac structures. Before being processed by the model, every frame of an input sequence is standardized based on the mean and standard deviation of the sequence.

4.2 Evaluation Metrics

Similarity, segmentation and regularization metrics are computed to evaluate the model. For similarity metrics, the Structural Similarity Index (SSIM) between the ED frame and registered frames of each sequence is computed. When it comes to segmentation evaluation, the segmentation label of each frame of a sequence are registered towards the ED frame using the predicted motion flow $F_{1,t}$. This is done for every sequence in each volume. Then the Dice score, Hausdorff Distance (HD) and Average Symmetric Surface Distance (ASSD) are computed between the registered segmentations for the whole volume and the ground truth segmentation of the ED volume (registrations are performed slice-wise but segmentation metrics are computed for the whole volume). We keep the largest connected component as a post-processing step before computing segmentation metrics. For regularization evaluation, the average percentage of pixels with negative Jacobian determinant $det(J_F(p))$ is computed on 2D motion flows for pixels within the heart. Ground truth segmentation masks are used to compute $det(J_F(p))$ per cardiac structure. For each kind of metric, results are reported only for the registered ES frame as well as average over all registered frames of the sequence. The Wilcoxon signed rank test is used to assess the statistical significance of results and a p-value < 0.05 is considered statistically significant.

4.3 Baseline Methods

The proposed approach is compared against the semi-supervised version of VoxelMorph presented by [1]. This model is retrained from scratch on our dataset using the online tensorflow implementation. Performances of the model trained with the MSE loss function (VM-MSE) and NCC loss (VM-NCC) are reported.

Results for the diffeomorphic version are also presented (VM-dif). These Voxelmorph models are only trained with the labeled ED and ES frames. The optimal set of hyperparameters reported in the article is kept. During inference, as Voxelmorph models take only 2 frames as input, we iterate over the sequence and, for each step t, the first frame of the sequence $I_1 = I_{ED}$ and the current frame I_t are passed to the network. The number of filters in the original Voxelmorph model is increased so that the network contains approximately 25 M parameters, the same number of parameters as our proposed model. This led to slightly better performance than the default configuration. In addition to these Voxelmorph models, comparisons with the classical medical image registration algorithm SyN (Symmetric Normalization) is also presented.
Moreover, our results also include 2 "naive" methods called IterWarpImg and IterWarpFlow which, during inference, iteratively compose deformation fields to register each image to the ED image. These models are trained with $N = 2$ where only adjacent frames are fed into the network. During inference, IterWarpImg iteratively deforms contour points from the current image I_t to the first image of the sequence I_1:

$$R_{t,1} = ((((I_t \circ F_{t-1,t}) \circ F_{t-2,t-1}) \circ \ldots \circ F_{3,2}) \circ F_{2,1}) \tag{7}$$

As described in [41], IterWarpFlow iteratively adds motion fields after aligning them to obtain deformation fields between non-adjacent frames. Thus, to obtain $R_{t,1}$, the image I_t registered to image I_1, the following equation is used:

$$F_{1,t} = F_{1,t-1} + (F_{t-1,t} \circ F_{1,t-1})$$
$$R_{t,1} = I_t \circ F_{1,t}$$

These two methods only uses f_1 to learn the motion between adjacent frames at training time.

4.4 Implementation Details

The model is implemented with Pytorch. Both networks f_1 and f_2 are trained simultaneously in an end-to-end manner. The AdamW optimizer is used with both a learning rate and weight decay of $1e^{-4}$. The number of features is doubled in each strided convolution of the encoder and halved in each up-convolution of the decoder. The maximum number of features at the bottleneck is 512. Our model is trained with a batch size of 1 and group normalization layers are used throughout the network. The number of frames during training is set to N=12. A 16G NVIDIA V100 GPU is used for training. Data augmentation includes: flipping rotation, zooming, translation, contrast adjustment, Gaussian noise injection, blurring, sharpening and intensity scaling. All models are trained for 100 epochs with 250 iterations per epoch.

5 Results and Discussion

5.1 Comparison with Baseline methods

Table 1 and 2 compare similarity, segmentation and regularization performances of the proposed approach and baseline methods presented in section 4.3. Our model achieved better results for each individual structure and for all segmentation and similarity metrics. The proposed approach and Voxelmorph models achieved better segmentation performance than other methods when the amount of motion between frame was high, as indicated by registration results for the ES frame. For the regularization measure, its performance is comparable to Voxelmorph without the diffeomorphic post-processing. In our experiments, Voxelmorph with the NCC loss achieves better segmentation and similarity results than other competing models. VM-Dif, which also uses the MSE loss function, achieved 0% of pixels with negative Jacobian determinant thanks to the scaling and squaring method applied to compute the exponential map of the predicted flow (refer to [6] for additional information). Thus, this latter model presents the best results in terms of regularity without deteriorating segmentation performance. Similarly, SyN also generates diffeomorphic deformation fields resulting in 0% of pixels with negative Jacobian determinant. Iterative warping methods did not reach other model's segmentation performance. However, they achieved satisfying similarity scores with ES and mean SSIM above VM-MSE and VM-dif. It should be noted that IterWarpFlow performed poorly when it came to flow regularization results, achieving the worse performance in terms of mean number of pixel with negative Jacobian determinant. This can likely be explained by the required interpolation step which takes place each time a deformation field is warped. On the contrary, since IterWarpImg only computed deformation fields between neighboring frames, flow smoothness metric was much better.

Figure 2 shows an example of a registered image using the optical flow computed for each baseline method as well as for our method. The end-systole image is registered to the end-diastole image. It can be seen that the IterWarpImg method produces a blurred image due to the successive application of the warping function, which performs bilinear interpolation at each step. The IterWarpFlow approach generates blocky patterns. The Syn model does not preserve the size of cardiac structures, with a thicker myocardium and irregularly shaped RV compared to the fixed image.

In Figure 3c is shown the cumulative frequency plot of VoxelMorph and the proposed method in terms of Dice score. As can be seen in this graph, the number of end-diastole volumes with Dice score below 94 amounts to around 5% for our model, compared to around 20% for VM-NCC, the best performing baseline method. All in all, this graph demonstrates that our approach contributes to a decrease in the number of failed predictions in terms of registered segmentations. Figure 3a and 3b presents the average Dice score and average SSIM against the distance to the ED frame as a percent of the cardiac cycle. Since not all videos have the same number of frames, results were interpolated over the maximum number of frames and averaged across sequences. These graphs demonstrate that,

Table 1. Segmentation and regularization results: comparison with baseline methods (± standard deviation). The segmentation label of each frame of a sequence is warped towards the segmentation label. P-values are computed between our method and VM-NCC

		ES frame	All frames			
	Method	Mean	Mean	RV	MYO	LV
Dice	VM-MSE	93.24 ± 0.02	94.79 ± 0.02	96.41 ± 0.02	91.51 ± 0.04	96.44 ± 0.02
	VM-NCC	92.88 ± 0.02	95.04 ± 0.02	96.32 ± 0.02	92.12 ± 0.03	96.67 ± 0.02
	VM-Dif	93.18 ± 0.02	94.73 ± 0.02	96.43 ± 0.02	91.37 ± 0.04	96.38 ± 0.02
	SyN	87.60 ± 0.06	93.59 ± 0.04	95.66 ± 0.03	90.12 ± 0.06	94.97 ± 0.05
	IterWarpImg	87.72 ± 0.06	93.18 ± 0.04	96.16 ± 0.02	89.05 ± 0.06	94.32 ± 0.04
	IterWarpFlow	88.15 ± 0.03	93.00 ± 0.04	95.99 ± 0.02	88.73 ± 0.06	94.28 ± 0.04
	Ours	$\mathbf{95.26 \pm 0.01}$	$\mathbf{95.82 \pm 0.01}$	$\mathbf{97.06 \pm 0.01}$	$\mathbf{93.14 \pm 0.02}$	$\mathbf{97.26 \pm 0.01}$
	p-value	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
ASSD (mm)	VM-MSE	0.13 ± 0.05	0.09 ± 0.06	0.07 ± 0.06	0.13 ± 0.10	0.08 ± 0.06
	VM-NCC	0.16 ± 0.08	0.09 ± 0.07	0.08 ± 0.08	0.12 ± 0.09	0.07 ± 0.08
	VM-Dif	0.13 ± 0.05	0.09 ± 0.06	0.07 ± 0.06	0.14 ± 0.11	0.08 ± 0.06
	SyN	0.41 ± 0.31	0.16 ± 0.20	0.12 ± 0.16	0.19 ± 0.21	0.16 ± 0.29
	IterWarpImg	0.34 ± 0.18	0.15 ± 0.13	0.08 ± 0.06	0.20 ± 0.16	0.17 ± 0.21
	IterWarpFlow	0.32 ± 0.15	0.15 ± 0.13	0.08 ± 0.06	0.20 ± 0.15	0.17 ± 0.20
	Ours	$\mathbf{0.08 \pm 0.03}$	$\mathbf{0.07 \pm 0.04}$	$\mathbf{0.05 \pm 0.03}$	$\mathbf{0.10 \pm 0.07}$	$\mathbf{0.05 \pm 0.04}$
	p-value	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
HD (mm)	VM-MSE	3.98 ± 1.00	3.09 ± 1.33	3.50 ± 1.84	3.01 ± 1.45	2.75 ± 1.22
	VM-NCC	4.78 ± 1.52	3.23 ± 1.50	3.84 ± 2.25	3.04 ± 1.47	2.80 ± 1.44
	VM-Dif	3.92 ± 1.02	3.04 ± 1.27	3.40 ± 1.73	2.97 ± 1.40	2.75 ± 1.19
	SyN	6.91 ± 2.51	3.83 ± 2.38	4.18 ± 2.60	3.78 ± 2.56	3.54 ± 2.50
	IterWarpImg	6.54 ± 1.47	3.92 ± 1.89	3.81 ± 1.86	4.06 ± 2.17	3.90 ± 2.17
	IterWarpFlow	6.46 ± 1.46	4.03 ± 1.86	3.84 ± 1.80	4.24 ± 2.20	4.03 ± 2.15
	Ours	$\mathbf{2.97 \pm 0.71}$	$\mathbf{2.59 \pm 0.83}$	$\mathbf{2.88 \pm 1.24}$	$\mathbf{2.57 \pm 0.98}$	$\mathbf{2.32 \pm 0.90}$
	p-value	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
$\%det(J_F) \leq 0$	VM-MSE	0.59 ± 1.98	0.17 ± 1.02	0.14 ± 0.96	0.06 ± 0.43	0.30 ± 2.32
	VM-NCC	0.40 ± 0.54	0.15 ± 0.34	0.11 ± 0.33	0.05 ± 0.20	0.29 ± 0.85
	VM-Dif	$\mathbf{0.0 \pm 0.0}$	$\mathbf{0.0 \pm 0.0}$	$\mathbf{0.0 \pm 0.0}$	$\mathbf{0.0 \pm 0.0}$	$\mathbf{0.0 \pm 0.0}$
	SyN	$\mathbf{0.0 \pm 0.0}$	$\mathbf{0.0 \pm 0.0}$	$\mathbf{0.0 \pm 0.0}$	$\mathbf{0.0 \pm 0.0}$	$\mathbf{0.0 \pm 0.0}$
	IterWarpImg	•	0.003 ± 0.02	0.004 ± 0.03	0.003 ± 0.02	0.003 ± 0.02
	IterWarpFlow	1.09 ± 1.19	0.44 ± 0.77	0.40 ± 0.88	0.22 ± 0.57	0.70 ± 1.21
	Ours	0.65 ± 0.91	0.16 ± 0.44	0.11 ± 0.43	0.07 ± 0.29	0.29 ± 0.95
	p-value	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001

Table 2. SSIM results (± standard deviation). Each frame of a sequence is warped towards the ED frame. P-values are computed between our method and VM-NCC.

Method	Mean SSIM	ES SSIM
VM-MSE	82.96 ± 0.05	76.80 ± 0.07
VM-NCC	88.79 ± 0.04	82.75 ± 0.06
VM-Dif	82.38 ± 0.05	76.13 ± 0.07
SyN	82.70 ± 0.05	75.22 ± 0.08
IterWarpImg	84.60 ± 0.05	77.75 ± 0.08
IterWarpFlow	86.43 ± 0.04	79.40 ± 0.07
Ours	$\mathbf{90.41 \pm 0.03}$	$\mathbf{86.00 \pm 0.05}$
p-value	< 0.0001	< 0.0001

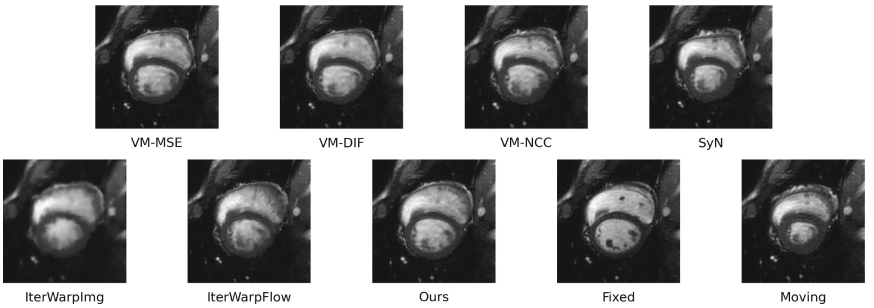
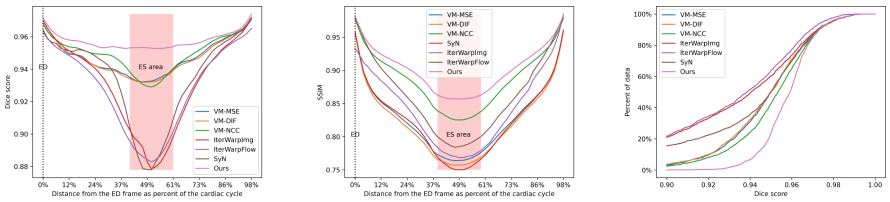


Fig. 2. Image registration example. The predicted optical flow is used to warp the ES (moving) image towards the ED (fixed) image. Best viewed zoomed-in.

as the amount of motion relative to the end-diastole frame increases (the frame that exhibits the maximum amount of motion relative to the ED frame is the ES frame), the performance gap between the proposed approach and other models widens, suggesting that our model better maintains the quality of estimated flows. This less significant drop in performance likely results from the use of intermediate frames between the moving and the fixed frame, which reduces the amount of motion present in each estimated flow.



(a) (b) (c)

Fig. 3. Average Dice score (a) and SSIM (b) of baseline methods and ours against the distance relative to the end-diastole frame as percent of the cardiac cycle. (c) cumulative frequency graph of Dice score for baseline methods and ours. Best viewed zoomed-in.

5.2 Limitations

Having two separate network f_1 and f_2 results in high GPU memory consumption which limits the number of frames used for training. Moreover, the first network f_1 is only used to compute motion between adjacent frames. As such it is a classical optical flow algorithm which could be replaced by a pre-trained network to free up GPU memory. This would also simplify the algorithm since it would reduce the number of losses. Instead of using a pretrained network, the size of f_1 could also be reduced by removing its decoder so that the fusion of flows is performed at lower resolution between the output of f_1 and f_2's encoders.

6 Conclusion

This work introduced a new method to perform image registration between distant frames of a video. An iterative motion aggregation process was presented to estimate motion flow between the first frame of the video sequence and all other frames. To perform this task, a first network was designed to predict motion between adjacent frames whereas a second network was in charge of fusing flows over the video sequence. The approach outperforms Voxelmorph and our pair-of-frame variant especially when the motion between two frames is large.

Acknowledgements. This work was funded by the grant number 965286 from the H2020 MAESTRIA project.

This work was performed using HPC resources from GENCI-IDRIS (Grant 20XX-AD011013634R1)

References

1. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: VoxelMorph: A Learning Framework for Deformable Medical Image Registration. IEEE Transactions on Medical Imaging **38**(8), 1788–1800 (Aug 2019). https://doi.org/10.1109/TMI.2019.2897538, arXiv:1809.05231 [cs]
2. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Gonzalez Ballester, M.A., Sanroma, G., Napel, S., Petersen, S., Tziritas, G., Grinias, E., Khened, M., Kollerathu, V.A., Krishnamurthi, G., Rohe, M.M., Pennec, X., Sermesant, M., Isensee, F., Jager, P., Maier-Hein, K.H., Full, P.M., Wolf, I., Engelhardt, S., Baumgartner, C.F., Koch, L.M., Wolterink, J.M., Isgum, I., Jang, Y., Hong, Y., Patravali, J., Jain, S., Humbert, O., Jodoin, P.M.: Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved? IEEE Transactions on Medical Imaging **37**(11), 2514–2525 (Nov 2018). https://doi.org/10.1109/TMI.2018.2837502, https://ieeexplore.ieee.org/document/8360453/
3. Butler, D., Wulff, J., Stanley, G., Black, M.: A Naturalistic Open Source Movie for Optical Flow Evaluation (Oct 2012). https://doi.org/10.1007/978-3-642-33783-3_44, pages: 625
4. Campello, V.M., Gkontra, P., Izquierdo, C., Martín-Isla, C., Sojoudi, A., Full, P.M., Maier-Hein, K., Zhang, Y., He, Z., Ma, J., Parreño, M., Albiol, A., Kong, F., Shadden, S.C., Acero, J.C., Sundaresan, V., Saber, M., Elattar, M., Li, H., Menze, B., Khader, F., Haarburger, C., Scannell, C.M., Veta, M., Carscadden, A., Punithakumar, K., Liu, X., Tsaftaris, S.A., Huang, X., Yang, X., Li, L., Zhuang, X., Viladés, D., Descalzo, M.L., Guala, A., Mura, L.L., Friedrich, M.G., Garg, R., Lebel, J., Henriques, F., Karakas, M., Çavuş, E., Petersen, S.E., Escalera, S., Seguí, S., Rodríguez-Palomares, J.F., Lekadir, K.: Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Segmentation: The M&Ms Challenge. IEEE Transactions on Medical Imaging **40**(12), 3543–3554 (Dec 2021). https://doi.org/10.1109/TMI.2021.3090082, conference Name: IEEE Transactions on Medical Imaging
5. Cao, X., Yang, J., Zhang, J., Nie, D., Kim, M.j., Wang, Q., Shen, D.: Deformable Image Registration based on Similarity-Steered CNN Regression. Medical image computing and computer-assisted intervention : MICCAI ... Interna-

tional Conference on Medical Image Computing and Computer-Assisted Intervention **10433**, 300–308 (Sep 2017). https://doi.org/10.1007/978-3-319-66182-7_35, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5731783/
6. Dalca, A.V., Balakrishnan, G., Guttag, J., Sabuncu, M.R.: Unsupervised Learning of Probabilistic Diffeomorphic Registration for Images and Surfaces. Medical Image Analysis **57**, 226–236 (Oct 2019). https://doi.org/10.1016/j.media.2019.07.006, arXiv:1903.03545 [cs]
7. Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Smagt, P.V.D., Cremers, D., Brox, T.: FlowNet: Learning Optical Flow with Convolutional Networks. In: 2015 IEEE International Conference on Computer Vision (ICCV). pp. 2758–2766. IEEE, Santiago (Dec 2015). https://doi.org/10.1109/ICCV.2015.316, https://ieeexplore.ieee.org/document/7410673/
8. Eppenhof, K., Pluim, J.: Supervised local error estimation for nonlinear image registration using convolutional neural networks (2017). https://doi.org/10.1117/12.2253859
9. Ferdian, E., Suinesiaputra, A., Fung, K., Aung, N., Lukaschuk, E., Barutcu, A., Maclean, E., Paiva, J., Piechnik, S.K., Neubauer, S., Petersen, S.E., Young, A.A.: Fully Automated Myocardial Strain Estimation from Cardiovascular MRI–tagged Images Using a Deep Learning Framework in the UK Biobank. Radiology: Cardiothoracic Imaging **2**(1), e190032 (Feb 2020). https://doi.org/10.1148/ryct.2020190032, http://pubs.rsna.org/doi/10.1148/ryct.2020190032
10. Gautama, T., Van Hulle, M.: A phase-based approach to the estimation of the optical flow field using spatial filtering. IEEE Transactions on Neural Networks **13**(5), 1127–1136 (Sep 2002). https://doi.org/10.1109/TNN.2002.1031944, https://ieeexplore.ieee.org/abstract/document/1031944, conference Name: IEEE Transactions on Neural Networks
11. Gehrig, M., Millhäusler, M., Gehrig, D., Scaramuzza, D.: E-RAFT: Dense Optical Flow from Event Cameras (Oct 2021), arXiv:2108.10552 [cs]
12. Hermann, S., Werner, R.: High Accuracy Optical Flow for 3D Medical Image Registration Using the Census Cost Function, vol. 8333 (Jan 2014). https://doi.org/10.1007/978-3-642-53842-1_3
13. Hu, Y., Modat, M., Gibson, E., Ghavami, N., Bonmati, E., Moore, C.M., Emberton, M., Noble, J.A., Barratt, D.C., Vercauteren, T.: Label-driven weakly-supervised learning for multimodal deformable image registration. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). pp. 1070–1074 (Apr 2018). https://doi.org/10.1109/ISBI.2018.8363756, https://ieeexplore.ieee.org/abstract/document/8363756, iSSN: 1945-8452
14. Hu, Y., Modat, M., Gibson, E., Li, W., Ghavami, N., Bonmati, E., Wang, G., Bandula, S., Moore, C.M., Emberton, M., Ourselin, S., Noble, J.A., Barratt, D.C., Vercauteren, T.: Weakly-supervised convolutional neural networks for multimodal image registration. Medical Image Analysis **49**, 1–13 (Oct 2018). https://doi.org/10.1016/j.media.2018.07.002, https://www.sciencedirect.com/science/article/pii/S1361841518301051
15. Huang, Z., Shi, X., Zhang, C., Wang, Q., Cheung, K.C., Qin, H., Dai, J., Li, H.: FlowFormer: A Transformer Architecture for Optical Flow (Sep 2022), arXiv:2203.16194 [cs] version: 4
16. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks (Dec 2016), arXiv:1612.01925 [cs]

17. Immanuel, S., Bala, D., George, A.: A Study on Block Matching Algorithms for Motion Estimation. International Journal on Computer Science and Engineering **3** (Jan 2011)
18. Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P.F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S., Maier-Hein, K.H.: nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation (Sep 2018), arXiv:1809.10486 [cs]
19. Jaderberg, M., Simonyan, K., Zisserman, A., kavukcuoglu, k.: Spatial Transformer Networks. In: Advances in Neural Information Processing Systems. vol. 28. Curran Associates, Inc. (2015), https://proceedings.neurips.cc/paper_files/paper/2015/hash/33ceb07bf4eeb3da587e268d663aba1a-Abstract.html
20. Jiang, S., Campbell, D., Lu, Y., Li, H., Hartley, R.: Learning to Estimate Hidden Motions with Global Motion Aggregation. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9752–9761. IEEE, Montreal, QC, Canada (Oct 2021). https://doi.org/10.1109/ICCV48922.2021.00963, https://ieeexplore.ieee.org/document/9711346/
21. Lamy, J., Soulat, G., Evin, M., Huber, A., de Cesare, A., Giron, A., Diebold, B., Redheuil, A., Mousseaux, E., Kachenoura, N.: Scan-rescan reproducibility of ventricular and atrial MRI feature tracking strain. Computers in Biology and Medicine **92**, 197–203 (Jan 2018). https://doi.org/10.1016/j.compbiomed.2017.11.015, https://www.sciencedirect.com/science/article/pii/S0010482517303852
22. Li, H., Fan, Y.: NON-RIGID IMAGE REGISTRATION USING SELF-SUPERVISED FULLY CONVOLUTIONAL NETWORKS WITHOUT TRAINING DATA. Proceedings. IEEE International Symposium on Biomedical Imaging **2018**, 1075–1078 (Apr 2018). https://doi.org/10.1109/ISBI.2018.8363757, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6070305/
23. Liu, S., Yang, B., Wang, Y., Tian, J., Yin, L., Zheng, W.: 2D/3D Multimode Medical Image Registration Based on Normalized Cross-Correlation. Applied Sciences **12**(6), 2828 (Jan 2022). https://doi.org/10.3390/app12062828, https://www.mdpi.com/2076-3417/12/6/2828, number: 6 Publisher: Multidisciplinary Digital Publishing Institute
24. Lucas, B.D., Kanade, T.: An Iterative Image Registration Technique with an Application to Stereo Vision. In: IJCAI'81: 7th international joint conference on Artificial intelligence. vol. 2, pp. 674–679. Vancouver, Canada (Aug 1981), https://hal.science/hal-03697340
25. Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4040–4048. IEEE, Las Vegas, NV, USA (Jun 2016). https://doi.org/10.1109/CVPR.2016.438, http://ieeexplore.ieee.org/document/7780807/
26. Meng, Q., Bai, W., O'Regan, D.P., Rueckert, a.D.: DeepMesh: Mesh-based Cardiac Motion Tracking using Deep Learning (Sep 2023), arXiv:2309.14306 [cs, eess]
27. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3061–3070. IEEE, Boston, MA, USA (Jun 2015). https://doi.org/10.1109/CVPR.2015.7298925, http://ieeexplore.ieee.org/document/7298925/
28. Morales, M.A., van den Boomen, M., Nguyen, C., Kalpathy-Cramer, J., Rosen, B.R., Stultz, C.M., Izquierdo-Garcia, D., Catana, C.: DeepStrain: A Deep Learning Workflow for the Automated Characterization of Cardiac Mechanics. Frontiers in

Cardiovascular Medicine **8** (2021), https://www.frontiersin.org/articles/10.3389/fcvm.2021.730316
29. Pauwels, K., Van Hulle, M.M.: Realtime phase-based optical flow on the GPU. In: 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. pp. 1–8. IEEE, Anchorage, AK, USA (Jun 2008). https://doi.org/10.1109/CVPRW.2008.4563090, http://ieeexplore.ieee.org/document/4563090/
30. Pock, T., Urschler, M., Zach, C., Beichel, R., Bischof, H.: A Duality Based Algorithm for TV-L1-Optical-Flow Image Registration, vol. LNCS 4792 (Nov 2007). https://doi.org/10.1007/978-3-540-75759-7_62, journal Abbreviation: Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention Pages: 8 Publication Title: Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention
31. Qin, C., Wang, S., Chen, C., Qiu, H., Bai, W., Rueckert, D.: Biomechanics-informed Neural Networks for Myocardial Motion Tracking in MRI (Jul 2020), arXiv:2006.04725 [cs, eess]
32. Qiu, H., Qin, C., Schuh, A., Hammernik, K., Rueckert, D.: Learning Diffeomorphic and Modality-invariant Registration using B-splines. In: Proceedings of the Fourth Conference on Medical Imaging with Deep Learning. pp. 645–664. PMLR (Aug 2021), https://proceedings.mlr.press/v143/qiu21a.html, iSSN: 2640-3498
33. Ranjan, A., Black, M.J.: Optical Flow Estimation Using a Spatial Pyramid Network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2720–2729. IEEE, Honolulu, HI (Jul 2017). https://doi.org/10.1109/CVPR.2017.291, http://ieeexplore.ieee.org/document/8099774/
34. Rohé, M.M., Datar, M., Heimann, T., Sermesant, M., Pennec, X.: SVF-Net: Learning Deformable Image Registration Using Shape Matching. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2017, vol. 10433, pp. 266–274. Springer International Publishing, Cham (2017). https://doi.org/10.1007/978-3-319-66182-7_31, https://link.springer.com/10.1007/978-3-319-66182-7_31, series Title: Lecture Notes in Computer Science
35. Shi, X., Huang, Z., Li, D., Zhang, M., Cheung, K.C., See, S., Qin, H., Dai, J., Li, H.: FlowFormer++: Masked Cost Volume Autoencoding for Pretraining Optical Flow Estimation. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1599–1610. IEEE, Vancouver, BC, Canada (Jun 2023). https://doi.org/10.1109/CVPR52729.2023.00160, https://ieeexplore.ieee.org/document/10204766/
36. Sun, D., Yang, X., Liu, M.Y., Kautz, J.: PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume (Jun 2018), arXiv:1709.02371 [cs]
37. Teed, Z., Deng, J.: RAFT: Recurrent All-Pairs Field Transforms for Optical Flow (Aug 2020), arXiv:2003.12039 [cs]
38. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is All you Need. In: Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
39. de Vos, B.D., Berendsen, F.F., Viergever, M.A., Staring, M., Išgum, I.: End-to-End Unsupervised Deformable Image Registration with a Convolutional Neural Network. In: Cardoso, M.J., Arbel, T., Carneiro, G., Syeda-Mahmood, T., Tavares,

J.M.R., Moradi, M., Bradley, A., Greenspan, H., Papa, J.P., Madabhushi, A., Nascimento, J.C., Cardoso, J.S., Belagiannis, V., Lu, Z. (eds.) Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. pp. 204–212. Lecture Notes in Computer Science, Springer International Publishing, Cham (2017). https://doi.org/10.1007/978-3-319-67558-9_24

40. Wang, J., Zhang, M.: DeepFLASH: An Efficient Network for Learning-Based Medical Image Registration. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4443–4451. IEEE, Seattle, WA, USA (Jun 2020). https://doi.org/10.1109/CVPR42600.2020.00450, https://ieeexplore.ieee.org/document/9157045/
41. Wu, G., Liu, X., Luo, K., Liu, X., Zheng, Q., Liu, S., Jiang, X., Zhai, G., Wang, W.: AccFlow: Backward Accumulation for Long-Range Optical Flow. pp. 12119–12128 (2023), https://openaccess.thecvf.com/content/ICCV2023/html/Wu_AccFlow_Backward_Accumulation_for_Long-Range_Optical_Flow_ICCV_2023_paper.html
42. Ye, M., Kanski, M., Yang, D., Chang, Q., Yan, Z., Huang, Q., Axel, L., Metaxas, D.: DeepTag: An Unsupervised Deep Learning Method for Motion Tracking on Cardiac Tagging Magnetic Resonance Images. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7257–7267. IEEE, Nashville, TN, USA (Jun 2021). https://doi.org/10.1109/CVPR46437.2021.00718, https://ieeexplore.ieee.org/document/9578105/
43. Yu, H., Chen, X., Shi, H., Chen, T., Huang, T.S., Sun, S.: Motion Pyramid Networks for Accurate and Efficient Cardiac Motion Estimation (Sep 2020), arXiv:2006.15710 [cs, eess]
44. Zhang, X., You, C., Ahn, S., Zhuang, J., Staib, L., Duncan, J.: Learning correspondences of cardiac motion from images using biomechanics-informed modeling (Sep 2022), arXiv:2209.00726 [cs, eess]
45. Zhu, S., Ma, K.K.: A new diamond search algorithm for fast block-matching motion estimation. IEEE Transactions on Image Processing **9**(2), 287–290 (Feb 2000). https://doi.org/10.1109/83.821744, https://ieeexplore.ieee.org/abstract/document/821744, conference Name: IEEE Transactions on Image Processing

EEG Data Augmentation Using Generative Adversarial Network for Improved Emotion Recognition

Raktim Acharjee(✉) and Shaik Rafi Ahamed

Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, Guwahati, Assam, India
{a.raktim,rafiahamed}@iitg.ac.in

Abstract. Emotion recognition from Electroencephalogram (EEG) signals has emerged as a promising method for understanding human affective states. However, Deep learning-based emotion recognition models suffer from overfitting and generalisation due to the variability in EEG signals and the scarcity of labelled data, which impede their performance. In this work, a Wasserstein Generative Adversarial Network with Gradient Penalty (WGAN-GP) based architecture was adopted for efficient EEG data augmentation. The publicly available "EEG Brainwave" dataset was used to train the WGAN-GP model to synthetically generate the fake EEG data. The generated synthetic data was mixed with the real data in different proportions to determine the optimum ratio of data augmentation for efficient emotion classification. The efficacy of the data augmentation was evaluated by proposing an LSTM-based classifier that efficiently classifies the three emotional states: positive, neutral, and negative. The experimental results show that the maximum classification accuracy of 99.14% was achieved with a precision of 0.9915, recall of 0.9914, and F1 score of 0.9914 when an equal quantity of real and synthetically generated EEG data was mixed to train the classifier. Our WGAN-GP-LSTM method not only enhances the robustness of emotion recognition models by utilizing data augmentation but also significantly improves the classification accuracy with limited labelled data and outperforms all other state-of-the-art techniques.

Keywords: Electroencephalogram (EEG) · Emotion Recognition · Data Augmentation · Generative Adversarial Network (GAN) · Deep Learning.

1 Introduction

Electroencephalography (EEG) is a non-invasive neuroimaging technique that measures electrical activity in the brain, offering insights into neurological activities. EEG offers a window into cognitive processes and emotional states by recording the electrical signals produced by the brain using electrodes applied

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15311, pp. 238–252, 2025.
https://doi.org/10.1007/978-3-031-78195-7_16

to the scalp [1]. Emotion is a complex psychological and physiological state involving subjective feelings, physiological arousal, expressive behaviours, and cognitive interpretations [9]. Emotion recognition plays a pivotal role in interpersonal communication and understanding. It holds immense potential across various domains, from improving human-computer interaction to enhancing mental health diagnostics and even aiding marketing strategies [8]. Emotion recognition from EEG signals has emerged as a promising method for understanding human affective states. This method typically involves collecting EEG data while individuals experience various emotional stimuli to elicit specific emotions [1,16]. EEG-based emotion recognition provides a non-invasive, cost-effective, and portable method with high temporal resolution for capturing neural activities of emotional states.

Deep learning techniques have shown remarkable performance in extracting complex patterns from EEG data for emotion classification. However, the efficacy of such models relies significantly on the availability and quality of training data. Emotion recognition models suffer from overfitting and generalisation due to the variability in EEG signals, and the scarcity of labelled data, which impede their performance [21]. Data augmentation has emerged as a crucial technique to address these challenges by artificially expanding the manifoldness and quantity of the training data [17]. This method prevents overfitting and improves the model's ability to identify patterns across various conditions and subjects. The traditional augmentation methods include time shifting, frequency warping, random noise addition, and data resampling. These techniques aim to increase the diversity and dataset size while preserving its underlying characteristics. Szczakowska et al. [22] used sliding windows of varying lengths, overlapping windows, and Gaussian noise for augmentation and found enhancement in the accuracies. However, the improvements were not very significant.

Generative Adversarial Networks (GANs) have become an effective tool in data augmentation which can generate new data instances with feature variations, enhancing the robustness and generalization of machine learning models [11]. GANs are performing well in creating artificial EEG signals that mimic real brain activity patterns. This technique aims to overcome dataset size and diversity limitations by generating additional labelled data for training deep learning models [25]. However, generating realistic synthetic EEG signals poses several issues, including capturing individual variability, noise characteristics, and the complexity of emotional dynamics [18]. To overcome these challenges, Fahimi et al. [10] developed a framework based on conditional Deep Convolutional Generative Adversarial Networks (DCGANs) to generate artificial EEG data from the recorded EEG signals. They incorporated Convolutional Neural Network (CNN) for both generator and discriminator networks. Their findings demonstrate that the classification tasks benefited greatly from the artificially generated samples added to the training set. To improve the stability of the model and to generate more diverse samples, Bao et al. [4] proposed a variational auto-encoder (VAE) based GAN architecture (VAE-D2GAN) that was trained on the Differential Entropy (DE) features of EEG data. They transformed the DE features

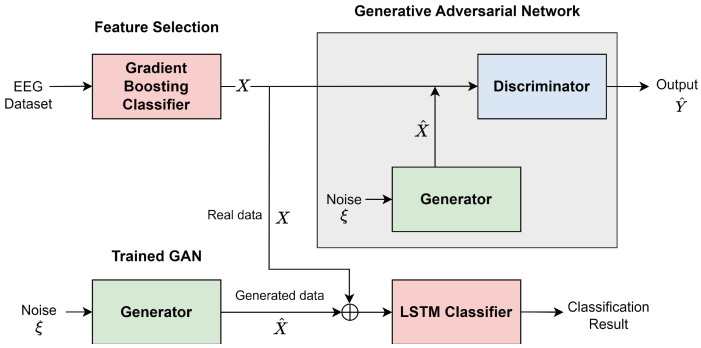


Fig. 1. Flowchart of the proposed method.

into topological images by mapping and interpolation, which helped the model learn relevant spatial distribution. Their method improved the emotion classification accuracy by 1.5% and 3.5% on the SEED and SEED-IV datasets, respectively. A similar approach has been taken by Tian et al. [24] in their work, where they have proposed a dual encoder variational-autoencoder-based GAN (DEVAE-GAN) architecture to take into account the EEG signals' spatiotemporal characteristics to generate high-quality synthetic data. They have reported a 5% improvement in the classification accuracy after using the augmented dataset along with the original dataset. Zhang et al. [27] have taken a different approach where they proposed a GAN-based self-supervised data augmentation method (GANSER). They introduced a multifactor training network (MTN) for masking transformation operation that forces the generator to synthesize a wide variety of samples, which ultimately improves the classification performance.

Conventional GANs often encounter challenges when applied to time-series data due to their inherent instability, mode collapse, and difficulty in training, which can lead to the generation of unrealistic samples and poor diversity in the synthesized data. Wasserstein GAN is a variant of GAN that uses the Wasserstein loss function to help with the mode collapse and vanishing gradient problems. Gradient Penalty (GP), on the other hand, is a regularization technique that enforces the 1-Lipschitz continuity, which ensures the stability of the WGAN with Wasserstein loss. In this work, these techniques were incorporated and adopted, and a Wasserstein Generative Adversarial Network with Gradient Penalty (WGAN-GP) based architecture was proposed for efficient EEG data augmentation to improve emotion recognition by generating realistic EEG data distribution. Publicly available "EEG Brainwave" dataset was used for this study [5,7]. The flowchart of the proposed technique is shown in Fig. 1. First, a Gradient Boosting Classifier was used for feature selection and extracted the fifteen most relevant features in contrast to efficient emotion recognition. Next, our proposed WGAN-GP model was trained on this feature-selected dataset, in which the discriminator learns to discriminate between produced and actual samples while the generator attempts to make data samples that are indistinguishable

from real data. Once the model had been trained, the discriminator was removed, and the generator started generating fake synthetic EEG data. These generated data were mixed with the real dataset into different proportions to create the augmented EEG dataset. To evaluate the efficacy of the data augmentation, an LSTM-based multi-class classifier was also proposed to efficiently classify the three emotional states: positive, neutral, and negative. Results show that the WGAN-GP-LSTM method not only enhances the robustness of emotion recognition models by utilizing data augmentation but also significantly improves the classification accuracy with the limited labelled dataset.

The remaining sections of the paper are arranged as follows: in the Methodology section, the "EEG Brainwave" dataset is first briefly discussed, followed by the feature selection method using the Gradient Boosting classifier. Then, a brief discussion about the traditional GANs and their variants is presented, followed by a comprehensive overview of our proposed WGAN-GP architecture. A detailed overview of the proposed LSTM-based classifier is also presented. The detailed training procedure of the proposed models is then discussed, followed by the performance evaluation and results in the Model Training and Evaluation section. The results are then compared with the state-of-the-art techniques. Finally, the results are analyzed and discussed, and the work is concluded in the Conclusion section.

2 Methodology

2.1 EEG Emotion Dataset

In this work, publicly available "EEG Brainwave Dataset: Feeling Emotions", created by J. J. Bird et al. [5,7] was used. This dataset contains EEG recordings of two subjects: one male and one female, under various emotion elicitation stimuli. The authors used a popular and commercially available MUSE EEG headband to record the data, which had four EEG electrodes, namely TP9, AF7, AF8, and TP10, and NZ was used as a reference point. The location of the electrodes in the Muse EEG headband is shown in Fig. 2 (a). A total of six emotional video clips (three positive and three negative) were used as a stimulus to elicit emotions in the two subjects. Also, neutral state brain activity was recorded without applying any external stimulus. For a total of 36 minutes, the authors collected EEG data at a sample frequency of 250 Hz (12 minutes for the classes of positive, neutral, and negative emotions). Five frequency bands are included in every recorded EEG: delta (1-4 Hz), theta (4-8 Hz), alpha (8-13 Hz), beta (13-30 Hz), and gamma (30-80 Hz). The recorded data was filtered to remove unwanted noises and artifacts. The authors used a sliding window of 1-second starting at t=0 with a window size of 0.5 seconds to extract various features to generate a statistical feature dataset after segmenting the originally recorded data into smaller data segments. Several statistical features such as mean, standard deviation, skewness, kurtosis, autocorrelations, maximum and minimum values, derivatives, log-covariance, Shannon entropy, log-energy entropy, and power spectral density were extracted at each time window for each of the five

frequency bands. The dataset's final shape became 2132×2547, where 2547 represents the total number of features that were extracted and 2132 represents the number of data samples.

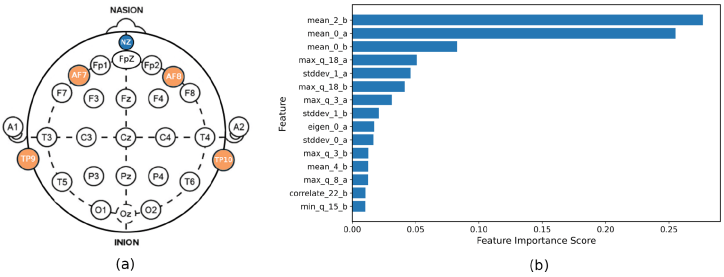


Fig. 2. (a) Location of the electrodes in the Muse EEG headband [7]. (b) Selected features with their feature importance score by the Gradient Boosting Classifier.

2.2 Feature Selection

Feature selection in deep learning is a method used to improve model performance by focusing on the most relevant features. In the EEG Brainwave dataset, there are a total of 2547 extracted features. To reduce the dimensionality and extract the most relevant features, the Gradient Boosting Classifier has been used for efficient feature selection. The Gradient Boosting Classifier is a robust machine learning technique that sequentially constructs an ensemble of simple models, usually decision trees, to form a highly accurate predictive model. This can be effectively utilized for feature selection to identify and prioritize important features. The gradient boosting classifier was chosen due to its robust performance and ability to handle the high dimensionality and complexity of EEG data. Several studies in the literature [2,23] have also reported that gradient boosting classifiers perform well for EEG feature selection compared to other feature selection methods. The optimization function of gradient boosting classifier is given as [26]:

$$\min_{\beta} \ell(\beta) + \lambda|\beta|_1 + \mu \sum_{f=1}^{d} q_\epsilon \left(\sum_{t=1}^{T} |F_{ft}\beta_t| \right) \tag{1}$$

where, β is a sparse linear vector that selects trees. The first penalty term of Eq. 1 is the plain l_1 norm that reduces overfitting, and the second penalty term is the capped l_1 norm that explicitly penalizes the extraction of features. The impurity function is given as [26]:

$$h_t = \underset{h_t \in \mathcal{H}}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{n} \Big(g_i - h_t(\mathbf{x}_i) \Big)^2 + \mu \sum_{f=1}^{d} \phi_f F_{ft} \tag{2}$$

where, the first term encourages feature splits to best fit the loss function's negative gradient, and the second term promotes splitting on features that have previously been used in prior iterations. The greedy CART algorithm was used to minimize the impurity function in Eq. 2 to learn regression trees h_t.

The dataset was normalized to a mean of 0 and a standard deviation of 1 before feeding into the gradient boosting classifier. Through preliminary experiments, it was observed that selecting the top 15 features yielded a significant improvement in both EEG data augmentation and emotion recognition accuracy while keeping the model complexities manageable. Additionally, different numbers of features (e.g., 10, 20, and 25) were experimented with, and it was found that the optimal results are obtained when the top 15 features are chosen. Therefore, the 15 most significant features were selected using the Gradient Boosting classifier on our dataset. Selected features with their feature importance score are shown in Fig. 2 (b). Following the feature selection process, the dataset now comprises 2132×15 dimensions, with 2132 representing the data samples and 15 denoting the chosen features.

2.3 Proposed Model

A Generative Adversarial Network (GAN) is a class of Deep Learning algorithms introduced by Goodfellow et al. [12] in 2014. GANs consist of two components: the generator and the discriminator, engaged in a competitive dynamic of creation and discrimination [13]. The generator receives random noise as input and endeavors to produce data that closely resembles real data. Conversely, the discriminator discerns between real data and data synthesized by the generator. Loss function of GAN is given below, where d and g play the two-player minimax game [12]:

$$\min_{d} \max_{g} -[\mathbb{E}(\log(d(\boldsymbol{x}))) + \mathbb{E}(1 - \log(d(g(\boldsymbol{z}))))] \tag{3}$$

where, $d(x)$ denote the likelihood that x originated from the real data, and g is a differentiable function that represents the generator network. Although the binary cross-entropy (BCE) loss function works well for conventional GANs, it can suffer from vanishing gradient and mode collapses. To overcome these issues, Arjovsky et al. [3] proposed the Wasserstein loss function, alternatively termed the Wasserstein distance or Earth Mover's Distance (EMD), as a solution. Wasserstein loss helps with the mode collapse and vanishing gradient problems, expressed as [3]:

$$\min_{g} \max_{c} \mathbb{E}(c(\boldsymbol{x})) - \mathbb{E}(c(g(\boldsymbol{z}))) \tag{4}$$

where, c is the critic, equivalent to the discriminator d in the conventional GAN, which has to be 1-Lipschitz continuous. To ensure the 1-Lipschitz continuity in the Wasserstein GAN, the Gradient Penalty (GP) is utilized to impose a penalty on the gradient norm of the critic's output concerning its input data. The loss function of the Wasserstein GAN with GP is as follows [14]:

$$\min_{g} \max_{c} \mathbb{E}(c(\boldsymbol{x})) - \mathbb{E}(c(g(\boldsymbol{z}))) + \lambda \mathbb{E}(\|\nabla c(\hat{\boldsymbol{x}})\|_2 - 1)^2 \tag{5}$$

where, the second term is the gradient penalty. The incorporation of gradient penalty in Wasserstein GAN represents a significant advancement in GAN training methodology, contributing to the improved robustness and effectiveness of WGANs in generating realistic data distributions.

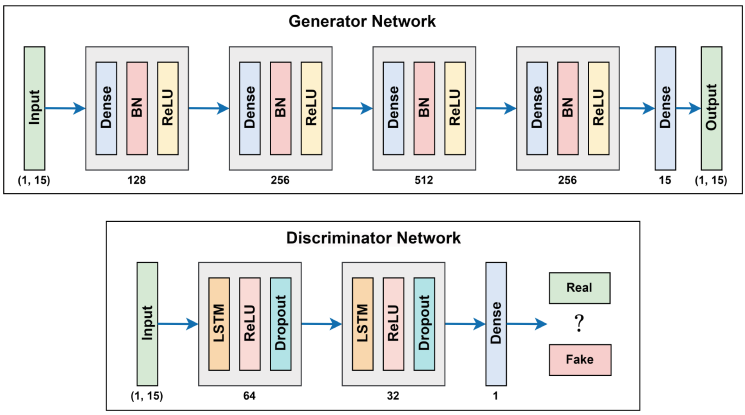


Fig. 3. The architecture of the proposed Wasserstein Generative Adversarial Network with Gradient Penalty (WGAN-GP) for EEG data augmentation. The generator and the discriminator networks are shown separately.

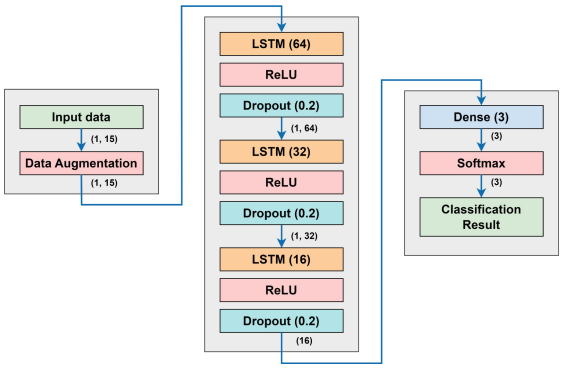


Fig. 4. The architecture of the proposed LSTM-based multiclass classifier network for efficient Emotion Recognition from EEG signals.

In this work, a Wasserstein Generative Adversarial Network with Gradient Penalty (WGAN-GP) based architecture has been proposed for efficient augmentation of EEG signals. Fig. 3 shows the design of both the Generator and

Discriminator networks of the proposed WGAN-GP model. The generator network comprises four blocks of dense neural networks, with each followed by a batch normalization layer and activated using ReLU activation functions. Beginning with a random noise vector as input, each dense layer applies a linear transformation, progressively refining the representation to resemble the target data distribution. After each dense layer, batch normalization is applied, which helps stabilize and accelerate the training process by normalizing the activations within each mini-batch, and also makes the optimization process more robust and efficient. ReLU activation functions are utilized following each batch normalization layer, introducing non-linearity to the network and enabling it to acquire more intricate features. The final dense layer of the generator network employs a linear activation function that learns to generate data closely resembling the distribution of the training data. The shape of the input noise vector and the generator's output is $(1, 15)$, similar to the shape of the training dataset.

The Discriminator network consists of two Long Short-Term Memory (LSTM) layers. LSTM is a type of recurrent neural network (RNN) designed to capture long-range dependencies in sequential data. There are 64 units in the first LSTM layer and 32 units in the second. Each LSTM unit maintains a cell state and hidden state, allowing the model to memorize information over multiple time steps. The expressions are as follows [15]:

$$c^{<t>} = \Gamma_u * \tilde{c}^{<t>} + \Gamma_f * c^{<t-1>} \tag{6}$$

$$a^{<t>} = \Gamma_o * \tanh(c^{<t>}) \tag{7}$$

where the input, forget, and update gates are denoted by Γ_u, Γ_f, and Γ_o, respectively and $\tilde{c}^{<t>}$, $c^{<t>}$ and $a^{<t>}$ are the candidate value, cell state and the hidden state, respectively. These LSTM layers process the input sequence iteratively, updating their internal states and capturing relevant temporal patterns in the data. After every LSTM layer, a ReLU activation function is implemented to introduce non-linearity, succeeded by a Dropout layer aimed at mitigating overfitting. The final output from the LSTM layer is fed into a dense layer, which is subsequently activated by a sigmoid function. During the iterative training process, the generator strives to generate data that closely resemble real data, while the discriminator learns to distinguish between real and generated samples.

To determine the effectiveness of the data augmentation and efficient emotion recognition, an LSTM-based multiclass classifier network is proposed to categorize EEG patterns into one of three emotion classes: positive, neutral, or negative. The architecture of this classifier is similar to the discriminator network of the WGAN-GP, shown in Fig. 4. The network consists of three blocks of Long Short-Term Memory (LSTM) units designed to capture temporal relationships and patterns within the EEG data. Following each LSTM block, a ReLU activation function is applied that introduces non-linearity into the network, and a dropout layer is applied with a rate of 0.2 after each ReLU activation to prevent overfitting. The output from the final LSTM block is fed into a dense layer containing three units, succeeded by a Softmax layer to normalize the output scores into probabilities. The emotion class with the highest probability is subsequently

predicted as the output. During the training process, the network learns to associate patterns in EEG data with their corresponding emotion classes, thereby achieving precise emotion recognition.

Table 1. The training parameters/hyperparameters for our proposed models.

Parameters/hyperparameters	**Value**		
	Generator	**Discriminator**	**Classifier**
Loss function	Wasserstein loss	BCE	CCE
Optimizer	Adam	RMSprop	Adam
Learning rate	0.001	0.001	0.001
Dropout	0.3	0.3	0.2
Batch size	32	32	32
Epochs	100	100	250
Trainable parameters	313,856	32,929	36,083

3 Model Training and Evaluation

Our proposed WGAN-GP and LSTM models were implemented in Keras using the TensorFlow 2.0 framework and trained on an NVIDIA Tesla V100-PCIE 32 GB GPU. Initially, the WGAN-GP model was trained on the training dataset. The Wasserstein loss function was applied to the generator network and Binary cross-entropy (BCE) was applied to the discriminator network during training. The optimization function employed for the generator was Adam optimizer, whereas Root Mean Squared Propagation (RMSprop) was used for the discriminator network. The training parameters and hyperparameters for our models are given in Table 1. Following the feature selection process, there were total 2132 data samples in the original dataset. This dataset was split into 80% for training and 20% for testing purposes. The training set was utilized to train the WGAN-GP network for 25 epochs with a batch size 32. In each training iteration, the discriminator is trained multiple times before updating the generator. During each iteration, the discriminator is provided with both real and synthetic EEG signal data. The loss is calculated by comparing the scores assigned to real and fake samples. Additionally, the gradient penalty term is computed to ensure smoothness in the discriminator's decision boundary. The gradients are then computed and used to update the discriminator's weights. After training the discriminator, the generator is trained once per iteration. The generator receives random noise as input and generates fake EEG signal data. The generated data is subsequently fed into the discriminator, and the loss is calculated according to the discriminator's assessment of the generated samples. The generator's weights are updated using the gradients computed from this loss.

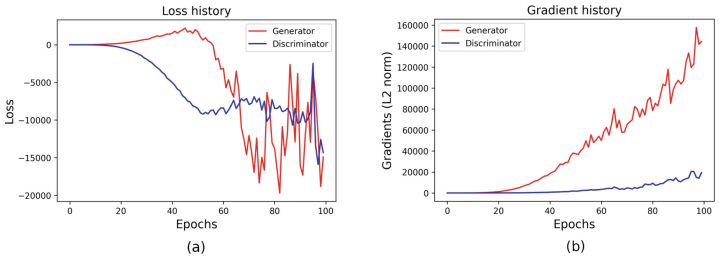


Fig. 5. Comparison of the (a) training loss curves and (b) training gradients (L2 norm) curves for the generator and discriminator.

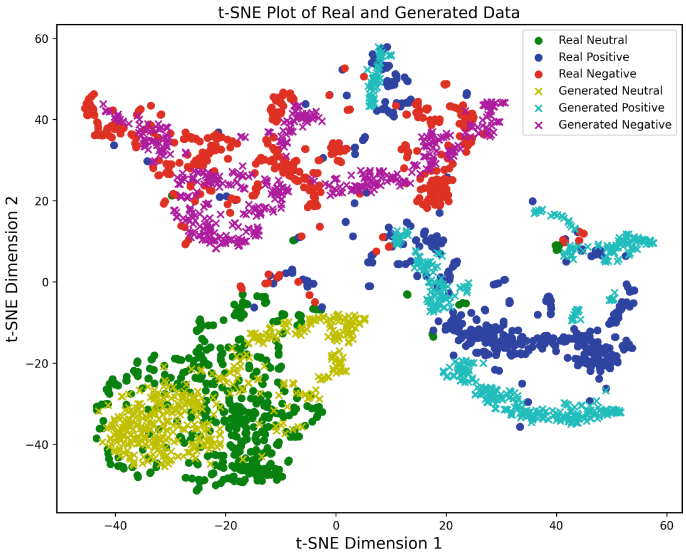


Fig. 6. Scatter plot of the real and generated EEG dataset (1×Dataset) using two components found by t-SNE.

During the training process, metrics like loss and gradient norms are monitored and recorded, offering insights into the training dynamics and the model's convergence. Fig. 5 shows the training loss curves and gradients (L2 norm) curves comparison between the generator and the discriminator. From the loss curves, it's observable that the training loss decreased swiftly for approximately sixty epochs for both the generator and the discriminator. Subsequently, the generator's loss began oscillating and exhibiting fluctuations beyond this point. Training GANs on EEG data is challenging due to the high complexity and variability of EEG signals, making it difficult to learn the underlying patterns. This complexity contributes to the observed instability in the loss curves. The primary metric for evaluating GAN performance is the quality of the generated samples. To test the effectiveness of our trained WGAN-GP model, fake EEG

data for all three classes was generated and plotted along with the real dataset. A scatter plot of the real and generated EEG dataset using two components found by t-SNE is shown in Fig. 6. Data from each emotion class are clustered together in the latent space, and it is evident that the generated data closely resemble the distribution of the corresponding real data. This implies that the generated data is suitable for data augmentation, as it encompasses adequate realistic information.

Table 2. Performance analysis for different generated sample sizes.

Generated Data	Accuracy	Precision	Recall	F1 Score
0×Dataset	0.9736	0.9753	0.9736	0.9734
0.5×Dataset	0.9847	0.9848	0.9847	0.9847
1×Dataset	**0.9914**	**0.9915**	**0.9914**	**0.9914**
2×Dataset	0.9906	0.9907	0.9906	0.9906
3×Dataset	0.9882	0.9883	0.9882	0.9882
4×Dataset	0.9867	0.9868	0.9866	0.9866

Table 3. Performance analysis of our proposed method.

Author	Method	Accuracy
Jordan et al. [6]	AdaBoosted LSTM	97.06%
Jordan et al. [5]	InfoGain, RandomForest	97.89%
Rahman et al. [20]	Light Gradient Boosting Machine	97.94%
Mridha et al. [19]	Deep Neural Network	98.44%
Our method	**WGAN-GP-LSTM**	**99.14%**

To assess the effectiveness of the data augmentation, an LSTM classifier was trained with the generated data. Categorical cross-entropy was employed as the loss function, and the Adam optimizer was utilized for optimization. Other training parameters and hyperparameters are given in Table 1. The dataset was divided into training and validation subsets, and the model was trained for 250 epochs with a batch size of 32, employing five-fold cross-validation. Each fold of cross-validation allocated 80% of the data for training and 20% for validation. The generated synthetic data from the trained WGAN-GP generator was mixed with the real EEG data in different proportions to determine the optimum ratio of data augmentation for efficient emotion classification. The results are given in Table 2. The results were assessed based on accuracy, precision, recall, and F1 score, and were evaluated using the original 20% testing dataset. The number of generated data is denoted by the times of the real dataset, where 0 represents

the only real dataset without any augmentation (baseline). Table 2 illustrates that adding an equal quantity of generated data with real data yields optimal performance, with a classification accuracy of 99.14%, precision of 0.9915, recall of 0.9914, and F1 score of 0.9914. The real dataset without any augmentation has the lowest classification accuracy of 97.36%. The proposed data augmentation method has led to a enhancement of 1.78% in classification accuracy. As the ratio of generated data increased in the training dataset, the classification accuracy started to improve. It gets maximum when the real and generated datasets are equal in the training data. Further increasing the number of generated data leads to a degradation in classification accuracy but still obtains higher accuracy than the original datasets. Fig. 7 shows the plot of the effect of data augmentation on classification accuracy for different generated sample sizes. The confusion matrices trained by the real dataset and the augmented dataset (1×Dataset) are shown in Fig. 8. The true label in each category is displayed as rows in confusion matrices, while the predicted labels are displayed as columns. When an equal quantity of generated data was added with the original training set, the classification accuracy for the three emotional states: positive, neutral, and negative was enhanced by 0.26%, 4.94%, and 0.17%, respectively. These results demonstrate that adding generated data to the training set enhances the performance of the classifier.

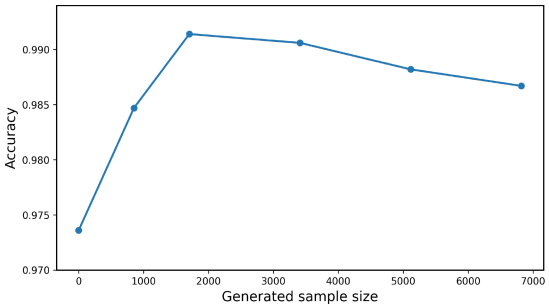


Fig. 7. The effect of data augmentation on classification accuracy for different generated sample sizes.

The findings on emotion classification accuracy were compared against several state-of-the-art methods to assess the effectiveness of our WGAN-GP-LSTM approach: (i) "A Deep Evolutionary Approach to Bioinspired Classifier Optimisation for Brain-Machine Interaction" proposed in [6]. (ii) "Mental Emotional Sentiment Classification with an EEG-based Brain-Machine Interface" introduced in [5]. (iii) "An Efficient Analysis of EEG Signals to Perform Emotion Analysis" suggested in [20]. (iv) "Emotion Recognition: A New Tool for Healthcare Using Deep Learning Algorithms" proposed in [19]. Table 3 presents the classification accuracies for the state-of-the-art techniques and our WGAN-GP-LSTM method. It can be seen that WGAN-GP-LSTM method archives a higher

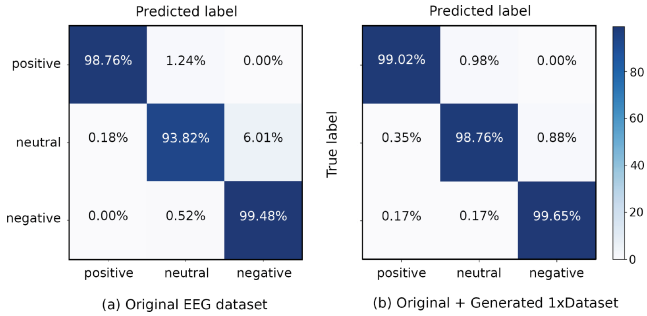


Fig. 8. The confusion matrices for emotion classification trained by (a) original EEG dataset (b) original and generated data added together (1×Dataset).

classification accuracy of 99.14% in comparison and outperforms all other state-of-the-art methods. Moreover, our method surpasses the approach suggested by Mridha et al. [19] with an improvement of 0.7% in the classification accuracy.

4 Conclusion

This paper proposes a Generative Adversarial Network (GAN) based EEG data augmentation method for improved emotion recognition. Deep learning-based emotion recognition models suffer from overfitting and generalisation due to the variability in EEG signals and the scarcity of labelled data that limits the overall performance. In this work, a Wasserstein Generative Adversarial Network with Gradient Penalty (WGAN-GP) based architecture was proposed for efficient EEG data augmentation. The publicly available "EEG Brainwave Dataset" was used for this study. First, a Gradient Boosted Classifier was used for feature selection to extract the fifteen most significant features from the dataset. This dataset was then used to train our proposed WGAN-GP model. To determine the effectiveness of the data augmentation and for efficient emotion recognition, an LSTM-based multiclass classifier was also proposed. Fake synthetic EEG data was generated from our trained WGAN-GP generator and mixed with the real EEG data in different proportions to determine the optimum ratio of data augmentation for efficient emotion classification. Results show that the generated data closely resembles the distribution of the corresponding real data, and mixing an equal number of generated data with the real data achieves the best performance in terms of classification accuracy of 99.14%. A comparison of our results was made with the state-of-the-art techniques, and our approach clearly outperformed all other methods in terms of classification accuracy. Our WGAN-GP-LSTM method was able to efficiently generate superior and realistic EEG data for augmentation that enhances the robustness of emotion recognition models and significantly improves the emotion recognition accuracy.

References

1. Alarcao, S.M., Fonseca, M.J.: Emotions Recognition using EEG Signals: A Survey. IEEE Trans. Affect. Comput. **10**(3), 374–393 (2017)
2. Alzahab, N.A., Baldi, M., Scalise, L.: Efficient feature selection for electroencephalogram-based authentication. In: 2021 IEEE International Symposium on Medical Measurements and Applications (MeMeA). pp. 1–6. IEEE (2021)
3. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein Generative Adversarial Networks. In: International conference on machine learning. pp. 214–223. PMLR (2017)
4. Bao, G., Yan, B., Tong, L., Shu, J., Wang, L., Yang, K., Zeng, Y.: Data augmentation for EEG-based emotion recognition using generative adversarial networks. Front. Comput. Neurosci. **15**, 723843 (2021)
5. Bird, J.J., Ekart, A., Buckingham, C.D., Faria, D.R.: Mental Emotional Sentiment Classification with an EEG-based Brain-machine Interface. In: Proceedings of theInternational Conference on Digital Image and Signal Processing (DISP'19) (2019)
6. Bird, J.J., Faria, D.R., Manso, L.J., Ekárt, A., Buckingham, C.D., et al.: A Deep Evolutionary Approach to Bioinspired Classifier Optimisation for Brain-Machine Interaction. Complexity (2019)
7. Bird, J.J., Manso, L.J., Ribeiro, E.P., Ekárt, A., Faria, D.R.: A Study on Mental State Classification using EEG-based Brain-Machine Interface. In: 2018 International Conference on Intelligent Systems (IS). pp. 795–800 (2018). https://doi.org/10.1109/IS.2018.8710576
8. Dzedzickis, A., Kaklauskas, A., Bucinskas, V.: Human emotion recognition: Review of sensors and methods. Sensors **20**(3), 592 (2020)
9. Ekman, P.: An argument for basic emotions. Cognition & emotion **6**(3–4), 169–200 (1992)
10. Fahimi, F., Dosen, S., Ang, K.K., Mrachacz-Kersting, N., Guan, C.: Generative adversarial networks-based data augmentation for brain-computer interface. IEEE transactions on neural networks and learning systems **32**(9), 4039–4051 (2020)
11. Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., Greenspan, H.: Synthetic data augmentation using GAN for improved liver lesion classification. In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018). pp. 289–293. IEEE (2018)
12. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative Adversarial Nets. Advances in neural information processing systems **27** (2014)
13. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative Adversarial Networks. Commun. ACM **63**(11), 139–144 (2020)
14. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved Training of Wasserstein GANs. Advances in neural information processing systems **30** (2017)
15. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
16. Jenke, R., Peer, A., Buss, M.: Feature extraction and selection for emotion recognition from EEG. IEEE Trans. Affect. Comput. **5**(3), 327–339 (2014)

17. Lashgari, E., Liang, D., Maoz, U.: Data augmentation for deep-learning-based electroencephalography. J. Neurosci. Methods **346**, 108885 (2020)
18. Luo, Y., Zhu, L.Z., Wan, Z.Y., Lu, B.L.: Data augmentation for enhancing EEG-based emotion recognition with deep generative models. J. Neural Eng. **17**(5), 056021 (2020)
19. Mridha, K., Sarker, T., Zaman, R., Shukla, M., Ghosh, A., Shaw, R.N.: Emotion Recognition: A New Tool for Healthcare Using Deep Learning Algorithms. In: International Conference on Electrical and Electronics Engineering. pp. 613–631. Springer (2023)
20. Rahman, A.A., Kabir, M.R., Ratul, R.H., Shamns, F.A., Nishat, M.M., Faisal, F.: An Efficient Analysis of EEG Signals to Perform Emotion Analysis. In: 2023 4th International Conference on Artificial Intelligence, Robotics and Control (AIRC). pp. 1–7 (2023). https://doi.org/10.1109/AIRC57904.2023.10303179
21. Suhaimi, N.S., Mountstephens, J., Teo, J.: EEG-based emotion recognition: A state-of-the-art review of current trends and opportunities. Computational intelligence and neuroscience **2020** (2020)
22. Szczakowska, P., Wosiak, A., et al.: Improving Automatic Recognition of Emotional States Using EEG Data Augmentation Techniques. Procedia Computer Science **225**, 4225–4234 (2023)
23. Thenmozhi, T., Helen, R.: Feature selection using extreme gradient boosting Bayesian optimization to upgrade the classification performance of motor imagery signals for BCI. J. Neurosci. Methods **366**, 109425 (2022)
24. Tian, C., Ma, Y., Cammon, J., Fang, F., Zhang, Y., Meng, M.: Dual-encoder vae-gan with spatiotemporal features for emotional eeg data augmentation. IEEE Transactions on Neural Systems and Rehabilitation Engineering (2023)
25. Wang, F., Zhong, S.h., Peng, J., Jiang, J., Liu, Y.: Data augmentation for EEG-based emotion recognition with deep convolutional neural networks. In: MultiMedia Modeling: 24th International Conference, MMM 2018, Bangkok, Thailand, February 5-7, 2018, Proceedings, Part II 24. pp. 82–93. Springer (2018)
26. Xu, Z., Huang, G., Weinberger, K.Q., Zheng, A.X.: Gradient Boosted Feature Selection. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 522–531 (2014)
27. Zhang, Z., Liu, Y., Zhong, S.h.: Ganser: A self-supervised data augmentation framework for eeg-based emotion recognition. IEEE Transactions on Affective Computing **14**(3), 2048–2063 (2023). https://doi.org/10.1109/TAFFC.2022.3170369

Automated Maternal Fetal Ultrasound Image Identification Using a Hybrid Vision Transformer Model

Thunakala Bala Krishna[1], Ajay Kumar Reddy Poreddy[1], Kolla Gnapika Sindhu[2], and Priyanka Kokil[1(✉)]

[1] Advanced Signal and Image Processing (ASIP) Lab, Department of Electronics and Communication Engineering, Indian Institute of Information Technology, Design and Manufacturing, Kancheepuram, Chennai 600127, India
{ec21d0004,edm20d012,priyanka}@iiitdm.ac.in

[2] Amrita School of Computing, Department of Computer Science and Engineering, Amrita Vishwa Vidyapeetham, Chennai, Tamil Nadu, India

Abstract. Ultrasound (US) technology has revolutionized prenatal care by offering noninvasive, real-time visualization of maternal-fetal anatomy. The accurate classification of maternal-fetal US planes is a critical segment of effective prenatal diagnosis. However, the inherent inter-class variance among different fetal US images presents a significant hurdle, making fetal anatomy detection a laborious and time-consuming task, even for experienced sonographers. This paper proposes a novel approach using a Hybrid Vision Transformer (H-ViT) for automated fetal anatomical plane classification to address these challenges. The proposed method utilizes hierarchical features extracted from DenseNet-121, which are then inputted into the vision transformer to analyze complex spatial relationships and patterns within fetal US images. By incorporating both global and local features, the proposed method enhances feature discriminability, thus alleviating low inter-class variance. The effectiveness of the H-ViT is evaluated using the largest publicly available maternal-fetal US image dataset. The experimental results rigorously demonstrate the superiority of our approach, achieving an accuracy of 96.60% compared to other state-of-the-art techniques.

Keywords: Fetal ultrasound classification · Vision transformer · Convolutional neural network · Deep learning · Maternal–fetal planes

1 Introduction

Ultrasound (US) imaging has revolutionized the field of prenatal care, providing healthcare professionals with a non-invasive and real-time visualization tool to assess the well-being of both the mother and the developing fetus [1]. It allows for examining critical anatomical structures, tracking fetal growth, and identifying potential abnormalities at various stages of pregnancy [2]. Traditionally, the

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15311, pp. 253–269, 2025.
https://doi.org/10.1007/978-3-031-78195-7_17

identification of maternal-fetal US planes has relied on the expertise of highly skilled sonographers, who manually analyze US images to locate and interpret specific anatomical landmarks [3]. However, this process is time-consuming, subject to inter-observer variability, and heavily reliant on the operator's experience and training. As a result, there is a growing need for automated and standardized approaches to enhance the efficiency and accuracy of maternal-fetal US plane detection. From a medical practitioner's point of view, the evaluation of fetal US images can be quite demanding because they often contain distortions like acoustic shadows, speckle noise, motion blur, and unclear boundaries. These distortions arise due to the intricate interplay between US waves and the tissues of both the mother and the fetus [4]. The comprehensive study and accurate classification of US fetal planes, encompassing crucial aspects such as the fetal abdomen, fetal brain, fetal femur, fetal thorax, maternal cervix, and other pertinent planes, hold immense potential to transform healthcare and positively impact society. They constitute a proactive approach towards enhancing prenatal care, estimating fetal weight [5], promoting maternal and fetal health [6], and fostering a higher quality of life for expectant mothers and their offspring [7]. Figure 1 presents the most frequently employed fetal anatomical structures in prenatal screening to estimate fetal well-being. Classifying US fetal planes aids in gestational age prediction by providing critical anatomical markers and measurements, enabling a more accurate estimation of the fetus's developmental stage and age [9].

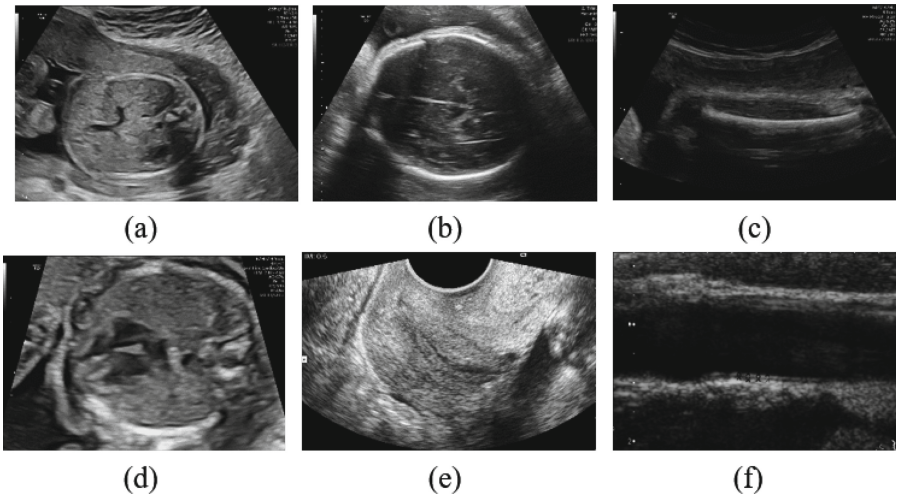


Fig. 1. Illustration of the commonly used fetal US images: (a) fetal abdomen (b) fetal brain (c) fetal femur (d) fetal thorax (e) fetal cervix (f) other [8].

Deep learning techniques, especially convolutional neural networks (CNNs), have demonstrated remarkable success in medical imaging tasks [10]. CNNs excel

at learning complex hierarchical representations from large datasets, enabling them to extract meaningful features and classify images with high accuracy [11]. Their application in US imaging has shown promising results in various areas, such as fetal biometry, organ segmentation, and anomaly detection. Xavier *et al.* [12] evaluated various deep CNNs on fetal US images and concluded that DenseNet-169 has a close correlation with human technicians. Sridar *et al.* [13] computed the local features of US images using pre-trained CNNs and the global attributes using the discriminant regions of the US planes. Further, the final decision is made by fusing the decisions computed from the individual attributes using the support vector machine (SVM) classifier. Rasheed *et al.* [14] computed the frame level predictions of fetal US videos using AlexNet and segmented the fetal head frames using UNET. Further, the segmented frames are utilized to calculate the biparietal diameter (BPD) and head circumference (HC) via segmented contours.

Thunakala *et al.* [15] developed a feature-fused model for fetal US planes using ResNet-50 and the AlexNet models. Further, the fused features are fed to SVM for final prediction. Yu *et al.* [16] developed a CNN model for fetal facial detection using augmentation and fine-tuning techniques. In [17], the authors utilized SonoNet architecture to automatically detect 13 fetal standard views and a bounding box localization mechanism using weak supervision from image-level labels. Yang *et al.* [18] developed a radial component mechanism (RCM) that visualizes the key geometric characteristics of fetal abdominal planes. Further, the critical attributes identified from RCM are given to a random forest (RF) classifier to identify fetal abdominal and non-abdominal classes. In [19], the authors extracted dense and region of interest attributes of fetal US images using Fisher vector, transform descriptor, and the Gaussian mixture model. Further, the computed attributes are given to SVM to identify sagittal, axial, and coronal US planes. In [20], the authors computed the spatial features of fetal US videos using MobileNet architecture and the temporal attributes using recurrent neural networks. Further, these cues are fused using a two-stage mechanism to identify the four classes of fetal US videos. In [21], the authors fused the deep features computed from AlexNet and VGG-19 model and fed to multilayer perceptron to classify the six planes of fetal US images. In [9,22], the authors developed an ensemble network by fine-tuning the top-performing CNNs, and the predictions from stacked CNNs are given to the majority voting classifier for final prediction. In [23], the authors reduced the interclass variance among the fetal US planes using VGG-16 and adaptive weighting mechanism, and then the features were fed to the softmax classifier to identify six classes of US planes. Although existing fetal US image classification methods [9,12–16,18–22] have produced satisfactory results, several limitations remain. The following highlights the limitations of the existing methods:

1. Only certain fetal anatomical structures, such as the fetal brain and maternal cervix, are accurately interpreted. However, other prominent planes like the fetal femur, abdomen, and thorax are often misclassified as non-standard planes.

2. The inherent low interclass variance of fetal US images needs to be addressed more effectively.
3. There is a necessity to develop DL models that focus primarily on significant regions of the fetal anatomy while minimizing attention to artifacts such as background information.

Recently, the introduction of hybrid vision transformer models has further advanced the field of image classification. Vision transformers (ViTs), originally proposed for natural image analysis, leverage the transformer architecture's ability to capture long-range dependencies and self-attention mechanisms. By incorporating both convolutional and transformer-based layers, hybrid Vision transformers (H-ViTs) combine the strengths of both approaches, potentially improving the performance of image classification tasks. In this paper, we have proposed to combine the effectiveness of CNN architectures and ViTs to automate the detection of common maternal-fetal US planes. We hypothesize that combining CNNs and ViTs will leverage the benefits of spatial feature extraction from CNNs and the attention-based capabilities of transformers, leading to enhanced accuracy and addressing the inherent low interclass variance of fetal US images. To the best of our knowledge, none of the works in the existing literature [9,12–24] proposed a hybrid model based on deep CNN and transformer to classify fetal US planes. This has motivated us to develop an H-ViT model to reduce inter-subject variability among the critical anatomical structures of fetal US images. Our study comprehensively evaluates these models on a diverse dataset, considering various factors such as computational efficiency, interpretability, and generalization performance. By automating the detection of common maternal fetal ultrasound planes, we envision a future where healthcare professionals can benefit from standardized and efficient analysis, enabling more accurate diagnoses, improved patient care, and ultimately, better outcomes for both the mother and the fetus. The following are the contributions of the proposed method:

1. A deep learning model called H-ViT, which improves the classification of commonly used fetal anatomical structures by enhancing the interclass variance among fetal planes is proposed.
2. The H-ViT model combines DenseNet-121 and Vision Transformer (ViT). Spatial feature maps extracted from the DenseNet-121 backbone are fed into the ViT, which further refines the fetal US attributes using the attention-based capabilities of transformers.
3. To ensure the robustness and reliability of the proposed model, we further assessed its performance using a speckle-introduced fetal US image dataset, notably under noisy conditions.

The remainder of the paper is structured as follows: Section 2 delve into the dataset details and introduce the H-ViT architecture proposed for classifying fetal US images. Section 3 reports the experimental results and validation studies, and Section 4 concludes the paper by summarizing the findings and discussing potential future research directions.

2 Methodology

This section describes the fetal US dataset, the preprocessing steps applied to the fetal US images, the feature extraction process, and an elucidation of the proposed H-ViT for fetal US image classification.

Table 1. Distribution of dataset over different classes

Fetal Plane	No of Samples
Fetal Abdomen	711
Fetal Femur	1040
Fetal Brain	3092
Fetal Thorax	1718
Maternal Cervix	1626
Other	4216
Total	12400

2.1 Dataset Description

A large and diverse dataset comprising routinely acquired maternal-fetal screening US images is utilized to evaluate the effectiveness of the proposed H-ViT [8]. The dataset was collected from two hospitals involving multiple operators and US machines. An expert maternal-fetal clinician has meticulously labeled all the images in the dataset to ensure accuracy. The dataset consists of 12,400 images, categorized into six distinct classes as mentioned in Table 1. Four classes represent the most commonly used fetal anatomical planes: abdomen, brain, femur, and thorax. Additionally, there is a class named the maternal cervix, often utilized for prematurity screening. Finally, a general category encompasses any other less frequently encountered image plane.

2.2 Data augmentation

In order to enhance the robustness and generalization of the proposed classification method, a systematic data augmentation strategy is employed on the maternal-fetal screening US dataset. We have increased the number of samples in each class of the fetal US dataset using data augmentation techniques such as cropping, rotating, translating, and flipping images [25]. Data augmentation enhances dataset diversity and reduces data imbalance, enabling assessing the robustness and generalization of the classification model [26]. This augmentation approach aims to diversify the dataset by generating variations of the original images while preserving their anatomical and clinical characteristics [25]. The original fetal US dataset exhibits variations in the number of samples across

the distinct fetal plane categories, with some classes having considerably fewer samples compared to others, as shown in Table 1. After augmenting the fetal US image dataset, the number of images corresponding to each class has increased to 2000. Additionally, the augmented fetal US dataset is divided into 70% for training, 15% for validation, and 15% for testing the CNN and ViT models.

2.3 Preprocessing

The proposed approach employs appropriate preprocessing techniques on the fetal US image dataset to facilitate training and evaluation. Preprocessing steps include rescaling the pixel values of the US images to a range of 0 to 1, shearing the images to introduce geometric variations, applying zooming transformations, and horizontal flipping. These steps enhance the ability of the classification model to generalize and enhance its performance. Furthermore, the images are resized to a standardized resolution of 224×224×3 pixels, ensuring consistency across the dataset. Additionally, as part of the preprocessing pipeline, speckle noise is introduced to the US images. Speckle noise, a multiplicative noise commonly encountered in US imaging due to interference patterns, is simulated to evaluate the proposed model strength to this artifact and assess its robustness in real-world scenarios.

2.4 Feature Extraction

The feature extraction process is crucial for capturing informative representations from input images in the proposed methodology. To achieve this, the pre-trained DenseNet-121 [27] CNN is utilized as our backbone feature extraction network. DenseNet-121 has demonstrated its effectiveness in various computer vision tasks due to its dense connectivity and feature reuse characteristics. The DenseNet-121 is pre-trained on a large ImageNet dataset to acquire a generalized understanding of low- to mid-level features. The pre-trained backbone CNN is fine-tuned on the target fetal US image dataset, enabling the extraction of hierarchical features with varying levels of abstraction [28]. During feature extraction, fetal US images are passed through the DenseNet-121 backbone, undergoing a series of convolutional and pooling layers. The dense blocks within the architecture generate feature maps that contain increasingly abstract representations as we move deeper into the network, as shown in Figure 2. These feature maps capture both low-level details and high-level semantic information, making them suitable for a wide range of downstream tasks.

2.5 Proposed H-ViT model

In this section, we introduce a novel and synergistic methodology for precise image classification, leveraging the remarkable capabilities of the ViT architecture. The ViT represents a groundbreaking approach in computer vision and image analysis [29]. Unlike traditional CNNs, ViT relies on self-attention mechanism, allowing it to effectively capture long-range dependencies in images, as

shown in Fig 2. This innovative architecture has shown remarkable performance in various image-related tasks, demonstrating its potential to revolutionize the field of image classification.

Building upon the strengths of ViT, our approach combines it with the computational prowess of the DenseNet-121 backbone [30]. The harmonious fusion of cutting-edge techniques results in an efficacious and proficient model explicitly designed for achieving precise and accurate image classification.

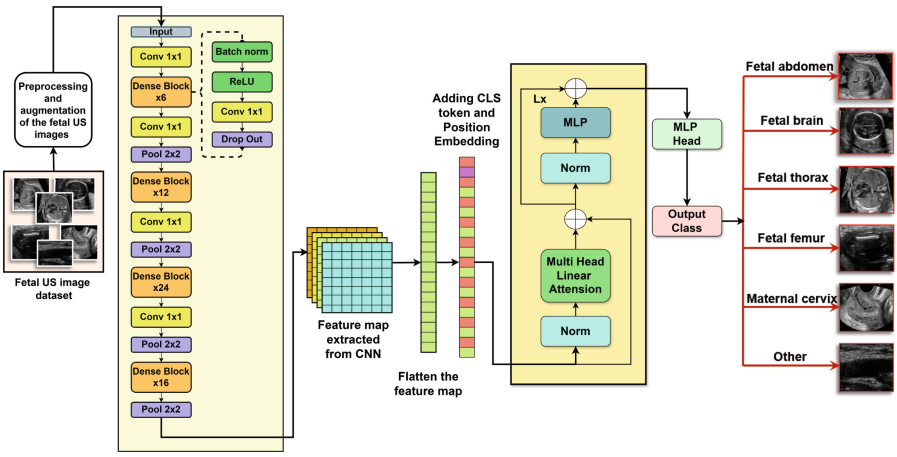


Fig. 2. Proposed model for Fetal Plane classification

Patch-based representation: Upon computing feature extraction by the backbone DenseNet-121 model, we strategically partition the resultant feature maps into discrete patches. The feature map partitioning is the foundation of our patch-based representation strategy. It facilitates a synergistic amalgamation of localized context and holistic perceptual insights derived from fetal US images. The partitioning strategy aims to capture fine-grained local intricacies and global contextual cues, fostering a comprehensive and nuanced understanding of fetal US image information.

Position embedding and class embedding: Position embedding is a strategy utilized in transformer models to encode the position of each image patch. CNNs, through their architecture, intrinsically preserve the spatial structure of the input images, but transformers process input data as a sequence. To address this, position embeddings are added to the input sequence to inject details about the position of each patch into the model. Class embedding is a unique token added to the input sequence in transformer models, particularly in the context of classification tasks. In ViT, a class embedding token (often called the [CLS]

token) is prepended to the sequence of image patches. In the proposed H-ViT model, position embeddings are added to each image patch to provide spatial information, and a class embedding token is prepended to the sequence. These embeddings allow the transformer model to understand the spatial relationships between patches and gather a holistic representation for classification purposes.

Transformer encoder: unveiling complex relationships: Operationally, the encoder initiates by taking the enriched patch-based representations. Through self-attention, the encoder perceptively estimates the significance of each patch in relation to its adjacent spatial values, thereby capturing intricate interdependencies that might elude conventional convolutional constructs. The intrinsic capability empowers the model to discern nuanced spatial hierarchies and complex cross-patch relationships, fostering a granular understanding of image content.

Furthermore, the self-attention process is perpetuated through the multi-head mechanism. The model adeptly encapsulates the fetal US image feature relationships and patterns by executing parallel attention computations, each capturing distinct aspects of inter-patch relationships. Subsequently, the outputs of these multi-faceted computations are harmoniously concatenated and subject to linear transformations, yielding a synthesis that matches diverse attentional perspectives.

The concatenated multi-head attention outputs are then channeled through position-wise feedforward neural networks, infusing the process with non-linearity and intricate processing. The iterative procedure unfolds across multiple encoder layers, thereby fostering successive refinement of patch relationships and representations.

Multi-layer perceptron The final fetal US image classification is estimated using the multi-layer perceptron (MLP) head. The MLP is a robust, interconnected structure comprised of various dense layers aimed at high-level feature refinement. MLP head subjects the amalgamated information to iterative transformations, fostering intricate non-linear abstractions. Combined with non-linearity, the sequential application of operations empowers the model to extract salient high-level semantic facets encoded within the input fetal US data.

2.6 Speckle Noise

Speckle noise is inherent to US imaging and can significantly impact the quality of acquired images [31]. By introducing synthetic speckle-noise at different levels, we simulate real-world US imaging conditions. Including speckle noise in the training dataset gives the model a more realistic representation of imaging scenarios. This helps the model adapt to noise patterns and improves its robustness against noisy inputs during inference. By training the model on a diverse dataset that includes speckle noise, we expect the model to generalize better to unseen real-world data, which typically exhibit speckle artifacts. The Rayleigh

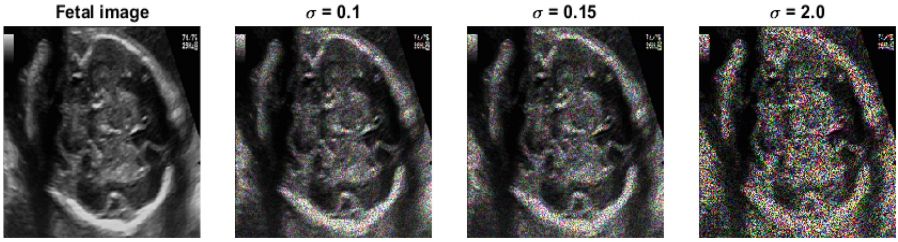


Fig. 3. Speckled noise fetal US images with different levels.

distribution is often appropriate for describing the amplitude of the received US signals, as speckle noise tends to exhibit a Rayleigh distribution when the signals are coherent [32]. The probability density function representing the Rayleigh distribution is given as:

$$\mathbb{P}(z;\sigma) = \frac{z}{\sigma^2}\exp\left(-\frac{z^2}{2\sigma^2}\right), \tag{1}$$

where z represents the row vectored input US image and σ denotes the spread of the speckle noise. In this paper, the generalization of the proposed model is evaluated by considering the different values of σ as mentioned in Figure 3.

3 Experimental results

In this section, we present the empirical results of our proposed methodology, which leverages a combination of DenseNet-121 and ViT for maternal fetal US image classification task. We conducted a comprehensive experimental analysis, including training and evaluation, to assess the effectiveness and robustness of our approach.

3.1 Experimental Setup

To compare the performance of the H-ViT model, we selected four prominent DL architectures: VGG19, Xception, InceptionV3, and DenseNet-121. After training on the fetal dataset, we observed that DenseNet-121 consistently outperformed the other architectures across a range of evaluation metrics. As a result, we selected DenseNet-121 as the backbone for our proposed approach. Further, to enhance the model's generalization capabilities and resilience to noise, we added different levels of speckle noise to the fetal US images. The incorporation of speckle noise not only improved the model's robustness but also contributed to achieving enhanced classification accuracy.

3.2 Evaluation Metrics

We assessed the performance of H-ViT and existing models using a comprehensive set of evaluation metrics, such as accuracy, precision, recall, and F1 score. Accuracy delivers insights into the ratio of correctly predicted instances to the total number of instances in the testing dataset, providing an overall measure of the model's correctness. The proportion of true positive predictions out of the total positive predictions made by the model gives the measure of precision. It measures the model's ability to avoid false positives. Recall indicates the proportion of true positive predictions from the actual positive instances in the testing dataset, which describes the model's capacity to capture positive instances. F1 score is the harmonic mean of precision and recall. It provides a balanced assessment of both metrics and is especially useful when dealing with imbalanced classes.

Table 2. Performance comparison of the proposed method (in %) with the backbone CNN and ViT model

Model	Accuracy	Precision	Recall	F1 score
DenseNet-201	93.00	93.07	93.00	93.01
ViT-base16	89.30	90.81	88.20	89.23
H-ViT	**96.33**	**96.77**	**96.36**	**96.56**

Table 3. Performance comparison of proposed model with state-of-the-art deep CNN models.

Model	Train Acc	Test Acc	Test Loss	Precision	Recall	F1-score	Test Time
VGG19	0.911	0.889	0.331	0.892	0.889	0.889	119s
InceptionV3	0.957	0.931	0.231	0.931	0.931	0.931	215s
Xception	0.969	0.927	0.222	0.927	0.927	0.927	150s
DenseNet-121	0.974	**0.940**	0.208	0.930	0.93	0.930	79s
ViT	0.997	0.893	0.311	0.908	0.882	0.892	50s
Proposed (H-ViT)	0.981	**0.963**	0.188	0.967	0.963	0.965	**37s**

3.3 Performance of the proposed H-ViT model

We have evaluated the performance of the backbone CNN model (DenseNet-121) and vision transformer (ViT) individually as an ablation study, and the corresponding results are tabulated in Table 2. The experimental results show

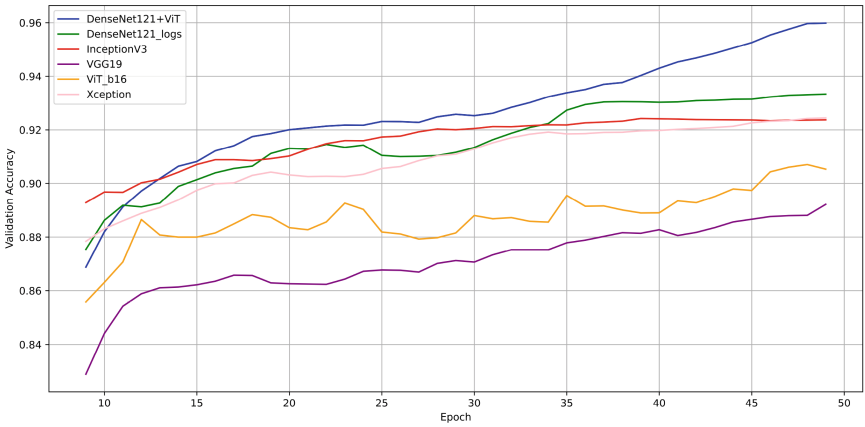


Fig. 4. Accuracy values of the state-of-the-art CNN models for different number of epochs.

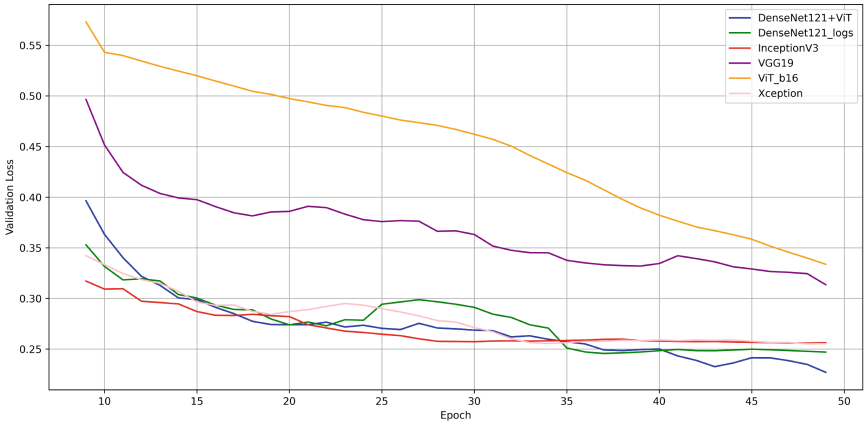


Fig. 5. Validation loss of the state-of-the-art CNN models for different number of epochs

that the CNN model has achieved better than the ViT base model. However, the proposed H-ViT model achieves superior results by leveraging the refined feature maps extracted from the backbone CNN model fed into the ViT. A comprehensive overview of the performance metrics for various existing models, including VGG19, InceptionV3, Xception, DenseNet-121, ViT, and the proposed hybrid model (H-ViT) is provided in Table 3. The metrics include training accuracy, testing accuracy, loss, precision, recall, F1-score, and testing time. From Table 3, one may observe that the proposed H-ViT model achieved remarkable improvements in terms of performance metrics showcasing its superior capabilities in capturing intricate visual patterns of fetal US planes. Figure 4 visualizes the accuracy values for different numbers of epochs on different state-of-the-art CNN

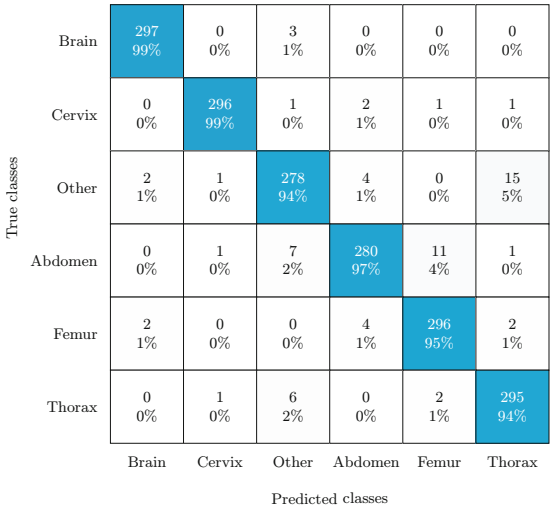


Fig. 6. Confusion matrix of the proposed H-ViT model for different classes of the fetal US dataset.

Table 4. Performance results of the DenseNet and the proposed H-ViT architecture for different levels of speckle noise.

Noise level	Test Accuracy	Test Loss	Precision	Recall	F1-score
DenseNet-121+Noise					
σ=0.1	0.878	0.366	0.887	0.875	0.876
σ=0.15	0.806	0.639	0.848	0.808	0.814
σ=0.2	0.725	0.978	0.814	0.727	0.735
DenseNet-121+ViT+Noise					
σ=0.1	0.928	0.210	0.929	0.928	0.929
σ=0.15	0.898	0.3000	0.900	0.893	0.894
σ=0.2	0.873	0.368	0.881	0.871	0.873

models. From Figure 4, it is clear that H-ViT (DenseNet-121+ViT) achieved the best accuracy values, demarcating its efficacy in identifying the key anatomical attributes of the fetal US planes. Similarly, Figure 5 depicts the validation loss of different CNN models across different numbers of epochs. The plot clearly shows that the H-ViT model's loss cure slowly converges toward zero as the number of epochs increases, indicating its suitability for precision-driven tasks. Figure 6 illustrates the confusion chart of the proposed fetal US image classification model, depicting the class-wise performance in terms of accuracy. Additionally, the confusion plot provides insights into the potential misinterpretations made by the proposed model for each class. It is concluded that the majority of fetal US images are accurately classified, except for the other class, which exhibits

structural similarities with the remaining classes, leading to misclassification. Also, we performed an ablation study by introducing speckle noise to the fetal US planes to check the generalization capability of the proposed H-ViT model. The models trained with speckle noise augmentation exhibited enhanced generalization, effectively mitigating the adverse effects of noise and variability in real-world scenarios. Table 4 illustrates the performance results of the DenseNet-121 model and the proposed H-ViT modeling in a noisy environment. The table depicts performance metrics that are slightly declined by increasing the level of speckle-noise across three different levels of speckle noise on DenseNet-121 architecture. However, embedding ViT into the DenseNet-121 model increases the performance values, and the performance results are slightly increased compared to the DenseNet-121 model. This indicates that the proposed H-ViT model performs consistently across different levels of speckle noise, highlighting its efficacy in the noisy environment.

3.4 Comparison with the existing works

The fetal US image classification performance of the proposed method is further compared with the competing works [12,16,17,21,23,33,34] to show the effectiveness. The quantitative performance measures are presented in Table 5. The results presented in the table show that the H-ViT model achieved superior outcomes than the competing methods.To compare the proposed and existing works fairly, all competing methods are trained and tested on the fetal US dataset used to evaluate the H-ViT model.

Table 5. Performance comparison of the proposed model (in %) with the competing methods.

Method	Accuracy	Precision	Recall	F1 score
Xavier et al. [12]	93.73	91.95	93.08	92.5
Zhen Yu et al [16]	94.12	92.87	94.25	93.48
Baumgartner et al. [17]	94.48	93.56	93.06	93.28
Krishna et al. [21]	95.10	93.83	95.00	94.38
Krishna et al. [23]	95.33	93.58	95.84	94.64
HaifaGhabri et al [33]	93.63	91.50	93.59	92.48
Sendra et al. [34]	94.94	93.17	95.36	94.19
Proposed method	**96.33**	**96.77**	**96.36**	**96.56**

3.5 Ablation study

The effectiveness of the proposed method is also evaluated through an ablation study using different dataset sizes and splits, including five fold cross-validation.

We acknowledge the importance of 5-fold cross-validation in evaluating machine learning models. To address this concern, we have conducted additional experiments using 5-fold cross-validation. The cross-fold validation results are provided in Table 6, and the experimental results provide a more comprehensive evaluation of the proposed model performance. Additionally, we evaluated the performance of the proposed model using different data splits: 60%, 70%, and 80% for training (train), 20%, 15%, and 10% for validation (vali), and 20%, 15%, and 10% for testing (test), respectively. The qualitative results for these experiments are provided in Table 7. From the table, one may observe that the 70:15:15 split yields superior performance due to more balanced class distribution across training, validation, and testing sets, enhancing the model's ability to learn and classify unseen instances accurately. The 15% validation set size provides sufficient evaluation data, improving performance while mitigating overfitting. This balanced approach ensures effective training and evaluation. These results are highlighted with brown color text in the revised manuscript.

Table 6. Five fold Cross-validation results

Model	Accuracy	Precision	Recall	F1 score
Fold 1	96.33	96.77	96.36	96.56
Fold 2	92.66	93.27	92.41	92.84
Fold 3	95.58	95.97	95.41	95.69
Fold 4	95.16	95.24	95.08	95.16
Fold 5	94.66	94.97	94.41	94.69

Table 7. Performance comparison of the proposed method under various dataset splits

Split ratio (train:vali:test)	Accuracy	Precision	Recall	F1 score
60:20:20	85.20	86.79	84.08	85.41
70:15:15	96.33	96.77	96.36	96.56
80:10:10	95.49	95.57	95.49	95.60

4 Conclusion

This article introduces an H-ViT that integrates a DenseNet-121 backbone with transformer architecture to enhance the automatic classification of maternal-fetal US planes during prenatal screening. To assess the effectiveness of the proposed approach, we utilized a publicly available fetal US image dataset obtained from high-resource settings. Analysis of experimental results, conducted with

various noise levels, demonstrates the classification model's generalization capability across diverse fetal US image qualities. The proposed method accurately identifies frequently investigated fetal structures, offering valuable support to sonographers and obstetricians in monitoring fetal development and early detection of complications.

In the future, it would be interesting to incorporate more advanced preprocessing techniques, such as noise reduction and artifact removal, which can further enhance the quality of input images. Consequently, the performance of the model might be improved.

References

1. Wells, P.N.: Ultrasound imaging. Physics in Medicine & Biology **51**(13), R83 (2006)
2. Levine, D.: Ultrasound versus magnetic resonance imaging in fetal evaluation. Top. Magn. Reson. Imaging **12**(1), 25–38 (2001)
3. Huang, Q., Zeng, Z., et al.: A review on real-time 3D ultrasound imaging technology. BioMed Research International 2017 (2017)
4. Meng, L., Zhao, D., Yang, Z., Wang, B.: Automatic display of fetal brain planes and automatic measurements of fetal brain parameters by transabdominal three-dimensional ultrasound. J. Clin. Ultrasound **48**(2), 82–88 (2020)
5. Hadlock, F.P., Harrist, R., Sharman, R.S., Deter, R.L., Park, S.K.: Estimation of fetal weight with the use of head, body, and femur measurements–a prospective study. Am. J. Obstet. Gynecol. **151**(3), 333–337 (1985)
6. Turan, S., Miller, J., Baschat, A.A.: Integrated testing and management in fetal growth restriction. In: Seminars in Perinatology. vol. 32, pp. 194–200. Elsevier (2008)
7. Nicolaides, K.H., Syngelaki, A., Ashoor, G., Birdir, C., Touzet, G.: Noninvasive prenatal testing for fetal trisomies in a routinely screened first-trimester population. Am. J. Obstet. Gynecol. **207**(5), 374-e1 (2012)
8. Burgos-Artizzu, X.P., Coronado-Gutierrez, D., Valenzuela-Alcaraz, B., Bonet-Carne, E., Eixarch, E., Crispi, F., Gratacós, E.: FETAL_PLANES_DB: Common maternal-fetal ultrasound images (Jun 2020), https://doi.org/10.5281/zenodo.3904280
9. Krishna, T.B., Kokil, P.: Standard fetal ultrasound plane classification based on stacked ensemble of deep learning models. Expert Syst. Appl. **238**, 122153 (2024)
10. Fiorentino, M.C., Villani, F.P., Di Cosmo, M., Frontoni, E., Moccia, S.: A review on deep-learning algorithms for fetal ultrasound-image analysis. Med. Image Anal. **83**, 102629 (2023)
11. Liu, S., Wang, Y., Yang, X., Lei, B., Liu, L., Li, S.X., Ni, D., Wang, T.: Deep learning in medical ultrasound analysis: A review. Engineering **5**(2), 261–275 (2019)
12. Burgos-Artizzu, X.P., Coronado-Gutiérrez, D., Valenzuela-Alcaraz, B., Bonet-Carne, E., Eixarch, E., Crispi, F., Gratacós, E.: Evaluation of deep convolutional neural networks for automatic classification of common maternal fetal ultrasound planes. Sci. Rep. **10**(1), 10200 (2020)
13. Sridar, P., Kumar, A., Quinton, A., Nanan, R., Kim, J., Krishnakumar, R.: Decision fusion-based fetal ultrasound image plane classification using convolutional neural networks. Ultrasound in Medicine & Biology **45**(5), 1259–1273 (2019)
14. Rasheed, K., Junejo, F., Malik, A., Saqib, M.: Automated fetal head classification and segmentation using ultrasound video. IEEE Access **9**, 160249–160267 (2021)

15. Krishna, T.B., Kokil, P.: Automated detection of common maternal fetal ultrasound planes using deep feature fusion. In: IEEE 19th India Council International Conference (INDICON). pp. 1–5. Kochi, India (2022)
16. Yu, Z., Tan, E.L., Ni, D., Qin, J., Chen, S., Li, S., Lei, B., Wang, T.: A deep convolutional neural network-based framework for automatic fetal facial standard plane recognition. IEEE J. Biomed. Health Inform. **22**(3), 874–885 (2017)
17. Baumgartner, C.F., Kamnitsas, K., Matthew, J., Fletcher, T.P., Smith, S., Koch, L.M., Kainz, B., Rueckert, D.: SonoNet: Real-time detection and localisation of fetal standard scan planes in freehand ultrasound. IEEE Trans. Med. Imaging **36**(11), 2204–2215 (2017)
18. Yang, X., Ni, D., Qin, J., Li, S., Wang, T., Chen, S., Heng, P.A.: Standard plane localization in ultrasound by radial component. In: 11th International Symposium on Biomedical Imaging (ISBI). pp. 1180–1183. IEEE (2014)
19. Lei, B., Zhuo, L., Chen, S., Li, S., Ni, D., Wang, T.: Automatic recognition of fetal standard plane in ultrasound image. In: 11th International Symposium on Biomedical Imaging (ISBI). pp. 85–88. IEEE (2014)
20. Pu, B., Li, K., Li, S., Zhu, N.: Automatic fetal ultrasound standard plane recognition based on deep learning and IIoT. IEEE Trans. Industr. Inf. **17**(11), 7771–7780 (2021)
21. Krishna, T.B., Kokil, P.: Automated classification of common maternal fetal ultrasound planes using multi-layer perceptron with deep feature integration. Biomed. Signal Process. Control **86**, 105283 (2023)
22. Sindhu, K.G., R, A.: Ensemble-based advancements in maternal fetal plane and brain plane classification for enhanced prenatal diagnosis. International Journal of Information Technology pp. 1–17 (2024)
23. Krishna, T.B., Kokil, P.: Integration of a deep convolutional neural network with adaptive channel weight technique for automated identification of standard fetal biometry planes. IEEE Trans. Instrum. Meas. **73**, 1–11 (2024)
24. Ma'Sum, M.A., Jatmiko, W., Tawakal, M.I., Al Afif, F.: Automatic fetal organs detection and approximation in ultrasound image using boosting classifier and Hough transform. In: International Conference on Advanced Computer Science and Information System. pp. 460–467. IEEE (2014)
25. Yang, S., Xiao, W., Zhang, M., Guo, S., Zhao, J., Shen, F.: Image data augmentation for deep learning: A survey. arXiv preprint arXiv:2204.08610 (2022)
26. Thabtah, F., Hammoud, S., Kamalov, F., Gonsalves, A.: Data imbalance in classification: Experimental evaluation. Inf. Sci. **513**, 429–441 (2020)
27. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4700–4708 (2017)
28. Varshni, D., Thakral, K., Agarwal, L., Nijhawan, R., Mittal, A.: Pneumonia detection using CNN based feature extraction. In: International Conference on Electrical, Computer and Communication Technologies (ICECCT). pp. 1–7. IEEE (2019)
29. Chen, C.F.R., Fan, Q., Panda, R.: Crossvit: Cross-attention multi-scale vision transformer for image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 357–366 (2021)
30. Angelina, C.L., Chou, Y.K., Lee, T.C., Kongkam, P., Han, M.L., Wang, H.P., Chang, H.T.: Hybrid vision transformer for classification of pancreatic cystic lesions on confocal laser endomicroscopy videos. In: International Conference on Consumer Electronics-Taiwan (ICCE-Taiwan). pp. 47–48. IEEE (2023)

31. Duarte-Salazar, C.A., Castro-Ospina, A.E., Becerra, M.A., Delgado-Trejos, E.: Speckle noise reduction in ultrasound images for improving the metrological evaluation of biomedical applications: An overview. IEEE Access **8**, 15983–15999 (2020)
32. Tuthill, T., Sperry, R., Parker, K.: Deviations from Rayleigh statistics in ultrasonic speckle. Ultrason. Imaging **10**(2), 81–89 (1988)
33. Ghabri, H., Alqahtani, M.S., Ben Othman, S., Al-Rasheed, A., Abbas, M., Almubarak, H.A., Sakli, H., Abdelkarim, M.N.: Transfer learning for accurate fetal organ classification from ultrasound images: A potential tool for maternal healthcare providers. Sci. Rep. **13**(1), 17904 (2023)
34. Sendra-Balcells, C., Campello, V.M., Torrents-Barrena, J., Ahmed, Y.A., Elattar, M., Ohene-Botwe, B., Nyangulu, P., Stones, W., Ammar, M., Benamer, L.N., et al.: Generalisability of fetal ultrasound deep learning models to low-resource imaging settings in five African countries. Sci. Rep. **13**(1), 2728 (2023)

Attention Dynamics: Estimating Attention Levels of ADHD using Swin Transformer

Debashis Das Chakladar[1(✉)], Anand Shankar[2], Foteini Liwicki[1], Shovan Barma[2], and Rajkumar Saini[1]

[1] Machine Learning Group, Luleå University of Technology, 97187 Luleå, Sweden
{debashis.das.chakladar,foteini.liwicki,rajkumar.saini}@ltu.se
[2] Indian Institute of Information Technology Guwahati, Guwahati 781039, Assam, India
{anand.shankar,shovan}@iiitg.ac.in

Abstract. Children diagnosed with Attention-Deficit/Hyperactivity Disorder (ADHD) face many difficulties in maintaining their concentration (in terms of attention levels) and controlling their behaviors. Previous studies have mainly focused on identifying brain regions involved in cognitive processes or classifying ADHD and control subjects. However, the classification of attention levels of ADHD subjects has not yet been explored. Here, a robust Swin Transformer (Swin-T) model is proposed to classify the attention levels of ADHD subjects. The experimental cognitive task 'Surround suppression' includes two events: Stim ON and Stim OFF related to the high and low attention levels of a subject. In the proposed framework, ADHD-specific channels are initially identified from input Electroencephalography (EEG). Next, the significant, non-noisy connectivity features are extracted from those channels through the Singular Value Decomposition (SVD) method. Finally, the non-noisy features are passed to the robust Swin-T model for attention-level classification. The proposed model achieves 97.28% classification accuracy with 12 subjects. The robustness of the proposed model leads to potential benefits in EEG-based research and clinical settings, enhancing the reliability of ADHD assessments.

Keywords: ADHD · Electroencephalography · Singular Value Decomposition · Granger causality · Deep learning · Swin Transformer.

1 Introduction

Attention disorder is a prevalent condition among both children and adults [33]. Most of the attention disorder studies focused on the diagnosis or classification of attention deficit hyperactivity disorder (ADHD) individuals. In a Near-infrared spectroscopy (NIRS) study [38], researchers compared typically developing (TD) children with ADHD individuals in different age groups. They found that children

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15311, pp. 270–283, 2025.
https://doi.org/10.1007/978-3-031-78195-7_18

with ADHD showed reduced activation in the right and middle parts of the prefrontal cortex compared to TD children. In a Functional magnetic resonance imaging (fMRI)-based neurofeedback study, authors found that the left inferior frontal gyrus and the left insular gyrus regions have significant interaction effects in the different age groups of ADHD subjects [11]. Due to high temporal resolution, Electroencephalography (EEG) is effectively used in ADHD [24] and other applications such as emotion recognition [26], neuromarketing [27,28], and workload estimation [4,7]. The absolute power of theta and beta EEG bands are used to diagnose ADHD over control groups [9,24]. High Theta band power leads to more attention of ADHD subjects [24]. In [25], authors developed a Convolution neural network (CNN)-based framework to classify ADHD and control subjects. They achieved 98.48% classification accuracy with their model. Unlike traditional CNN, which relies on local spatial information, vision transformer (ViT) processes images globally, enabling them to grasp complex patterns and relationships across the entire image [29]. EEG-transformer has been efficiently used to classify ADHD and control subjects [15]. They achieved 95.85% classification accuracy with their model and they found that their model outperforms several traditional classifiers such as EEG-NET, shallow CNN, and deep CNN. Swin Transformer (Swin-T) offers advantages over ViT and EEG Transformer with its hierarchical processing, efficient handling of long-range dependencies, scalability to high-resolution inputs, and improved feature representation [22]. In [37], a framework combining EEG-channel attention and the Swin-T model has been developed for identifying the patterns of motor imagery tasks. The model has been effectively used in Brain-Computer Interface (BCI) rehabilitation purposes, where it achieved 87.67% average classification accuracy. Correlation-based EEG connectivity features and the Swin-T model have been used for cognitive workload assessment [21]. EEG-based weighted connectivity features (i.e.; functional/effective) have been used in different applications such as emotion recognition [39], face perception [5], and workload estimation [6] tasks. However, these connectivity features may contain noisy or insignificant connections between EEG channels [10]. Singular Value Decomposition (SVD) is useful in removing noisy connections from brain connectivity matrices while preserving the essential information [17]. SVD has been effectively used for removing artifacts [17] from raw EEG and finding dipole orientation in brain connectivity analysis [30].

Most of the existing studies [16,38] focused either on the activated brain regions of ADHD individuals or classification analysis between ADHD and control group using deep learning techniques [12,13]. However, identifying the attention levels of ADHD subjects is of utmost importance as it is related to the behavior of ADHD individuals. Therefore, in this study, the attention level of ADHD individuals is classified with the Swin-T model. Implementing a Swin-T model, an advanced deep learning model, showcases the application of cutting-edge technology in the field of cognitive neuroscience. This can inspire further research and development of new tools in the neuroscience domain. In the proposed model, initially, specific EEG channels used in ADHD diagnosis are identified based on existing studies [19,31], and weighted connectivity features are

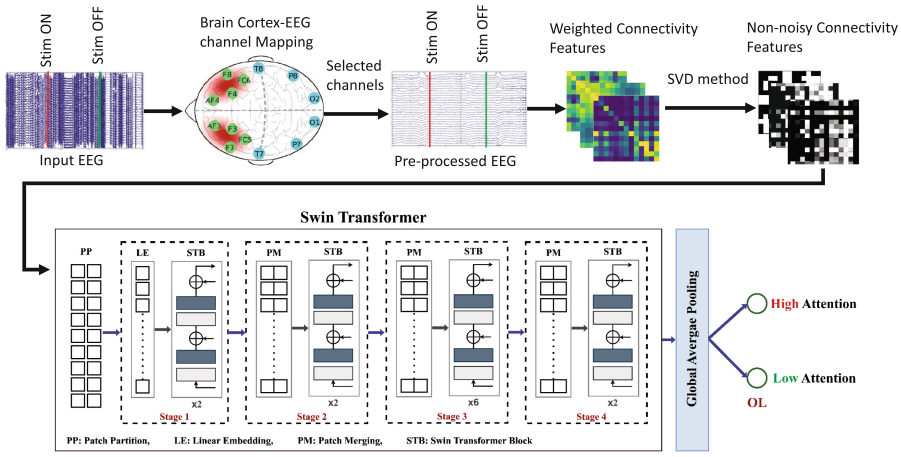


Fig. 1. Proposed framework for attention level classification using non-noisy connectivity features and Swin Transformer model.The events "Stim ON" and "Stim OFF" related to high and low attention levels respectively. OL is the output layer.

extracted from those selected EEG channels. Then SVD method is applied to eliminate the noisy connections from weighted features. The non-noisy features are passed to the Swin-T model for attention-level classification. The proposed model is shown in Fig. 1. The contributions of the proposed model are listed as follows.

1. To the best of our knowledge, this is the first approach that combines functional-effective EEG connectivity features and the Swin-T model to perform attention-level classification for ADHD participants.
2. SVD-based application mostly highlights noise reduction from input signals [17] or images [32]. Nevertheless, the implementation of non-noisy connectivity features using SVD for attention-level classification has not been discovered to date.
3. For generalization, the proposed model is also evaluated on other ADHD datasets (for the same experimental cognitive task) and our model achieves significant results on all datasets.

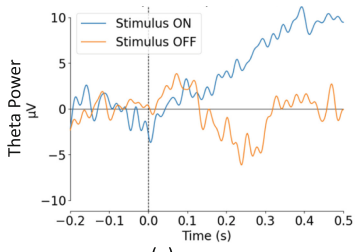
2 Methods

The proposed model is divided into three subsections: (a) Dataset and preprocessing, (b) Feature extraction & Noise reduction, and (c) Swin Transformer for classification. The details of each subsection are mentioned below.

2.1 Dataset and Preprocessing

This research used a public EEG dataset [20] containing data from developing brains and individuals with ADHD. EEG data of 12 ADHD participants (7 children: 6–10 years and 5 adolescents: 11–17 years) were recorded while they were performing the 'Surround suppression' cognitive task. In each sequence of discrete 2.4 s trials, four circular 'foreground' stimuli were flickered on and off at 25 Hz. These flickering on and off are represented as Stimulus ON/Stim ON and Stimulus OFF/Stim OFF events. The terms Stim ON and Stim OFF will be used in the subsequent sections. The high value of theta band power leads to more attention for ADHD individuals [24]. In Fig. 2(a), similar findings are found with experimental subjects, where increased theta band power of the Stim ON event leads to higher attention than the Stim OFF event. The experiment consists of 64 trials, where each trial duration is 3.4 s (0.5 s for fixation cross-presentation followed by stimulus presentation of 2.4 s followed by an inter-trial interval of 0.5 s). EEG recordings were captured using a 128-channel EEG Geodesic Hydrocel system at a sampling rate of 500 Hz.

From the existing ADHD studies [3,19,31], the relevant brain regions for ADHD diagnosis were found. Thus, the focus is placed on such EEG channels out of 128 which are related to those brain regions. The Brodmann areas (BAs) of those regions were identified, and 12 EEG channels were selected based on those BAs [1] (last access date of footnote's link: 9th April 2024). The mapping of brain regions/cortexes corresponding to EEG channels is shown in Fig. 2(b). Next, the raw EEG of those 12 channels is preprocessed using a 1-40 Hz passband filter to eliminate high-frequency noise, and then the Independent Component Analysis (ICA) is applied to remove EEG artifacts.



Brain Regions	EEG Channels (Brodmann Area)
Prefrontal cortex	Fp1(10L),Fp2(10R),F1(24L),F2(24R) AFz(32)
Posterior cingulate cortex	Pz (23)
Inferior parietal cortex	P5(39L), P6(39R),CP3(40L), CP4(40R)
Occipital cortex	O1(18), O2(18)

(b)

Fig. 2. (a): Theta band power of EEG signals for Stimulus ON/Stim ON and Stimulus OFF/Stim OFF events. Attention levels of ADHD subjects are related to Theta band power [24], (b): Mapping between brain regions and EEG channels. Here L/R denote the left and right hemispheres.

[1] https://brmlab.cz/project/brain_hacking/broadmannarea

2.2 Feature extraction and Noise reduction

In the feature extraction phase, five connectivity features (two power-based: Correlation, Coherence and three causality-based: Granger causality (GC), Granger causality spectral (Spec_GC), Phase Transfer Entropy (PTE)) are extracted for each cognitive state Stim ON and Stim OFF events (duration of 2.4 s). A detailed description of each feature is mentioned in the following subsections.

2.2.1 Power-Based Connectivity Features: Linear dependency between the time-series signals of two EEG channels [10] is measured by Pearson's correlation coefficient. Correlation coefficients typically range from -1 to 1. A high correlation value between brain regions suggests strong functional integration [10].

On the other hand, coherence indicates how similar two signals are for the same frequency component, ranging from 0 to 1. High coherence values indicate strong synchronization between brain signals at specific frequencies, suggesting a functional coupling between the corresponding brain regions [10].

2.2.2 Causality-Based Connectivity Features: GC anticipates the behavior of one time-series data using the past values of another time-series data. The GC is measured by univariate and bivariate autoregressive (AR) models, and it is calculated using variance (Var) of both the AR models' residuals (e_x, e_{xy}) (1).

$$GC(x_i \rightarrow x_j) = ln\left(\frac{Var(e_x)}{Var(e_{xy})}\right) \tag{1}$$

Spec_GC ($X_i \rightarrow X_j$) is interpreted as the fraction of the total spectral power at frequency f of X_i that is contributed by X_j. Here, the GC model of order six is chosen for proper connectivity analysis [34].

PTE is an advanced version of the transfer entropy [23]. Let's assume, for a given EEG frequency band, the instantaneous phase time series of the signal $x(t)$ is $\theta(t)$. The complex filtered signal of $x(t)$ is represented by $S(t) = x(t)exp(i\theta(t))$, which is obtained by Morlet wavelet or Hilbert transformation of the original time series signal $x(t)$. As the high activation of theta band power is mostly used in ADHD diagnosis [18], so theta frequency band is used for calculating PTE. Similarly for another time series $y(t)$ the same transformation is applied and the PTE with lag δ from $x(t)$ to $y(t)$ is computed as below [23]:

$$PTE_{x\rightarrow y} = H(\theta_y(t), \theta_y(t')) + H(\theta_y(t'), \theta_x(t')) - H(\theta_y(t')) - H(\theta_y(t), \theta_y(t'), \theta_x(t')) \tag{2}$$

where, $\theta_x(t')$ and $\theta_y(t')$ are the past states at the time point $t' : t - \delta$. The $H(\theta_y(t'))$ and $H(\theta_y(t), \theta_y(t'))$ are the marginal entropy. Joint entropy are represented by $H(\theta_y(t'), \theta_x(t'))$ and $H(\theta_y(t), \theta_y(t'), \theta_x(t')$ respectively. This equation captures the information flow from $x(t)$ to $y(t)$ at a specific delay δ through the phase dynamics of the time series.

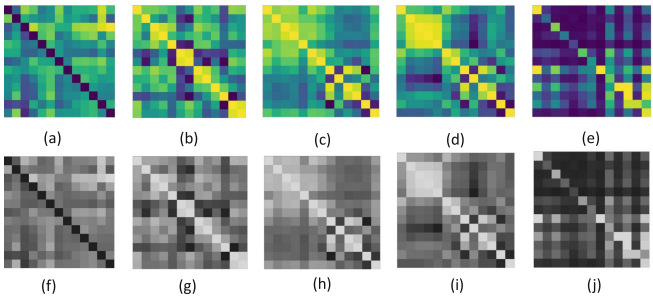


Fig. 3. Weighted and its corresponding non-noisy feature matrices of the selected EEG channels (refer to Fig. 2(b)). Weighted feature matrices (a): correlation, (b): coherence, (c): GC, (d): Spec_GC, (e): PTE. Non-noisy features: (f): correlation, (g): coherence, (h): GC, (i): Spec_GC, (j): PTE. Channel ordering in feature matrices is as follows: F2, Fp2, AFz, F1, Fp1, CP3, P5, Pz, O1, O2, CP4, P6.

2.2.3 Noise Reduction using SVD: The SVD eliminates noisy or insignificant connections from the weighted connectivity matrix [2]. SVD breaks down a matrix (A) into three smaller matrices: U, V^T, and Σ, where U and V^T hold the singular vectors, and Σ contains the singular values in order. This process uncovers hidden patterns in the original matrix. For a connectivity feature matrix A (size $N \times N$), SVD decomposes it as:

$$A = U\Sigma V^T \tag{3}$$

Where N is the optimum EEG channel. SVD can help to remove noise from data by focusing on the dominant signal components and filtering out irrelevant variations [32]. The left singular matrices (U) are truncated, retaining only the first k components to reduce data dimensionality. These truncated matrices are used to construct a non-noisy connectivity feature matrix. The weighted and its corresponding non-noisy connectivity features are shown in Fig. 3. In the SVD-based non-noisy feature matrices (Fig. 3 (f-j)), the darker cells (associated with larger singular values) represent more important or meaningful features. Conversely, the lighter cells (associated with smaller singular values) correspond to noisy connections or less significant features. The noise-free feature matrices (power/causality) are passed to the Swin-T model for attention-level classification.

2.3 Swin Transformer for classification

The Swin-T architecture having hierarchical attention mechanisms along with the utilization of shifted windows captures intricate local and global features within data, leading to effective performance in image classification [22]. In this work, the attention levels of ADHD participants are discriminated by the Swin-T model. The proposed framework of the Swin-T model is shown in Fig. 4(a). The input image size is 128×128. Initially, non-overlapping patches (patch size

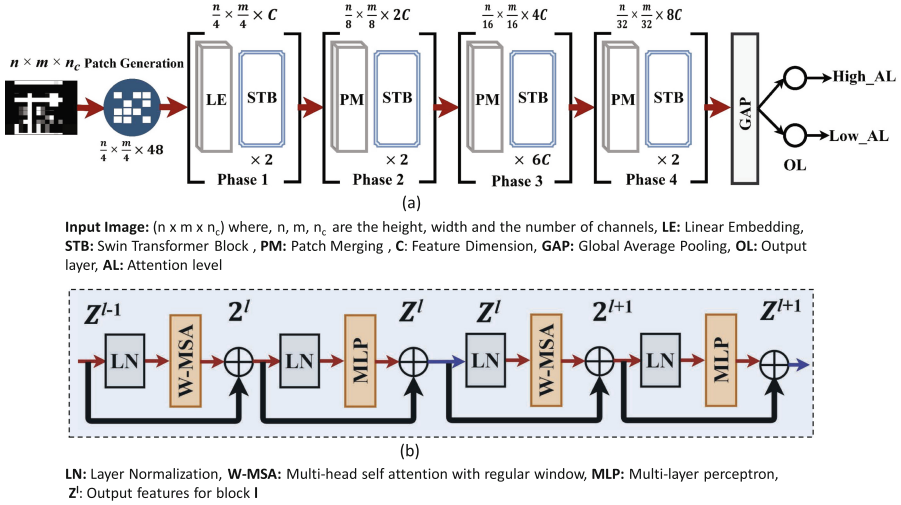


Fig. 4. (a): The proposed Swin transformer framework to classify attention levels of ADHD individuals. (b): Framework of the STB.

$= 4 \times 4$) are generated from the input images. Each patch is considered as a token, which is used to measure the feature (patch sizes $\times$ number of channels in input image=$4 \times 4 \times 3 = 48$). In phase 1, sequence-wise features are extracted and passed through a linear embedding (LE) layer to project the features into arbitrary dimensions (C). The resultant data is fed into the Swin-T block (STB). In this phase, the transformer block maintains the number of tokens $(n/4, m/4)$. In phase 2, the merging of tokens and transformation of features is performed by patch merging (PM) and STB layers respectively. PM is performed by shifted window-based multi-head self-attention technique (W-MSA) followed by MLP layer. Features of each group of 2×2 neighboring patches are concatenated using the PM layer and a linear layer is applied to the 4C dimensional concatenated features. Additionally, it reduces the tokens number by a multiple of $2 \times 2 = 4$ along with output features dimension to 2C. In phase 2, the final image resolution and output feature dimension are $n/8 \times m/8$ and $2C$. In phase 3 and phase 4, the number of output feature dimensions is set to $4C$ and $8C$ respectively, while keeping the input image resolution as $n/16 \times m/16$ and $n/32 \times m/32$ respectively to generate the hierarchical representation of the features maps. Finally, the feature maps of phase 4 are passed to the global average pooling (GAP) layer followed by the output layer (OL) for the classification of attention levels. The pipeline of individual STB is shown in Fig. 4(b). In the proposed STB, the number of window size, shift size, and number of MLP units are set to 7, 0, and 1024 respectively, along with the embedding dimension of 64. The hierarchical feature representation enhances the scalability of the proposed model by reducing the computational complexity.

3 Results

This section is divided into four subsections: (a) Optimizing classification performance through the SVD method, (b) Classification analysis, (c) Performance analysis, and (d) Comparison analysis. A detailed description of each subsection is mentioned below.

3.1 Optimizing classification performance through SVD method

SVD reduces noisy connections from the weighted matrix and this reduction is done by selecting the most significant k components of the left singular matrices. Non-noisy feature matrices are then created from these k components. A smaller k value retains only the crucial singular values and vectors, compressing the matrix. Conversely, a larger k value preserves more information but increases computational complexity and memory usage [32]. Therefore, it's crucial to find the optimal k value during SVD computation to minimize noise and maximize classification performance. Non-noisy connectivity matrices with different k components are generated, and classification is performed. The classification results of the Swin-T model for different k values are shown in Fig. 5. It's evident that a k value of 15 yields the highest classification accuracy. Thus, all non-noisy feature matrices are computed using that k value.

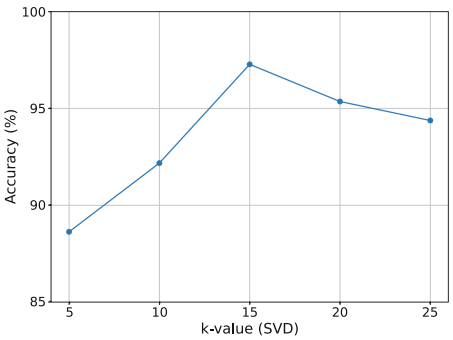
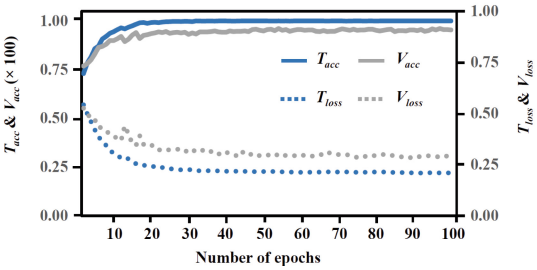


Fig. 5. Performance of the proposed Swin-T model with different k values of SVD method. It is a trade-off graph between the performance of the Swin-T model and information loss during the SVD process.

3.2 Classification analysis

The proposed Swin-T model is trained with a batch size of 50 for 100 epochs. The binary cross-entropy loss function and the Adam optimizer (learning rate of 1e-03) are used during the training process. Initially, input data is divided into the ratio of 80:20 for training and testing of the model; further, 20% of training data is used for the validation set. Once the training is completed, the

prediction is performed on the test set. The loss /accuracy curve (training and validation) of the proposed model is shown in Fig. 6(a). From the curve, it can be concluded that both training and validation losses are decreased with the epochs. In contrast, both accuracies are increased over the epochs. The classification accuracy of 97.28% is obtained using the proposed Swin-T model. The class-wise classification result is shown in Fig. 6(b). The result is presented based on different classification parameters, such as precision, recall, and F_1 score.



(a)

Class	Precision	Recall	F_1 Score
Stim ON	0.9782	0.9829	0.9805
Stim OFF	0.9638	0.9610	0.9623

(b)

Fig. 6. Classification results of the proposed model: (a) Accuracy/Loss curve of the Swin-T model. Here, T and V in legends/y-axis denote the training and validation. (b) Class-wise classification output of the Swin-T model.

3.3 Performance analysis

For generalization, the proposed model is evaluated using three versions (Release 8, Release 9, and Release 10) of the CMI-HBN dataset [1]. All the datasets use the same cognitive task as the experimental task (i.e.; 'Surround suppression'). Here, 15 subjects with attention disorders were randomly selected from each dataset. Cognitive event (Stim ON/Stim OFF)-wise non-noisy features have been extracted for all subjects, and those features are passed to the Swin-T model for attention-level classification. Table 1 displays the performance analysis of the proposed model for each dataset. It's evident that our model performs notably well across all datasets.

Table 1. Performance analysis of the Swin-T model based on other EEG-based attention disorder datasets

Dataset	Accuracy(%)
CMI-HBN (Release 8)[1]	91.09
CMI-HBN (Release 9)[1]	87.14
CMI-HBN (Release 10)[1]	88.22

3.4 Comparison analysis

This section is divided into three subsections: (a) Comparison between weighted and non-noisy features, (b) Comparison with Deep learning models, and (c) Comparison with existing studies. A detailed discussion of each subsection is mentioned below.

3.4.1 Comparison between weighted and non-noisy features: In this section, the effectiveness of the SVD method is evaluated by comparing the classification performance using weighted and non-noisy features (extracted from SVD). ADHD subject-wise, all the weighted connectivity features were combined and passed to the Swin-T model for evaluation. The same process was performed for non-noisy features. The same training parameters as mentioned in subsection 3.2 were kept. Removing noisy features can prevent overfitting in classification models. Additionally, SVD-based feature extraction enhances the discriminative power of the features, leading to better classification performance. In Fig. 7, a significant improvement of 11.05% is achieved in the classification result, with the SVD-based non-noisy features over weighted connectivity features.

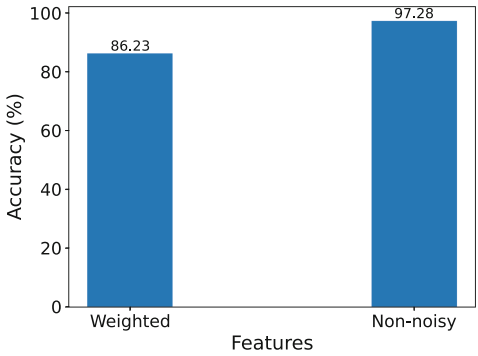


Fig. 7. Comparative analysis between classification performance of Swin-T model using weighted and non-noisy connectivity features (extracted from SVD).

3.4.2 Comparison with Deep Learning Models: Here, the performance of the proposed Swin-T model is compared with other deep learning models and the ViT model (refer to Table 2). The same set of connectivity features is used for this comparison. During training, similar training parameters are reserved as discussed in subsection 3.2. The model structure/ parameters of each model are mentioned in Table 2. It can be concluded that the proposed model performs better than all the other deep models and ViT.

Table 2. Comparison with other deep learning models. Note: Convolution(C), MaxPooling(M), Flatten(F), Dense(D), Multihead Self Attention(MSA), Transformer Encoder(TE), Patch Encoder(PE), Layer Normalization(LN), Multilayer perception(MLP), SwinTransfomrerBlock(STB), GlobalAveragePooling(GAP), Output layer(OL), Window Size(WS), Shift size(S_size), Accuracy(Acc)

Deep Model	Configuration/Parameters	Acc(%)
ShallowConvNet	C16-M2-C32-M2-C64-M2-C128-M2-F-D500-OL	90.09
DeepConvNet	C16-M2-C32-M2-C64-M2-C128-M2-C256-M2-C512-M2-F-D500-OL	92.25
ViT	PE-TE(LN-MSA-LN-MLP128-MLP64)-LN-F-MLP2048-MLP1024-OL, Transformer layers=6, Patch size =6, Projection dimension=64	94.37
Swin-T(Proposed)	STB1-STB2-STB3-STB4-GAP-OL. WS=7, S_size=0, number of MLP units =1024	97.28

3.4.3 Comparison with Existing Studies: In this section, existing ADHD classification studies [8,14,35,36] have been compared with the proposed model. The comparison (refer to Table 3) is performed based on the methodology used in those studies for ADHD classification. The performance of those studies is also reported in Table 3. From this table, it can be concluded that the proposed model outperforms all the existing studies of ADHD classification.

Table 3. Comparison with existing studies of ADHD classification. Note: power spectral density(PSD), local mode decomposition (LMD), and variational mode decomposition (VMD)

Studies	Method	Acc(%)
Tosun *et al.* [35]	PSD and LSTM	88.88
Vahid *et al.* [36]	EEGNET	83.00
Chen *et al.* [8]	EEG-based brain network with the CNN	94.67
Esas *et al.* [14]	LMD, VMD and CNN	95.00
Swin-T(Proposed)	Connectivity features and Swin Transformer	97.28

4 Conclusion

This study combined two aspects in the attention level classification of ADHD participants. First, finding significant connections between EEG channels through the SVD method and second, classifying the attention level using the robust Swin-T model. The proposed model achieved 97.28% classification accuracy. As noisy connections often lead to spurious connectivity and influence wrong information flow among brain regions, therefore, removing such noisy

connections is of utmost importance. The SVD method eliminates such connections and increases the classification performance (refer to Fig. 7). A significant improvement of 11.05% in the classification result is achieved after removing the noisy, non-significant connections from the feature matrices. For generalization, the proposed model is also evaluated using other CMI-HBN datasets, and significant results are achieved for all the datasets. In this study, the experiment is performed with a small number (i.e., 12) of ADHD subjects. Therefore, in the near future, the proposed model will be evaluated using ADHD datasets with more subjects, and the attention level of an ADHD individual will be measured with more cognitive tasks.

References

1. Alexander, L.M., Escalera, J., Ai, L., Andreotti, C., Febre, K., Mangone, A., Vega-Potler, N., Langer, N., Alexander, A., Kovacs, M., et al.: An open resource for transdiagnostic research in pediatric mental health and learning disorders. Scientific data **4**(1), 1–26 (2017)
2. Boschi, A., Brofiga, M., Massobrio, P.: Thresholding functional connectivity matrices to recover the topological properties of large-scale neuronal networks. Front. Neurosci. **15**, 705103 (2021)
3. Cao, M., Martin, E., Li, X.: Machine learning in attention-deficit/hyperactivity disorder: new approaches toward understanding the neural mechanisms. Transl. Psychiatry **13**(1), 236 (2023)
4. Chakladar, D.D., Datta, S., Roy, P.P., Vinod, A.: Cognitive workload estimation using variational auto encoder & attention-based deep model. IEEE Transactions on Cognitive and Developmental Systems (2022)
5. Chakladar, D.D., Pal, N.R.: Brain connectivity analysis for EEG-based face perception task. IEEE Transactions on Cognitive and Developmental Systems (2024)
6. Chakladar, D.D., Roy, P.P., Iwamura, M.: EEG-based cognitive state classification and analysis of brain dynamics using deep ensemble model and graphical brain network. IEEE Transactions on Cognitive and Developmental Systems **14**(4), 1507–1519 (2021)
7. Chakladar, D.D., Samanta, D., Roy, P.P.: Multimodal deep sparse subspace clustering for multiple stimuli-based cognitive task. In: 2022 26th International Conference on Pattern Recognition (ICPR). pp. 1098–1104. IEEE (2022)
8. Chen, H., Song, Y., Li, X.: A deep learning framework for identifying children with adhd using an EEG-based brain network. Neurocomputing **356**, 83–96 (2019)
9. Clarke, A.R., Barry, R.J., Johnstone, S.J., McCarthy, R., Selikowitz, M.: EEG development in attention deficit hyperactivity disorder: From child to adult. Clin. Neurophysiol. **130**(8), 1256–1262 (2019)
10. Cohen, M.X.: Analyzing neural time series data: theory and practice. MIT press (2014)
11. Criaud, M., Wulff, M., Alegria, A., Barker, G., Giampietro, V., Rubia, K.: Increased left inferior fronto-striatal activation during error monitoring after fMRI neurofeedback of right inferior frontal cortex in adolescents with attention deficit hyperactivity disorder. NeuroImage: Clinical **27**, 102311 (2020)
12. Dong, Q., Qiang, N., Lv, J., Li, X., Liu, T., Li, Q.: Spatiotemporal attention autoencoder (STAAE) for ADHD classification. In: Medical Image Computing and

Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VII 23. pp. 508–517. Springer (2020)
13. Dubreuil-Vall, L., Ruffini, G., Camprodon, J.A.: Deep learning convolutional neural networks discriminate adult ADHD from healthy individuals on the basis of event-related spectral EEG. Front. Neurosci. **14**, 251 (2020)
14. Esas, M.Y., Latifoğlu, F.: Detection of adhd from EEG signals using new hybrid decomposition and deep learning techniques. J. Neural Eng. **20**(3), 036028 (2023)
15. He, Y., Wang, X., Yang, Z., Xue, L., Chen, Y., Ji, J., Wan, F., Mukhopadhyay, S.C., Men, L., Tong, M.C.F., et al.: Classification of attention deficit/hyperactivity disorder based on eeg signals using a EEG-transformer model. J. Neural Eng. **20**(5), 056013 (2023)
16. Hong, J., Park, B.y., Cho, H.h., Park, H.: Age-related connectivity differences between attention deficit and hyperactivity disorder patients and typically developing subjects: a resting-state functional MRI study. Neural regeneration research **12**(10), 1640 (2017)
17. Judith, A.M., Priya, S.B., Mahendran, R.K.: Artifact removal from EEG signals using regenerative multi-dimensional singular value decomposition and independent component analysis. Biomed. Signal Process. Control **74**, 103452 (2022)
18. Kiiski, H., Bennett, M., Rueda-Delgado, L.M., Farina, F.R., Knight, R., Boyle, R., Roddy, D., Grogan, K., Bramham, J., Kelly, C., et al.: EEG spectral power, but not theta/beta ratio, is a neuromarker for adult ADHD. Eur. J. Neurosci. **51**(10), 2095–2109 (2020)
19. Konrad, K., Eickhoff, S.B.: Is the ADHD brain wired differently? a review on structural and functional connectivity in attention deficit hyperactivity disorder. Hum. Brain Mapp. **31**(6), 904–916 (2010)
20. Langer, N., Ho, E.J., Alexander, L.M., Xu, H.Y., Jozanovic, R.K., Henin, S., Petroni, A., Cohen, S., Marcelle, E.T., Parra, L.C., et al.: A resource for assessing information processing in the developing brain using EEG and eye tracking. Scientific data **4**(1), 1–20 (2017)
21. Li, Z., Zhang, R., Zeng, Y., Tong, L., Lu, R., Yan, B.: Mst-net: A multi-scale swin transformer network for EEG-based cognitive load assessment. Brain Res. Bull. **206**, 110834 (2024)
22. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
23. Lobier, M., Siebenhühner, F., Palva, S., Palva, J.M.: Phase transfer entropy: a novel phase-based measure for directed connectivity in networks coupled by oscillatory interactions. Neuroimage **85**, 853–872 (2014)
24. Markovska-Simoska, S., Pop-Jordanova, N.: Quantitative in children and adults with attention deficit hyperactivity disorder: comparison of absolute and relative power spectra and theta/beta ratio. Clin. EEG Neurosci. **48**(1), 20–32 (2017)
25. Moghaddari, M., Lighvan, M.Z., Danishvar, S.: Diagnose ADHD disorder in children using convolutional neural network based on continuous mental task EEG. Comput. Methods Programs Biomed. **197**, 105738 (2020)
26. Panda, D., Chakladar, D.D., Dasgupta, T.: Multimodal system for emotion recognition using EEG and customer review. In: Proceedings of the global ai congress 2019. pp. 399–410. Springer (2020)

27. Panda, D., Chakladar, D.D., Rana, S., Parayitam, S.: An EEG-based neuro-recommendation system for improving consumer purchase experience. J. Consum. Behav. **23**(1), 61–75 (2024)
28. Panda, D., Chakladar, D.D., Rana, S., Shamsudin, M.N.: Spatial attention-enhanced EEG analysis for profiling consumer choices. IEEE Access (2024)
29. Qi, N., Piao, Y., Zhang, H., Wang, Q., Wang, Y.: Seizure prediction based on improved vision transformer model for EEG channel optimization. Computer Methods in Biomechanics and Biomedical Engineering pp. 1–12 (2024)
30. Rubega, M., Carboni, M., Seeber, M., Pascucci, D., Tourbier, S., Toscano, G., Van Mierlo, P., Hagmann, P., Plomp, G., Vulliemoz, S., et al.: Estimating EEG source dipole orientation based on singular-value decomposition for connectivity analysis. Brain Topogr. **32**, 704–719 (2019)
31. Rubia, K.: Cognitive neuroscience of attention deficit hyperactivity disorder (ADHD) and its clinical translation. Front. Hum. Neurosci. **12**, 100 (2018)
32. Scetbon, M., Elad, M., Milanfar, P.: Deep k-svd denoising. IEEE Trans. Image Process. **30**, 5944–5955 (2021)
33. Slater, J., Joober, R., Koborsy, B.L., Mitchell, S., Sahlas, E., Palmer, C.: Can electroencephalography (EEG) identify ADHD subtypes? a systematic review. Neuroscience & Biobehavioral Reviews **139**, 104752 (2022)
34. Tadel, F., Baillet, S., Mosher, J.C., Pantazis, D., Leahy, R.M.: Brainstorm: a user-friendly application for MEG/EEG analysis. Computational intelligence and neuroscience **2011** (2011)
35. Tosun, M.: Effects of spectral features of EEG signals recorded with different channels and recording statuses on adhd classification with deep learning. Physical and Engineering Sciences in Medicine **44**(3), 693–702 (2021)
36. Vahid, A., Bluschke, A., Roessner, V., Stober, S., Beste, C.: Deep learning based on event-related EEG differentiates children with ADHD from healthy controls. J. Clin. Med. **8**(7), 1055 (2019)
37. Wang, H., Cao, L., Huang, C., Jia, J., Dong, Y., Fan, C., De Albuquerque, V.H.C.: A novel algorithmic structure of EEG channel attention combined with swin transformer for motor patterns classification. IEEE Transactions on Neural Systems and Rehabilitation Engineering (2023)
38. Yasumura, A., Omori, M., Fukuda, A., Takahashi, J., Yasumura, Y., Nakagawa, E., Koike, T., Yamashita, Y., Miyajima, T., Koeda, T., et al.: Age-related differences in frontal lobe function in children with ADHD. Brain Develop. **41**(7), 577–586 (2019)
39. Yu, M., Xiao, S., Hua, M., Wang, H., Chen, X., Tian, F., Li, Y.: EEG-based emotion recognition in an immersive virtual reality environment: From local activity to brain network features. Biomed. Signal Process. Control **72**, 103349 (2022)

CycleGAN Based Bone Suppression Techniques for Standard CXR Images

V. Santhosh Kumar Tangudu(✉), Jagadeesh Kakarla, A. Sam Prabhu, and Bala Vikranth Kumar

Indian Institute of Information Technology, Design and Manufacturing, Kancheepuram, Chennai, India
t.v.santhoshkumar@gmail.com

Abstract. CycleGAN is an image translation technique that can successfully suppress bones in dual-energy X-ray images. This study introduces two novel variations of the CycleGAN model, namely CycleGAN-Idempotent Loss (CGIL) and CGIL with additional layers (CGILAL). The PSNR, MS-SSIM, and VIF measures are used to compare the performance of CGIL and CGILAL models with CycleGAN and Pix2Pix GAN models. CGIL and CGILAL showed better performance than CycleGAN and Pix2Pix in terms of these metrics when applied to the dual-energy chest x-ray dataset. Here, these trained models are used to construct bone-suppressed images from standard X-rays in the CovidX dataset, where there are no equivalent bone-suppressed images available. The standard CXR images without bone suppression and bone-suppressed standard X-ray images are currently being used for the multi-class categorization of Covid, Pneumonia, and Normal images. A total of 16,965 images in each dataset are used for the study, with 5,655 images from each class. A total of 13,572 images are used for training, whereas 3,393 images are used for testing. The accuracy of standard CXR images without bone suppression is 90.07%. The accuracy of CGIL and CGILAL is better than that of CycleGAN and Pix2Pix, with CGIL achieving 95.14% accuracy and CGILAL achieving 95.76% accuracy, compared to CycleGAN's accuracy of 94.49% and Pix2Pix's accuracy of 94.22%. The findings indicate that models trained on bone-suppressed images exhibit superior accuracy in comparison to standard X-ray images without bone suppression.

Keywords: GAN, Dual-energy X-ray, Classification

1 Introduction

Deep learning has transformed medical image processing by enabling tasks such as segmentation, classification, localization, and disease identification tasks. However, there are challenges with dataset generation such as noise, uncurated, lack of trust, and data anonymity issues for medical image data. Researchers are

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15311, pp. 284–294, 2025.
https://doi.org/10.1007/978-3-031-78195-7_19

exploring collaborations for data-sharing and enhancement strategies with open-source datasets. Data augmentation techniques increase training set diversity without the addition of new patient data, allowing deep learning to advance disease diagnosis. Data augmentation in medical image processing can be achieved using image processing as well as deep learning-based techniques. Image processing libraries can perform operations such as cropping, blurring, zooming, flipping, and rotating. Deep learning techniques like GANs create synthetic data by learning patterns from training data, which makes the model work better in medical image processing applications.

GANs[1] are made up of a generator and a discriminator, which together generate realistic data. The generator outputs the data that matches the real data to fake the discriminator, whereas the discriminator distinguishes between actual and fake data. In the GAN training procedure, these two models will be adjusted based on the losses sustained. The block diagram of GAN is shown in Fig. 1.

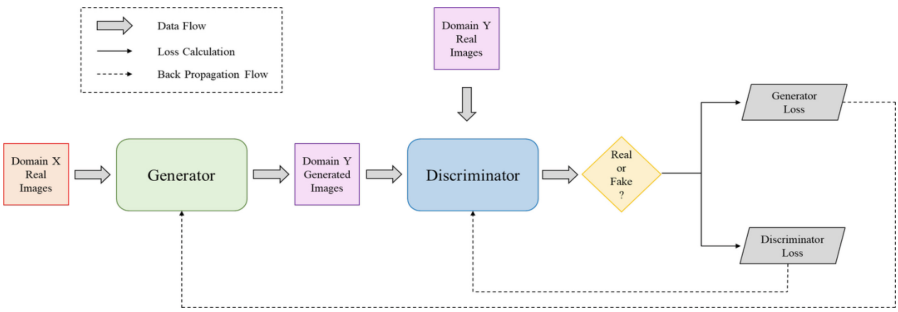


Fig. 1. Work-flow of the Generative Adversarial Networks

Standard chest X-rays (CXRs) are affordable even in distant places and are effective for diagnosing pneumonia, tuberculosis, lung tumors, and COVID. Research shows that dual-energy chest radiographs (CXRs) may detect abnormalities more accurately than standard chest X-rays. Dual-energy chest X-rays use calcium to reduce X-ray radiation to identify bones from soft tissues. Dual-energy chest X-rays increase patient's exposure to radiation and require expensive, specialized equipment. Thus, researchers investigated chest X-ray bone suppression techniques. GANs are promising deep-learning approaches for suppressing bones in chest X-ray images, improving the visibility of soft tissues, and facilitating accurate diagnosis. This task can be achieved with CycleGAN[2], Pix2Pix GAN[3], and others.

The deep learning models trained on bone-suppressed standard CXR images have given better results in comparison to standard CXR images with bones for the classification tasks. It is recommended to convert standard chest X-ray (CXR) images with bones into bone-suppressed standard CXR images to improve classification results. In this study, a multi-class classification task is

used to demonstrate the effect of the bone suppression techniques on standard CXR images.

In section.2, the existing literature on bone suppression techniques is discussed. In section.3, CycleGAN technique, including proposed variants for bone suppression, and a classification task to validate the performance of bone suppression techniques are discussed. In section.4, the results are discussed. Finally, the conclusion is discussed in the section.5

2 Related Work

Image-to-image translation is a technique that preserves context while mapping between two sets of images. These two sets can be paired or unpaired. Training using paired images yields better outcomes compared to training with unpaired images. However, the accessibility of paired images is not always feasible[2]. The image-to-image translation approach has been applied in several applications such as super-resolution, segmentation, and object transformation, among others [4]. Several deep learning methods for image-to-image translation exist, including generative adversarial networks (GANs) [5], variational autoencoders (VAEs) [6], and conditional generative models [7].

The incorporation of more layers into a deep learning model can profoundly impact its performance and behavior. One advantage of deeper models is their ability to capture complex features, such as detailed patterns and visual representations. Each layer in these models learns progressively more abstract representations[8]. An example is the VGG19 model, which exhibited reduced percentages of top-5 and top-1 errors compared to VGG16 and VGG13 when trained on the ImageNet dataset[9]. The vanishing gradient problem arises when the number of layers is added beyond a specific threshold[10], resulting in a decrease in the efficacy of further layers. In light of this, this paper showcases the comparison of our model with its multiple-layer equivalent.

Rajaraman S *et al.*[11] demonstrated the effect of bone suppression by assessing the gain in performance in detecting pulmonary abnormalities consistent with COVID-19 disease. Geeta Rani *et al.*[12] used bone suppression techniques to enhance the performance of COVID-19 classification. CycleGAN was used for bone suppression on unpaired image data[13]. Mochizuki *et al.*[14] used a CycleGAN-based deep learning model to generate the bone-suppressed images, which improved image recognition, making it possible to achieve accurate motion tracking irradiation. Similarly, Luyi *et al.*[15] proposed RS-GAN for rib suppression. The results of the model generated from CXR images lead to a better performance in lung disease classification and tuberculosis area detection.

Kida *et al.*[16] proposed a modified U-net for enhancing CBCT quality of the image and removing any unwanted artifacts. Y. Yan *et al.*[17] proposed a Lung Tumor Localization methodology in which multiple U-Net models were used to perform various tasks including Bone Suppression. Xiaotang *et al.*[18] proposed a modified U-Net architecture integrated with residual skip connections to remove obstructing bony structures in CTS projections to reconstruct bone-suppressed

projections. K. Sato *et al.*[19] developed a method for Lung Tumor detection that has improved performance due to the use of U-Net for the suppression of bones in a particular step. J. Xie [20] proposed a model based on Pix2pix for bone-suppression. It consists of a U-Net structure for the generator and PatchGAN for the discriminator.

3 Methodology

3.1 Proposed Models

CycleGAN model depicted in Fig. 2 has identical designs for both generators (G_{BS-B} and G_{B-BS}) and both discriminators (D_{BS} and D_B). Assuming there are two domains, B represents dual-energy X-ray images with bones, and BS represents dual-energy X-rays with bone suppression. Here, the generator G_{B-BS} converts the original image B to the image BS, and the other generator G_{BS-B} converts the image BS to the image B. The discriminator D_{BS} and D_B are used to ascertain whether the images from the BS and B domains, respectively, are real or fake. As shown in Fig. 2, the forward cycle indicated by a red line, the image B_input fed into G_{B-BS} to generate $BS\_generated$, which is then used as an input to G_{BS-B} to generate B_cyclic. If the model is working as expected, there shouldn't be any difference for B_input and B_cyclic. Otherwise, the weights in the model are to be adjusted to reduce the loss, represented by Cycle_Loss-1 in the image. Similarly, in the backward cycle represented by the blue line, the image BS_input given as an input to G_{BS-B} to generate $B\_generated$, which is then used as an input to G_{B-BS} to generate BS_cyclic. The loss between BS_input and BS_cyclic is represented by Cycle_Loss-2.

Cycle_Loss is the sum of Cycle_Loss-1 and Cycle_Loss-2. The discriminator D_{BS} takes BS_input and $BS\_generated$ images and detects whether $BS\_generated$ is real or fake. Similarly, discriminator D_B takes B_input and $B\_generated$ and detects whether $B\_generated$ is real or fake. The GAN loss, represented by the sum of GAN_Loss-1 and GAN_Loss-2, is an adversarial loss computed using the discriminator's prediction. The GAN loss is determined by the discriminator's accuracy in differentiating between real and fake images. The CycleGAN uses an additional loss function called identity loss, which is the sum of Identity_Loss-1 and Identity_Loss-2. The generators in [2] are designed to perform an almost perfect mapping to the target domain when given images from that domain as input. Otherwise, there will be an identity loss. The identity loss as described in equation (1), helps maintain the color composition of the images.

$$\begin{aligned}\mathcal{L}_{identity}(G_{B-BS}, G_{BS-B}) = \\ \mathbb{E}_{BS\_input \sim p_{data(BS\_input)}} [\|G_{B-BS}(BS\_input) - BS\_input\|_1] \\ + \mathbb{E}_{B\_input \sim p_{data(B\_input)}} [\|G_{BS-B}(B\_input) - B\_input\|_1] \end{aligned} \quad (1)$$

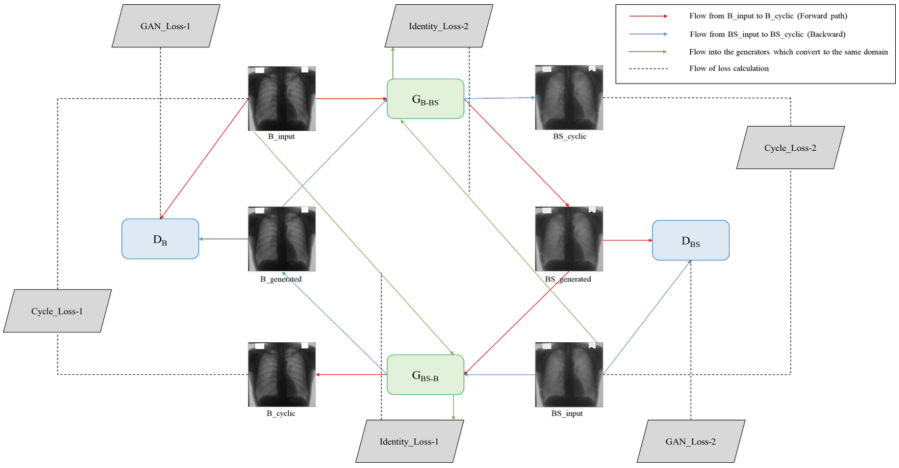


Fig. 2. Work-flow of CycleGAN

In this study, two new models, CycleGAN with Idempotent Loss (CGIL) and CGIL with additional layers (CGILAL), proposed to suppress bones in CXR images. In the conventional CycleGAN framework, identity loss is used as a metric to quantify the loss incurred during the process of identity mapping. To enhance stability, the identity loss indicated in equation (1) can be substituted with an idempotent loss shown in equation (2) in CycleGAN. In case of identity loss, the generator takes the original image from the target domain as an input and is expected to make no modifications. The idempotent loss is calculated based on the changes made to the image produced by the generator, which is then used as input to the same generator. Idempotent loss helps in maintaining consistency by limiting excessive modification of the image during translation. The CycleGAN with Idempotent Loss (CGIL) model is an identical version of the CycleGAN model proposed in [2], except for one of the loss functions. In this case, the identity loss function is substituted with an idempotent loss function. The generator shouldn't modify the image when the target domain image is given as input.

The CGILAL model is an enhanced version of CGIL that has been trained using a greater number of residual blocks. The generator in traditional CycleGAN models originally consisted of nine residual blocks. Multiple experiments have been undertaken to test the impact of increasing the number of residual blocks on performance. The best performance was achieved while using twelve residual blocks in the generator.

The CGIL and CGILAL models are trained using the dataset acquired from [18]. The training dataset is enhanced with an additional 35 pairs of dual-energy chest x-ray images. Data augmentation involves performing transformations such

as rotations, as well as horizontal and vertical shifts to increase the dataset size to 4080 dual-energy CXR images.

$$\mathcal{L}_{idempotent}(G_{B-BS}, G_{BS-B}) = \mathbb{E}_{B\_input \sim p_{data(B\_input)}} \left[\| G_{B-BS}(B\_input) - G_{B-BS}(G_{B-BS}(B\_input)) \|_1 \right] + \mathbb{E}_{BS\_input \sim p_{data(BS\_input)}} \left[\| G_{BS-B}(BS\_input) - G_{BS-B}(G_{BS-B}(BS\_input)) \|_1 \right] \quad (2)$$

3.2 Datasets

Two distinct datasets are used in this study to assess the proposed models. The first dataset outlined in [21] comprises of dual-energy X-ray images. The purpose of this dataset is to train and evaluate the bone-suppression capabilities of the proposed models. The dataset has a total of 4,080 medical images. The trained models have been used on conventional CXR images in the second dataset, known as COVIDx CXR-3 dataset[22]. This dataset is a publicly available collection of 30,882 CXR images from 17,026 patient cases. It was created by combining various other publicly available datasets. The dataset comprises 16,194 CXR images of COVID-19, 5,655 CXR images of Pneumonia, and 8,185 CXR images of Normal cases. BRISQUE [23] image scoring technique has been used to downsample and choose 5655 images from each category, resulting in a total of 16965 images.

3.3 Application

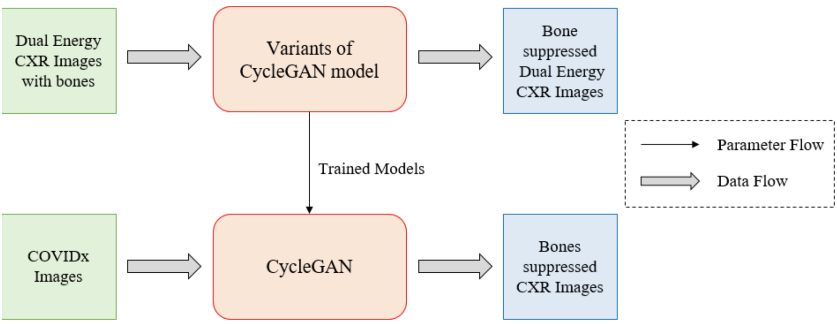


Fig. 3. Application of these trained models to suppress the bones in COVIDx dataset

Dual-energy CXR images with bone and bone suppression are necessary for training the proposed models, as it is not possible to obtain bone-suppressed CXR images. The models trained with dual-energy CXR images will be used to suppress the bones from standard CXR images. Bone-suppressed images will help in the rapid diagnosis of chest-related conditions such as pneumonia, tuberculosis,

lung cancer, and COVID-19. A multi-class classification pipeline as shown in Fig. 4, classifies COVID, Pneumonia, and Normal images, is used to evaluate the effectiveness of bone suppression in CXR images.

CXR Images can be used to detect COVID-19 and for treatment, essential for effective COVID-19 management. Early detection of COVID-19 enables appropriate medical intervention to be taken to lessen the severity of the illness and isolation to control its spread. Nonetheless, the most common methods for detecting COVID-19 are the RT-PCR test and medical imaging techniques. The RT-PCR gives high false-negative results[24]. Therefore, quicker and more cost-effective detection mechanisms are required. Standard CXR images are useful in the early identification of COVID, but there is a high possibility of false positives and false negatives. Due to these inaccuracies, CXR is being utilized as a supplementary rather than a primary instrument. Increasing the accuracy of COVID-19 detection using only CXR images expedites treatment.

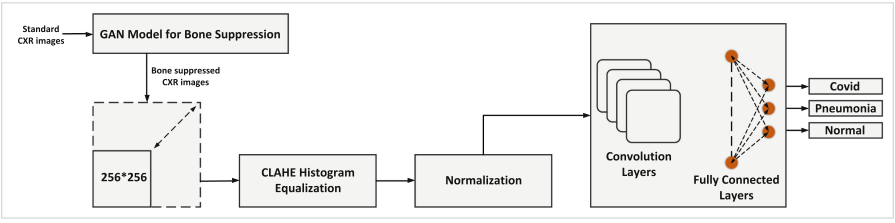


Fig. 4. Multi-class classification pipleine

Bone-suppressed CXR images contribute to improved classification outcomes. However, the existing public CXR COVID-19 datasets lack bone-suppressed CXR images to enhance the model accuracy. As a result, the method depicted in Fig. 3 is proposed to generate bone-suppressed images for COVID-related datasets. The proposed models will be trained using dual-energy CXR image datasets to get the bone suppression capabilities, and the resulting models will be used to generate bone-suppressed standard CXR images. Here, the bone-suppressed images for standard CXR images in COVIDx CXR-3 [22] dataset are generated using proposed trained models.

The bone-suppressed images obtained from the models used for multi-class classification using the pipeline illustrated in Fig. 4 involve several stages. Prior to model training, the bone-suppressed images will undergo several preprocessing tasks. The CXR images have been scaled to a dimension of 256x256. They have undergone histogram equalization using the CLAHE[25] algorithm and normalized. The preprocessed images are trained using MobileNetV2. The purpose of this is to accurately classify the CXR images into three classes: COVID, Pneumonia, and Normal. Additionally, it aims to demonstrate that using bone-suppressed images yields higher accuracy as compared to using the original images

4 Results

The proposed models are being evaluated in comparison to CycleGAN and Pix2Pix. The four models, namely CGIL, CGILAL, CycleGAN, and Pix2Pix are trained and assessed using the dual-energy X-ray dataset. The dataset comprises 4,080 medical images, with 75% allocated for training and 25% for testing. This means that approximately 3,000 images are used for training and 1,000 images are used for testing. The training of these models is conducted on the Google Colab Platform, which is equipped with GPU capabilities. Each model underwent training for 30 epochs using the training data in batches of two during the training process. For the sake of maintaining comparable input dimensions across models during the training and testing process, all images were uniformly resized to 256x256 pixels.

The performance of the models is evaluated using three key metrics such as the Peak Signal-to-Noise Ratio (PSNR), the Multi-Scale Structural Similarity Index Measure (MS-SSIM), and the Visual Information Fidelity (VIF). PSNR is used to compare the image quality of the bone-suppressed images generated by the models to the image quality of the original images containing bone. MS-SSIM quantifies the quality of an image by evaluating it at various levels of detail. The computation involves a weighted combination of contrast, structural comparison, and luminance comparison. The Visual Information Fidelity (VIF) metric quantifies the degree of similarity between visual information in a distorted image and the corresponding original image.

Table 1 demonstrates the PSNR, MS-SSIM, and VIF values for all four models. These results are obtained from 25% of dual-energy CXR images which is approximately 1000 images. The results show that CGIL outperforms CycleGAN across all metrics. Similarly, CGILAL outperforms the CGIL model across all metrics.

Table 1. Comparision of PSNR, MS-SSIM, VIF metrics across various models as Mean±SE

	Pix2Pix	CycleGAN	CGIL	CGILAL
PSNR	29.8±1.24	30.05±3.37	31.00±3.21	32.4±2.26
MS-SSIM	0.972±0.008	0.976±0.008	0.979±0.006	0.98±0.002
VIF	0.72±0.045	0.67±0.040	0.66±0.037	0.66±0.072

The CovidX CXR-3 dataset is used to evaluate the impact of these bone suppression strategies. Out of the total 30,882 CXR images in this dataset, 16,965 images are chosen for ternary classification. The objective is to classify these images into three categories: Covid, Pneumonia, and Normal CXR images. There are a total of 16,965 images selected for this analysis, with each class containing 5,655 CXR images. The data is divided into 80% training data and

20% testing data. There are 13,572 images allocated for training and 3,393 images allocated for testing.

The trained models are used to generate bone-suppressed images from the standard CXR images. A total of five datasets are created for this experiment. Four sets of bone-suppressed images are generated using trained models, while one dataset does not feature bone suppression. Classification pipelines are used to train each dataset. The dataset is trained using the MobilenetV2 model on the Google Colab platform. The model is trained using the following hyperparameters: an Adam optimizer with a learning rate of 0.01, and trained for 25 epochs. The confusion matrices for each dataset are displayed in Fig. 5. The bone-suppressed standard CXR images generated by Pix2Pix, CycleGAN, CGIL, and CGILAL has classification accuracies of 94.22, 94.49, 95.14, and 95.76, respectively. The accuracy of classifying CXR images without bone suppression is 90.07.

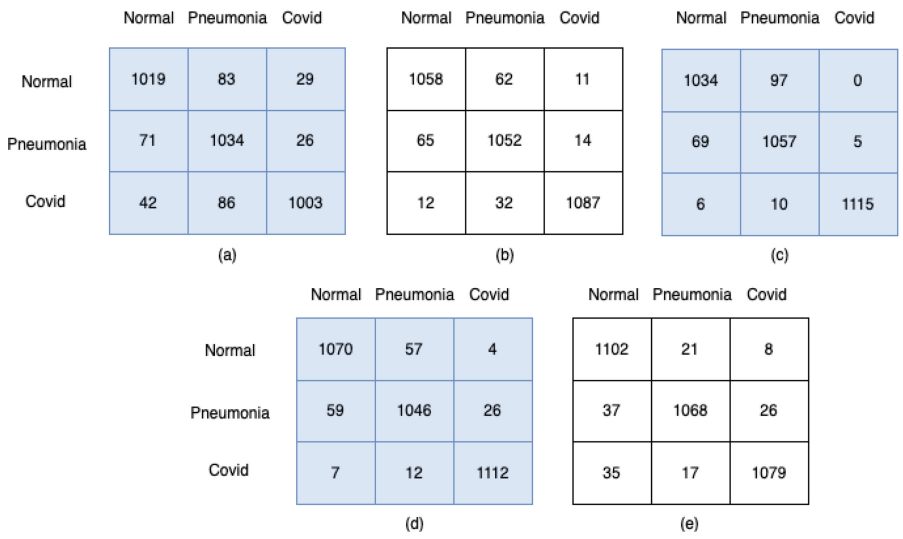


Fig. 5. Confusion Matrices for datasets trained on classification pipeline (a) Standard CXR images without bone suppression (b) Pix2Pix bone suppressed CXR images (c) CycleGAN bone suppressed CXR images (d) CGIL bone suppressed CXR images (e) CGILAL bone suppressed CXR images

5 Conclusion

Translating standard CXR images into bone-suppressed CXR images is not possible. Dual-energy CXR images are used to generate bone-suppressed images from standard CXR images. This study uses Pix2Pix, CycleGAN, and two enhanced versions of CycleGAN (CGIL and CGILAL) to accomplish bone suppression in standard CXR images. The models are trained using dual-energy CXR images and then used to suppress the bones in standard chest X-rays.

The bone-suppressed images created are used for multi-class classification. The results indicate that the classification accuracy of CXR images with bone suppression is superior to that of CXR images without bone suppression. Moreover, the recently suggested models exhibited superior performance compared to the existing models.

References

1. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. Commun. ACM 63, 11 (November 2020), 139–144. https://doi.org/10.1145/3422622
2. J. -Y. Zhu, T. Park, P. Isola and A. A. Efros, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 2242-2251, https://doi.org/10.1109/ICCV.2017.244
3. Isola, Phillip, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. "Image-to-image translation with conditional adversarial networks." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1125-1134. 2017
4. Alotaibi, A.: Deep Generative Adversarial Networks for Image-to-Image Translation: A Review. Symmetry **12**, 1705 (2020). https://doi.org/10.3390/sym12101705
5. Chlap, P., Min, H., Vandenberg, N., Dowling, J., Holloway, L., Haworth, A.: A review of medical image data augmentation techniques for deep learning applications. J. Med. Imaging Radiat. Oncol. **65**, 545–563 (2021). https://doi.org/10.1111/1754-9485.13261
6. Biswas, B., Ghosh, S.K., Ghosh, A. (2020). DVAE: Deep Variational Auto-Encoders for Denoising Retinal Fundus Image. In: Bhattacharyya, S., Konar, D., Platos, J., Kar, C., Sharma, K. (eds) Hybrid Machine Intelligence for Medical Image Analysis. Studies in Computational Intelligence, vol 841. Springer, Singapore. https://doi.org/10.1007/978-981-13-8930-6_10
7. Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784
8. Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556
9. Russakovsky, O., Deng, J., Su, H., et al.: ImageNet Large Scale Visual Recognition Challenge. Int. J. Comput. Vis. **115**, 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
10. Glorot, X., & Bengio, Y. (2010, March). Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the thirteenth international conference on artificial intelligence and statistics (pp. 249-256). JMLR Workshop and Conference Proceedings
11. Rajaraman, S., Cohen, G., Spear, L., Folio, L., Antani, S.: DeBoNet: A deep bone suppression model ensemble to improve disease detection in chest radiographs. PLoS ONE **17**(3), e0265691 (2022). https://doi.org/10.1371/journal.pone.0265691
12. Rani, G., Misra, A., Dhaka, V.S., Buddhi, D., Sharma, R.K., Zumpano, E., Eugenio Vocaturo, A., multi-modal bone suppression, lung segmentation, and classification approach for accurate COVID-19 detection using chest radiographs, Intelligent Systems with Applications, Volume 16,: 200148. ISSN **2667–3053**,(2022). https://doi.org/10.1016/j.iswa.2022.200148

13. Jia Liang, Yu-Xing Tang, You-Bao Tang, Jing Xiao, and Ronald M. Summers "Bone suppression on chest radiographs with adversarial learning", Proc. SPIE 11314, Medical Imaging 2020: Computer-Aided Diagnosis, 1131409 (16 March 2020); https://doi.org/10.1117/12.2550868
14. Mochizuki, Z., Saito, M., Suzuki, T., Mochizuki, K., Hasegawa, J., Nemoto, H., Satani, K., Takahashi, H., Onishi, H.: Cycle-generative adversarial network-based bone suppression imaging for highly accurate markerless motion tracking of lung tumors for cyberknife irradiation therapy. J. Appl. Clin. Med. Phys. **25**(1), e14212 (2024)
15. Han, L., Lyu, Y., Peng, C., Zhou, S.K.: GAN-based disentanglement learning for chest X-ray rib suppression. Med. Image Anal. **77**, 102369 (2022)
16. Kida, S., Kaji, S., Nawa, K., Imae, T., Nakamoto, T., Ozaki, S., Nakagawa, K.: Visual enhancement of cone-beam CT by use of CycleGAN. Med. Phys. **47**(3), 998–1010 (2020)
17. Yan, Y., Fujii, F., Shiinoki, T., Liu, S.: Markerless Lung Tumor Localization From Intraoperative Stereo Color Fluoroscopic Images for Radiotherapy. IEEE Access **12**, 40809–40826 (2024). https://doi.org/10.1109/ACCESS.2024.3376744
18. Xiaotong Xu, Qian Li, Shuang Jin, Zhe Su, Yu Zhang, Chest tomosynthesis image enhancement by bone suppression using convolutional neural networks with synthetic data, Journal of Radiation Research and Applied Sciences, Volume 17, Issue 2, 2024, 100866, ISSN 1687-8507, https://doi.org/10.1016/j.jrras.2024.100866
19. K. Sato, T. Ishii and Y. Saijo, "The Effectiveness of the Bone Suppression Algorithm and Concentration Conversion in the Detection of Lung Tumors," 2023 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), Nadi, Fiji, 2023, pp. 1-4, https://doi.org/10.1109/CSDE59766.2023.10487738
20. J. Xie, "Multi-Task Medical Image-to-Images Translation using Transformer for Chest X-Ray Radiography," 2021 2nd International Conference on Artificial Intelligence and Computer Engineering (ICAICE), Hangzhou, China, 2021, pp. 708-715, https://doi.org/10.1109/ICAICE54393.2021.00139
21. M. Gusarev, R. Kuleev, A. Khan, A. Ramirez Rivera and A. M. Khattak, "Deep learning models for bone suppression in chest radiographs," 2017 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), Manchester, UK, 2017, pp. 1-7, https://doi.org/10.1109/CIBCB.2017.8058543
22. Wang, L., Lin, Z.Q., Wong, A.: COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. Sci. Rep. **10**, 19549 (2020). https://doi.org/10.1038/s41598-020-76550-z
23. Chow, L.S., Rajagopal, H.: Modified-BRISQUE as no reference image quality assessment for structural MR images. Magn. Reson. Imaging **43**, 74–87 (2017)
24. Manoucher Teymouri, Samaneh Mollazadeh, Hamed Mortazavi, Zari Naderi Ghale-noie, Vahideh Keyvani, Farzaneh Aghababaei, Michael R Hamblin, Ghasem Abbaszadeh-Goudarzi, Hossein Pourghadamyari, Seyed Mohammad Reza Hashemian, Hamed Mirzaei, Recent advances and challenges of RT-PCR tests for the diagnosis of COVID-19, Pathology - Research and Practice, Volume 221: 153443. ISSN **0344–0338**,(2021). https://doi.org/10.1016/j.prp.2021.153443
25. Reza, A.M.: Realization of the contrast limited adaptive histogram equalization (CLAHE) for real-time image enhancement. Journal of VLSI signal processing systems for signal, image and video technology **38**, 35–44 (2004)

BCN: Batch Channel Normalization for Image Classification

Afifa Khaled(✉)

School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China
I202022059@hust.edu.cn

Abstract. Normalization techniques have been extensively utilized in deep learning due to their ability to enable higher learning rates and reduce the sensitivity to initialization. However, the effectiveness of commonly used normalization methods is often constrained to specific domains. In contrast to the standard Batch Normalization (BN) and Layer Normalization (LN), where BN computes the mean and variance across the (N, H, W) dimensions and LN computes them across the (C, H, W) dimensions (where N, C, H, and W represent the batch, channel, spatial height, and width dimensions, respectively), this paper introduces a novel normalization technique called Batch Channel Normalization (BCN). BCN is designed to leverage both channel and batch dependencies, thereby combining the advantages of BN and LN in an adaptive manner, depending on the dataset or task at hand. BCN normalizes inputs independently along the (N, H, W) and (C, H, W) axes, subsequently combining the normalized outputs based on adaptive parameters. As a fundamental building block, BCN can be seamlessly integrated into existing models for a wide range of applications in computer vision. Empirical results demonstrate that the proposed technique can be effectively applied to various versions of Convolutional Neural Networks (CNNs) or Vision Transformer architectures. The code is publicly available at https://github.com/AfifaKhaled/Batch-Channel-Normalization.

Keywords: Layer Normalization · Batch Normalization · Vision Transformer

1 Introduction

In recent decades, machine learning (ML) has become one of the most widely used techniques in the field of artificial intelligence. More recently, deep learning (DL) has emerged as a prevalent topic, with deep neural networks (DNNs) being extensively applied across various domains, including natural language processing, computer vision, and graph mining. Typically, DNNs consist of stacked layers with learnable parameters and non-linear activation functions. While the deep and complex structure of DNNs enables them to learn intricate features,

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15311, pp. 295–308, 2025.
https://doi.org/10.1007/978-3-031-78195-7_20

it also introduces challenges during training due to the randomness in parameter initialization and variations in input data, a phenomenon known as internal covariate shift [11]. This problem becomes more pronounced in deeper networks, where slight modifications in deeper hidden layers are amplified as they propagate through the network, leading to significant shifts in these layers.

To address the aforementioned issue, several normalization methods have been introduced. Specifically, Batch Normalization (BN)[11], Layer Normalization (LN)[3], and Group Normalization (GN) [20] have achieved remarkable success in deep learning models. Among these, BN is the most widely used for deep neural networks. Despite their success in many applications, popular normalization methods still have certain limitations. For example, Batch Normalization (BN) requires large batch sizes [20], making it unsuitable for online learning tasks and large distributed models where mini-batches must be small. To address these issues, Layer Normalization (LN) was proposed to avoid relying on the batch dimension, thereby eliminating restrictions on mini-batch size [3]. However, LN does not perform as effectively as BN on convolutional layers.

To overcome the limitations of Batch Normalization (BN) and Layer Normalization (LN), as well as to fully leverage the advantages of both techniques, we have developed a new normalization method called Batch Channel Normalization (BCN). Unlike previous techniques, BCN aims to normalize along the (C, N, H, W) axes. However, directly computing the mean and variance along the (N, C, H, W) dimensions overlooks the differing significance between the batch dimension and the channel dimension. Consequently, as illustrated in Fig 1, BCN first computes the mean μ_1 and variance σ_1^2 of the layer inputs along the (C, H, W) axes. It then computes the mean μ_2 and variance σ_2^2 along the (N, H, W) axes. Finally, the normalized outputs are combined based on adaptive parameters. To evaluate the effectiveness of the proposed method, we apply BCN to several popular models, including ResNet [7], DenseNet [9], Vision Transformer [5], and BYOL [6], on the image classification task. Our experiments demonstrate that BCN produces promising results, leading to improved training speed and enhanced generalization performance. Our main contributions can be summarized as follows:

- We introduce a new normalization technique, termed Batch Channel Normalization (BCN), as a straightforward alternative to BN and LN techniques.
- BCN leverages the dependencies between channels and batches, adaptively combining information from both the channel and batch dimensions, thereby capturing the advantages of both BN and LN.
- Empirical results demonstrate that BCN can significantly improve the generalization performance of neural networks compared to existing normalization techniques.

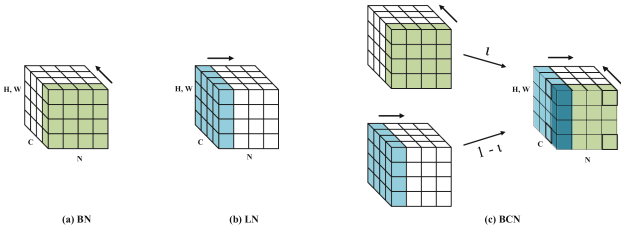


Fig. 1. Visualization of various normalization techniques. Each subplot displays a feature map tensor, where N represents the batch axis, C represents the channel axis, and (H, W) represent the spatial height and width axes, respectively.

2 Related Work

2.1 Dimension Normalization

The first group of methods involves normalizing different dimensions of the output. Examples include Layer Normalization [3], which normalizes inputs across features; Instance Normalization [17], which normalizes over spatial locations in the output; and Group Normalization [20], which independently normalizes along spatial dimensions and feature groups. Here, we introduce two techniques that are most relevant to our work, namely, Batch Normalization [11] and Layer Normalization [3].

Batch Normalization (BN) facilitates faster convergence and stabilizes the learning process. During training, BN computes the mean μ_B and variance σ_B^2 of the layer inputs as follows:

$$\mu_B = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{1}$$

$$\sigma_B^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_B)^2, \tag{2}$$

$$\bar{x}_B = \gamma\frac{(x_i - \mu_B)}{\sqrt{(\sigma_B^2 + \epsilon)}} + \beta. \tag{3}$$

We can observe that BN computes the mean μ_B and variance σ_B^2 along the (N, H, W) axes[11]. During testing, BN uses the μ_B and σ_B^2 values computed by the exponential moving average during training:

$$\mu = \alpha\mu + (1 - \alpha)\mu_B, \tag{4}$$

$$\bar{x} = \gamma\frac{(x_i - \mu)}{\sqrt{(\sigma^2 + \epsilon)}} + \beta, \tag{5}$$

where n is the batch size, and γ and β are learnable parameters. Here, α is typically set to 0.9, and ϵ is a small constant.

Layer Normalization (LN) computes the mean μ_L and variance σ_L^2 along the (C, H, W) axes as follows:

$$\mu_L = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{6}$$

$$\sigma_L^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2, \tag{7}$$

$$\bar{x}_L = \gamma\frac{(x_i - \mu_L)}{\sqrt{(\sigma_L^2 + \epsilon)}} + \beta. \tag{8}$$

Unlike BN, LN performs the same computations during both training and inference. Additionally, LN is particularly effective in stabilizing the dynamics of hidden states in recurrent neural networks.

Our method belongs to this group. The proposed Batch Channel Normalization (BCN) aims to normalize along the (C, N, H, W) axes, combining the benefits of both BN and LN while mitigating their respective deficiencies.

2.2 Normalization Improvement

The second group modifies the original Batch Normalization method [11]. This group includes methods such as Ghost Batch Normalization (Ghost BN) [8], which normalizes independently across different splits of batches, and Batch Renormalization [10] and Streaming Normalization [12], both of which adjust the approach to utilize global averaged statistics instead of batch statistics.

While these normalization techniques have gained practical popularity and success, significant advancements in them have only begun to emerge recently. Meanwhile, Batch Normalization (BN) [11] remains the most widely used normalization technique to date. Moreover, these alternative normalization methods have not yet matched BN's accuracy in many tasks, such as segmentation, detection, and video classification.

2.3 Weight Normalization

The third group consists of methods that normalize weights rather than activations. This group includes Weight Normalization [16] and Normalization Propagation [2], both of which divide weights by their ℓ_2 norm, differing only in minor details.

Recently, several papers have proposed techniques to enhance weight normalization across a wider range of CNNs. One such approach involves implicit regularization and convergence for weight normalization [19]. This study explored the weight normalization technique and reparametrized projected gradient descent for over-parameterized least squares regression. The authors demonstrated that the non-convex formulation exhibits beneficial regularization effects compared to gradient descent on the original objective.

3 Methodology

The motivation behind the success of the normalization techniques has been an important research topic. In this section, we investigate the motivation for developing the new normalization technique.

3.1 Method Formulation

The concept of normalizing along the (N, H, W) axes and (C, H, W) axes has been explored in previous works [3,11,17,20]. Earlier approaches typically perform normalization along either the (N, H, W) axes or the (C, H, W) axes independently. Our objective, however, is to normalize along the (N, C, H, W) axes. Directly computing the mean and variance along the (N, C, H, W) dimensions fails to account for the differing significance of the batch and channel dimensions. Therefore, we propose to normalize separately along the (N, H, W) and (C, H, W) axes, and then combine the normalized outputs based on adaptive parameters ι. This approach could enhance training, validation, and test accuracy, as we demonstrate experimentally in the next section.

In a manner similar to how Batch Normalization (BN) normalizes layer inputs, during training, Batch Channel Normalization (BCN) first computes the mean μ_1 and variance σ_1^2 of the layer inputs along the (N, H, W) axes:

$$\mu_1 = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{9}$$

$$\sigma_1^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_1)^2. \tag{10}$$

Second, BCN computes the mean μ_2 and variance σ_2^2 along the (C, H, W) axes:

$$\mu_2 = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{11}$$

$$\sigma_2^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_2)^2. \tag{12}$$

Next, $\bar{x}_1$ and $\bar{x}_2$ are normalized using μ_1, σ_1^2 and μ_2, σ_2^2, respectively:

$$\bar{x}_1 = \frac{(x_i - \mu_1)}{\sqrt{(\sigma_1^2 + \epsilon)}}, \tag{13}$$

$$\bar{x}_2 = \frac{(x_i - \mu_2)}{\sqrt{(\sigma_2^2 + \epsilon)}}. \tag{14}$$

BCN introduces an additional learnable parameter ι to adaptively balance the normalized outputs along the (N, H, W) and (C, H, W) axes.

$$\bar{y} = \iota\bar{x}_1 + (1 - \iota)\bar{x}_2, \tag{15}$$

The output of BCN normalization can then be formulated as follows:

$$Y = \gamma \bar{y} + \beta, \tag{16}$$

where γ and β are learnable parameters, and ϵ is a small constant for numerical stability.

At the inference stage, since μ and σ are pre-computed and fixed, the normalization can be fused into the convolution operation.

Following previous works [4,13], BCN normalizes along the (N, H, W) axes using an exponential moving average [14] during training, as follows:

$$\mu = \alpha\mu + (1 - \alpha)\mu_1, \tag{17}$$

$$\sigma^2 = \alpha\sigma^2 + (1 - \alpha)\sigma_1^2, \tag{18}$$

$$\bar{x} = \frac{(x_i - \mu)}{\sqrt{(\sigma^2 + \epsilon)}}, \tag{19}$$

where α is set to 0.9 in our experiments.

The key difference between BCN and existing normalization techniques is that, under BCN, all channels in a layer share the same normalization terms, μ and σ^2.

3.2 Implementation

BCN can be implemented with just a few lines of Python code in PyTorch [15] or TensorFlow [1], where the computation of $\bar{x}_1$ along the (N, H, W) axes and $\bar{x}_2$ along the (C, H, W) axes is performed. The overall BCN process is outlined in Algorithm 1, with the corresponding Python code shown in Fig 2.

3.3 When Should the Output Resemble BN? When Should It Resemble LN?

If the normalization is along the (N, H, W) axes, BCN closely resembles BN. Conversely, if the normalization is along the (C, H, W) axes, BCN more closely resembles LN. This straightforward combination of two existing normalization techniques allows BCN to directly balance the normalized outputs, requiring only a few additional parameters.

4 Experiments and Discussion

Algorithm 1 Batch Channel Normalization (BCN)

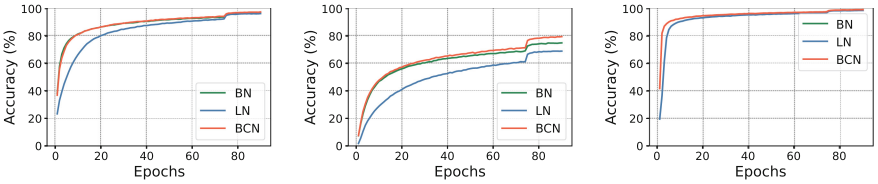
Require:
Input $x = \{x_1, x_2, ..., x_n\}$, Parameters to be learned: ι, β and γ
Ensure:
$Y = BCN_{\gamma,\beta,\iota}(x_i)$
1: Calculate μ_1 and σ_1^2 based on Eq. 9 and 10
2: Calculate μ_2 and σ_2^2 based on Eq. 11 and 12
3: Calculate the normalized output $\bar{x}_1$ along (N, H, W) and $\bar{x}_2$ along (C, H, W) axes by Eq. 13 and 14
4: Adaptively combine $\bar{x}_1$ and $\bar{x}_2$ based on Eq. 15
5: Calculate the final output Y based on Eq. 16
6: **return** Y

```
def BatchChannelNorm(x, gamma, beta, momentum=0.9, num_channels, eps=1e-5):
    self.num_channels = num_channels
    self.epsilon = epsilon
    self.x1 = BCN_1(self.num_channels, epsilon=self.epsilon) # normalized along (N, H,
        W) axes.
    self.x2 = BCN_2(self.num_channels, epsilon=self.epsilon) # normalized along (C, H,
        W) axes.
    # x: input features with shape [N,C,H,W]
    # gamma, beta: scale and offset
    self.gamma = nn.Parameter(torch.ones(num_channels))
    self.beta = nn.Parameter(torch.zeros(num_channels))
    # iota is the BCN variable to be learned to adaptively balance the normalized
        outputs along (N, H, W) axes and (C, H, W) axes.
    self.iota = nn.Parameter(torch.ones(self.num_channels))
    X = self.x1(x)
    Y = self.x2(x)
    Result = self.iota.view([1, self.num_channels, 1, 1]) * X + ( 1 - self.iota.view
        ([1, self.num_channels, 1, 1])) * Y
    Result = self.gamma.view([1, self.num_channels, 1, 1]) * Result + self.beta.view
        ([1, self.num_channels, 1, 1])
    return Result
```

Fig. 2. Python code of Batch Channel normalization (BCN) based on PyTorch.

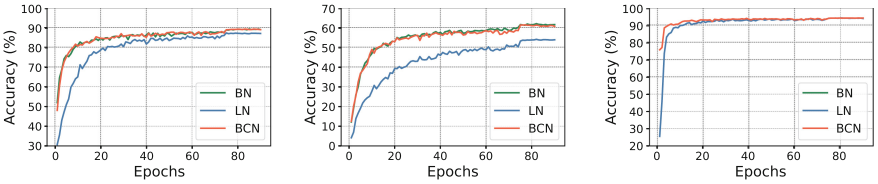
4.1 Comparison with Normalization Techniques

In this subsection, we compare our method with conventional normalization techniques using popular datasets and neural networks. Specifically, we evaluate the image classification performance of ResNet with different normalization techniques on the CIFAR-10/100 and SVHN datasets. Additionally, we compare the performance of various normalization techniques on DenseNet using the ImageNet dataset and on self-supervised learning with BYOL using the CIFAR-10



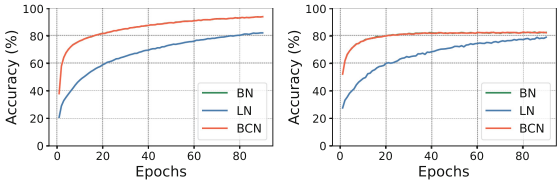
(a) CIFAR-10 (b) CIFAR-100 (c) SVHN

Fig. 3. Training accuracy of ResNet with different normalization techniques on (a) CIFAR-10, (b) CIFAR-100, (c) SVHN.



(a) CIFAR-10 (b) CIFAR-100 (c) SVHN

Fig. 4. Validation accuracy of ResNet with different normalization techniques on (a) CIFAR-10, (b) CIFAR-100, (c) SVHN.



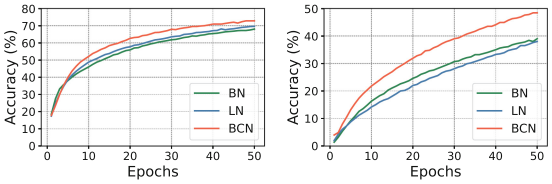
(a) Train (b) Validation

Fig. 5. Training and validation accuracy curves for different normalization techniques in BYOL on the CIFAR-10 dataset. Note that the results for BN are identical to those for BCN.

dataset. Furthermore, we explore the application of BCN on new models, such as Vision Transformers, using the CIFAR-10/100 and SVHN datasets.

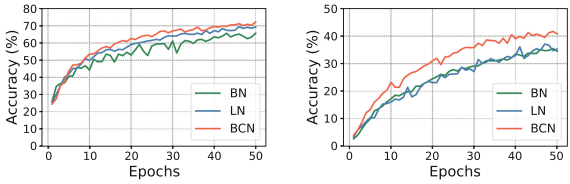
Results on ResNet We conduct experiments on ResNet [7] for the image classification task. The model is trained using stochastic gradient descent (SGD), starting with a learning rate of 0.1, which is reduced by a factor of 10 at the 75th and 85th epochs, respectively. A batch size of 8 and a momentum of 0.9 are employed.

We present the results of BCN, BN, and LN during training and validation on the three datasets in Fig 3 and Fig 4. As shown, BCN demonstrates the fastest learning rate. On CIFAR-10, within approximately 20 epochs, it achieves



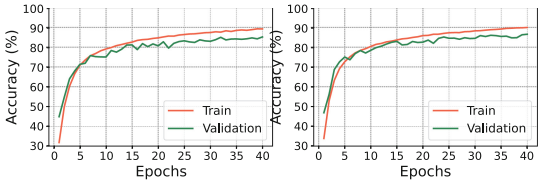
(a) CIFAR-10 (b) CIFAR-100

Fig. 6. Training accuracy of ViT with different normalization techniques on (a) CIFAR-10 and (b) CIFAR-100.



(a) CIFAR-10 (b) CIFAR-100

Fig. 7. Testing accuracy of ViT with different normalization techniques on (a) CIFAR-10, (b) CIFAR-100.



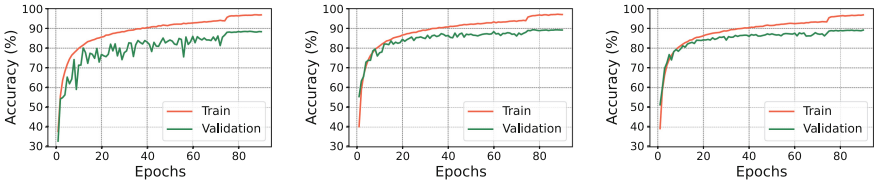
(a) 2-batch size (b) 4-batch size

Fig. 8. The accuracy curves for mini-batch size on the CIFAR-10 dataset.

about 86.12% training accuracy and 84.16% validation accuracy. In comparison, BN and LN reach 86.04% and 79.74% training accuracy and 82.87% and 78.58% validation accuracy, respectively, within the same number of epochs.

Additionally, Table 1 presents the test accuracy results for BCN and other representative normalization techniques (BN, LN, IN, and GN) on the CIFAR-10, CIFAR-100, and SVHN datasets. All experiments were conducted under the same conditions, including learning rate, loss function, batch size, etc. The results indicate that BCN is generally effective, often achieving the best or second-best performance. For instance, on the CIFAR-100 dataset, BCN shows significant improvement over state-of-the-art techniques.

Results on BYOL We apply BCN to a recent state-of-the-art method, BYOL [6], for self-supervised learning. We have implemented BYOL in PyTorch,



(a) 128-batch size (b) 16-batch size (c) 8-batch size

Fig. 9. The training and validation accuracy curves for different batch sizes on the CIFAR-10 dataset.

Table 1. Comparison of test accuracy across three datasets. The best results are highlighted in **bold**, and the second-best results are underlined.

Method	CIFAR-10	CIFAR-100	SVHN
BN	96.11	<u>74.50</u>	98.22
LN	95.76	68.61	97.62
IN	<u>96.68</u>	73.42	**98.93**
GN	95.91	70.15	98.49
BCN	**96.97**	**79.09**	<u>98.63</u>

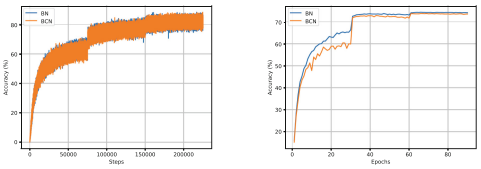
using the same hyperparameter settings as in the original paper [6]. BCN is applied to both the online and target models. As shown in Fig 5, applying BCN improves the performance of BYOL.

Results on ViT There is growing interest in developing Vision Transformer (ViT) methods [5] for a wider range of applications. We implemented ViT from scratch and tested it with different batch sizes and embedding dimensions on the CIFAR-10/100 and SVHN datasets. As shown in Fig 6, the performance of ViT improved when the normalization technique was replaced by BCN. Specifically, replacing the existing normalization with BCN led to an increase of approximately 0.73% in training accuracy and 0.72% in testing accuracy.

Furthermore, Fig 7 illustrates the performance of BCN on the testing set. Overall, these results support our hypothesis that BCN can outperform existing normalization techniques like BN and LN, particularly in new models such as ViT.

4.2 Results on DenseNet

DenseNet is a well-known Dense Convolutional Network. We implemented DenseNet-201 in PyTorch. Careful selection of the learning rate can lead to improved performance. To this end, we experimented with different learning rates to determine the most suitable values for our datasets and topology. In this study, we initially set the learning rate to 3×10^{-2}. Similarly, we conducted



(a) Training accuracy (b) Testing accuracy

Fig. 10. Training and testing accuracies of DenseNet on the ImageNet dataset.

experiments to identify the optimal batch size, ultimately using a batch size of 512. The maximum batch size was used to fully utilize GPU memory during training. In our experiments, we trained for 90 epochs, with each epoch consisting of 2503 iterations. Additionally, we recorded training accuracy by steps and testing accuracy by epochs. We compared the training, testing, and validation performance of different normalization techniques by combining DenseNet-201 with BN and BCN on the ImageNet dataset. As illustrated in Fig 10 (a) and Fig 10 (b), the proposed BCN demonstrated strong performance.

4.3 Experiments on Brain Segmentation

We conducted experiments on brain segmentation datasets, specifically the *MICCAI iSEG* dataset [18] and the *MRBrainS* dataset[1]. Table 2 presents the accuracy metrics of the proposed model for the 13 subjects in the test set. As shown in Table 2, the Dice Coefficient (DC) values with BCN indicate strong performance.

Table 2. Segmentation performance (Dice Coefficient, *DC*) on the *MICCAI iSEG* dataset. The best performance for each tissue class is highlighted in **bold**.

Model	Dice Coefficient (DC)		
	CSF	GM	WM
BN	0.95	0.92	0.90
IN	0.93	0.90	0.90
GN	0.90	0.90	0.90
LN	0.90	0.90	0.90
BCN	**0.96**	**0.94**	**0.92**

[1] https://mrbrains13.isi.uu.nl/

4.4 Ablation Study

In this subsection, we explore the impact of batch size. We evaluate various batch sizes: 128, 16, and 8. Our findings are presented in Fig 9, indicating that BCN yields favorable results across different batch sizes. To examine whether BCN mitigates the weaknesses of BN and LN, particularly as noted by [11], where BN performs poorly with small batch sizes, we focus on addressing the minibatch problem of BN in this subsection. The experiments demonstrate that BCN alleviates the performance issues of BN with small batch sizes. Specifically, BCN achieves strong performance with batch sizes of 4 and 2, as shown in Fig 8. For example, with a batch size of 8 and after 20 epochs, BCN reaches approximately 86.12% training accuracy and 84.16% validation accuracy, whereas with a batch size of 2 over the same number of epochs, BCN achieves about 84.58% training accuracy and 81.42% validation accuracy. These results confirm that the proposed technique is effective even with small batch sizes.

4.5 Analysis Optimization Landscape

In this section, we study and analyze the optimization landscape. Our starting point is identifying the key impact that BCN has on the training process. BCN reparametrizes the underlying optimization problem, making its landscape significantly smoother. The first aspect of this impact is the enhancement in the Lipschitz continuity of the loss function. Recall that a function f is L-Lipschitz if $|f(x_1) - f(x_2)| \leq L||x_1 - x_2||$ for all x_1 and x_2. This means that the loss changes at a slower rate, and the magnitudes of the gradients are smaller as well. Additionally, the loss exhibits significantly better "effective" β-smoothness. Recall that a function f is β-smooth if its gradient is β-Lipschitz. It is important to note that, due to the presence of non-linearities, one should not expect β-smoothness to be bounded in an absolute, global sense. At a particular training step, we measure the variation (shaded region) in loss and the "effective" β-smoothness, which refers to the maximum difference (in ℓ_2-norm) in gradient over the distance moved in that direction.

5 Conclusion

In this paper, we proposed a new normalization technique termed Batch Channel Normalization (BCN). BCN simultaneously exploits both the channel and batch dimensions and adaptively combines the normalized outputs. Our experiments on various models and datasets demonstrate that BCN consistently outperforms state-of-the-art normalization techniques, indicating that BCN is a versatile and generalizable normalization method. As future work, an ablation study could be conducted to directly compute the mean and variance along the (N, C, H, W) axes. Additionally, we plan to investigate the applicability of the BCN technique across a broader range of applications and evaluate its effectiveness across a wider variety of CNN architectures.

Acknowledgements. We would like to thank Sergey Ioffe for helpful discussions.

References

1. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al.: TensorFlow: A system for large-scale machine learning. In: 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16). pp. 265–283 (2016)
2. Arpit, D., Zhou, Y., Kota, B., Govindaraju, V.: Normalization propagation: A parametric technique for removing internal covariate shift in deep networks. In: Proceedings of The 33rd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 48, pp. 1168–1176. New York, New York, USA (20–22 Jun 2016)
3. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. Advances in NIPS Deep Learning Symposium (2016)
4. Cai, Z., Ravichandran, A., Maji, S., Fowlkes, C.C., Tu, Z., Soatto, S.: Exponential moving average normalization for self-supervised and semi-supervised learning. CoRR **abs/2101.08482** (2021), https://arxiv.org/abs/2101.08482
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. CoRR **abs/2010.11929** (2020), https://arxiv.org/abs/2010.11929
6. Grill, J., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.Á., Guo, Z.D., Azar, M.G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M.: Bootstrap your own latent: A new approach to self-supervised learning. CoRR **abs/2006.07733** (2020), https://arxiv.org/abs/2006.07733
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
8. Hoffer, E., Hubara, I., Soudry, D.: Train longer, generalize better: Closing the generalization gap in large batch training of neural networks. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. pp. 1731–1741 (2017)
9. Huang, G., Liu, Z., Weinberger, K.Q.: Densely connected convolutional networks. CoRR **abs/1608.06993** (2016), http://arxiv.org/abs/1608.06993
10. Ioffe, S.: Batch renormalization: Towards reducing minibatch dependence in batch-normalized models. In: Advances in Neural Information Processing Systems. vol. 30 (2017)
11. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning. pp. 448–456. pmlr (2015)
12. Liao, Q., Kawaguchi, K., Poggio, T.A.: Streaming normalization: Towards simpler and more biologically-plausible normalizations for online and recurrent learning. CoRR **abs/1610.06160** (2016), http://arxiv.org/abs/1610.06160
13. Luo, C., Zhan, J., Wang, L., Gao, W.: Extended batch normalization. CoRR **abs/2003.05569** (2020), https://arxiv.org/abs/2003.05569
14. Mukhoti, J., Dokania, P.K., Torr, P.H.S., Gal, Y.: On batch normalisation for approximate bayesian inference. CoRR **abs/2012.13220** (2020), https://arxiv.org/abs/2012.13220
15. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)

16. Salimans, T., Kingma, D.P.: Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In: Lee, D.D., Sugiyama, M., von Luxburg, U., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain. p. 901 (2016)
17. Ulyanov, D., Vedaldi, A., Lempitsky, V.S.: Instance normalization: The missing ingredient for fast stylization. CoRR **abs/1607.08022** (2016), http://arxiv.org/abs/1607.08022
18. Wang, L., Nie, D., Li, G., Puybareau, É., Dolz, J., Zhang, Q., Wang, F., Xia, J., Wu, Z., Chen, J.W., et al.: Benchmark on automatic six-month-old infant brain segmentation algorithms: the iseg-2017 challenge. IEEE Trans. Med. Imaging **38**(9), 2219–2230 (2019)
19. Wu, X., Dobriban, E., Ren, T., Wu, S., Li, Z., Gunasekar, S., Ward, R., Liu, Q.: Implicit regularization of normalization methods. CoRR **abs/1911.07956** (2019), http://arxiv.org/abs/1911.07956
20. Wu, Y., He, K.: Group normalization. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 3–19 (2018)

EEG-Based Mental Imagery Task Adaptation via Ensemble of Weight-Decomposed Low-Rank Adapters

Taveena Lotey[1(✉)], Aman Verma[2], and Partha Pratim Roy[1]

[1] Department of CSE, Indian Institute of Technology Roorkee, Roorkee, India
taveena@cs.iitr.ac.in , partha@cs.iitr.ac.in
[2] Department of ME, Indian Institute of Technology Roorkee, Roorkee, India
averma1@me.iitr.ac.in

Abstract. Electroencephalography (EEG) is widely researched for neural decoding in Brain Computer Interfaces (BCIs) as it is non-invasive, portable, and economical. However, EEG signals suffer from inter- and intra-subject variability, leading to poor performance. Recent technological advancements have led to deep learning (DL) models that have achieved high performance in various fields. However, such large models are compute- and resource-intensive and are a bottleneck for real-time neural decoding. Data distribution shift can be handled with the help of domain adaptation techniques of transfer learning (fine-tuning) and adversarial training that requires model parameter updates according to the target domain. One such recent technique is Parameter-efficient fine-tuning (PEFT), which requires only a small fraction of the total trainable parameters compared to fine-tuning the whole model. Therefore, we explored PEFT methods for adapting EEG-based mental imagery tasks. We considered two mental imagery tasks: speech imagery and motor imagery, as both of these tasks are instrumental in post-stroke neurorehabilitation. We proposed a novel ensemble of weight-decomposed low-rank adaptation methods, EDoRA, for parameter-efficient mental imagery task adaptation through EEG signal classification. The performance of the proposed PEFT method is validated on two publicly available datasets, one speech imagery, and the other motor imagery dataset. In extensive experiments and analysis, the proposed method has performed better than full fine-tune and state-of-the-art PEFT methods for mental imagery EEG classification.

Keywords: Electroencephalography · Deep Learning · Transfer Learning · Fine-tuning · Low-Rank Adaptation.

1 Introduction

Brain-computer interfaces (BCIs) represent a developing field of research aimed at establishing direct communication pathways between the brain and external devices. The physiological measure often used for decoding neural signals

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15311, pp. 309–324, 2025.
https://doi.org/10.1007/978-3-031-78195-7_21

is electroencephalography (EEG), as it is non-invasive, portable, and economical. There are various experimental BCI paradigms to control external devices through robotic arm [5], speller system [32], and exoskeleton [6] via particular brain activity in a specific task. Among these paradigms, mental imagery has undergone extensive investigation as a mechanism for controlling BCIs, leveraging the intrinsic brain activity that arises from the voluntary imagination of users [1]. Additionally, mental imagery operates independently of external stimuli. Therefore, it supports the development of a user-friendly interface. It reduces fatigue and enhances users' awareness of their surroundings more naturally.

Deep learning (DL) has advanced rapidly in recent years. Many deep learning-based methods have been proposed for enhanced EEG classification. Deep learning-based methods such as Convolutional neural networks (CNN) and Transformers have a large number of parameters, and they require large amounts of data to learn and extract discriminative features. EEG signals have high dimensionality and labor-extensive recording procedures. These challenges result in small datasets that make DL methods prone to overfitting. Additionally, EEG signals suffer from inter- and intra-subject variability [25]. One of the possible solutions to reduce inter-subject variability is Domain adaptation (DA). It is defined as using a classifier/model learned on one task with sufficient labeled samples, and this classifier/model adapts to another related task with only a limited amount of training data.

Many works have utilized DA approaches to reduce the distribution shift in two domains of mental imagery tasks. One section of the studies tends to use domain adaptation for intra-subject/cross-session distribution and another for inter-subject/cross-subject distribution among various subjects [16,20]. Motor Imagery (MI) is the most researched modality of mental imagery tasks. Moreover, several studies have employed domain adaptation for MI-EEG signal classification. Many recent studies utilized source and target domain correlation assessment approaches to select the source/s similar to target domain [11,16,34]. CNN-based feature extractors and classifiers are extensively used in multiple studies on domain adaptation of MI-EEG task classification [18,21,33]. However, only a few studies explored transformer-based approaches for MI task adaptation [28]. The other modality of mental imagery tasks is speech imagery/imagined speech (SI), which has yet to be explored in domain adaptation.

In the case of inaccessibility of source data and availability of a pre-trained source model, the target model is initialized with parameters of a pre-trained model and subsequently fine-tuned on target domain data [12,30]. With the rapid increase in the number of parameters and depth of the deep neural networks, fine-tuning such large models is a computationally expensive task. This issue is addressed in natural language processing (NLP) and computer vision (CV) domain with parameter-efficient adaptation [8]. In parameter-efficient adaptation, new modules are added in between the pre-trained model layers, and these modules are known as adapters. One such adapter method is Low-Rank Adaptation (LoRA) [9]. While fine-tuning, only these adapters are trained instead of

training the whole pre-trained model on the target domain. Therefore, training only these adapters results in a reduction in the number of trainable parameters.

To the best of our knowledge, the impact of parameter-efficient adapters is not explored in the EEG classification task. Therefore, our proposed method is based on a high-performance weight-decomposed low-rank adaptation method [17]. Additionally, convolutional transformer-based methods have not been explored in mental imagery task classification. Inspired by this, we utilized EEG Conformer, a convolutional transformer method for feature extraction and classification [27]. In literature, numerous studies have been proposed to decode the neural activity on a specific mental imagery task [19]. Only a few studies proposed methods to decode neural activity on multiple mental imagery tasks [1]. Our method validated the performance of the proposed low-rank adaptation method on two mental imagery datasets, where parameters of a model trained on one dataset are used to fine-tune the model on the other dataset and vice-versa.

The contributions of this paper are listed as follows:

- We demonstrate the parameter efficient fine-tuning (PEFT) based approaches for enhanced mental imagery task classification. To the best of our knowledge, PEFT-based low-rank adaptation (LoRA) is first explored for mental imagery classification tasks in our work.
- We propose EDoRA, a novel ensemble of weight-decomposed low-rank adapters for mental imagery EEG classification tasks for enhanced adaptation performance.
- To the best of our knowledge, low-rank adaptation on two categories of mental imagery tasks, i.e., speech and motor imagery, is first explored in our work.
- A detailed analysis of parameter-efficient fine-tuning is performed on speech and motor imagery tasks of mental imagery EEG signal.

2 Related Work

2.1 MI Classification

Several studies have contributed in addressing the problem of inter-subject and intra-subject data distribution variability in motor imagery EEG signal classification. Most of the work in recent years employed source and target domain correlation assessment methods to select the closest source domain for better target domain classification performance [11,13,16,24,31,33,34], and convolutional neural network (CNN) based methods are utilized as classifier [3,4,7,13,18,21,28,33,34]. Only a few works have studied the impact of transfer learning-based domain adaptation approaches on latest transformer based models [10].

Several studies explored the efficiency of CNN based approaches for domain adaptation in MI-EEG classification. Zhong *et al.* proposed a domain adaptation framework based on correlation alignment of the source and target domain motor imagery EEG data [34].Then conventional CNN based classifier is used to classify the features in cross-subject settings. Hang et al. proposed a deep

domain adaptation network based on CNN with MMD to minimize the source and target distrubution distance, and then applied center-based discriminative feature learning approach to maximize the inter-class distance [7]. They jointly optimized the source and target domain data to align the features. Liu *et al.* proposed a framework for subject adaptation and it includes a CNN based feature extractor, a subject adapter based on MMD to align the source and target domains and reduce the feature distribution shifts [18].

Most of the studies have jointly optimized the source and target domain for MI-EEG classification task. Few studies used transfer learning, i.e., to use model trained on source domain data to optimize the model trained on target domain data. Phunruangsakao *et al.* proposed a deep domain adaptation framework that selects multiple source domains to optimize label classification of single target domain [24]. This work further experiments by making source parameters inaccessible and making the privacy policy stricter. Inspired by this work, Huang *et al.* proposed a multi-source free domain adaptation framework with attention weighted module for better source and target domain alignment [12]. To keep the privacy of the source domain data, our work uses the parameters of pre-trained source model to fine tune the target model.

2.2 SI Classification

There are few works exploring the classification performance of speech recognition using traditional methods [26]. However, the research area of domain adaptation of speech imagery/imagined speech tasks is less explored in the literature. Jimenez *et al.* proposed a deep unsupervised domain adaptation method based on standardization-refinement approach [14]. The research area of domain adaptation in EEG-based speech imagery is still required to be explored. With this inspiration, we chose speech imagery as one task of mental imagery EEG classification to validate the effectiveness of our proposed approach.

In literature, only a few works explored the decoding efficiency of transformer based approaches for domain adaptation in MI EEG signal classification [10, 28]. Transformer based methods are proved to be efficient than CNN based method in the domain of NLP and CV. With this inspiration, our method used a convolutional transformer based method curated for EEG signal classification with spatial feature extraction power of CNN and temporal feature extraction efficiency of transformers [27].

Also, to the best of our knowledge, mental imagery task adaptation from speech to motor imagery and vice-versa is not yet explored, where model pre-trained on one type of mental imagery task is used to fine-tune the target model of other type of mental imagery task. Also, fine-tuning requires all parameters of the pre-trained model to be trained while optimization on target domain data. Nowadays, the size of the models keep getting bigger to attain human brain level intelligence. Therefore, full fine-tuning of such large models becomes compute and space extensive. In recent literature of NLP and CV, parameter efficient fine-tuning methods such as LoRA [9], are proposed that requires only a small number of trainable parameters compared to full fine-tuning while also maintaining the

performance. Inspired by a similar work [17], we propose an ensemble of weight-decomposed low rank adaptation method for EEG-based mental imagery task classification.

3 Methods

3.1 Definitions and Notations

We denote $D^S_{d_1} = \{(X^i_{d_1}, y^i_{d_1}) \mid X^i_{d_1} \in R^{N_{c_1} \times N_{t_1}},\ y^i_{d_1} \in Y_{d_1}\}$ as the source domain, where $X^i_{d_1}$ is a pre-training EEG dataset trial in the source domain with N_{c_1} spatial channels and N_{t_1} temporal sampling points, and $Y_{d_1} = \{0 : out, 1 : in, 2 : up\}$ is the label set if d_1 is SI dataset and $Y_{d_1} = \{0$: *Left Hand*, 1: *Right Hand*, 2: *Both Feet*, 3: *Tonque*$\}$ if d_1 is MI dataset.

Similarly, the target domain is defined as $D^T_{d_2} = \{(X^i_{d_2}, y^i_{d_2}) \mid X^i_{d_2} \in R^{N_{c_2} \times N_{t_2}},\ y^i_{d_2} \in Y_{d_2}\}$, where $X^i_{d_2}$ is a fine-tuning EEG dataset trial in the target domain (MI or SI Dataset).

3.2 Low-Rank Adaptation

Low-rank adaptation is the one of the PEFT method that does not change the model architecture and gaining popularity due to its simplicity and efficiency. LoRA (Low-Rank Adaptation) strategy utilizes a simple design that brings practical benefits to dense layers in deep learning models [9]. The experimental focus of this paper was primarily on Transformer language models, however, the principles can be applied to other models.

Neural networks often incorporate dense layers that perform matrix multiplications using fully-ranked weight matrices. Inspired by the concept that pre-trained language models function within a low hidden dimension, the authors of this paper proposed that weight updates during model adaptation also exhibit a low intrinsic rank. Consider a pre-trained weight matrix W_0 in $R^{d \times k}$. The low-rank decomposition of this weight matrix is given by:

$$W_0 + \Delta W = W_0 + BA \tag{1}$$

where $B \in R^{d \times r}$, $A \in R^{r \times k}$, and the rank r satisfies $r \leq \min(d, k)$. During training, W_0 is held constant, with only A and B (which contain the trainable parameters) being updated. This yields a modified forward pass equation:

$$h = W_0 x + \Delta W x = W_0 x + BAx \tag{2}$$

If r is chosen as a very small number, the number of trainable parameter reduces significantly. LoRA approach is resource efficient and have low inference latency. LoRA can be applied to any subset of weight matrices within a neural network to minimize the number of trainable parameters.

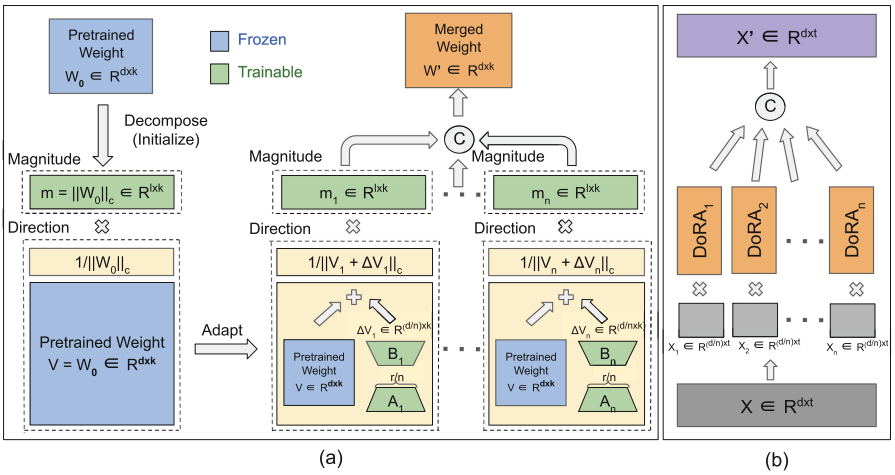


Fig. 1. (a) Overview of proposed EDoRA paramter-efficient fine-tuning approach. It depicts the overall parameter updation process of EDoRA, before and after fine-tuning (b) Feature updation via EDoRA adaptation. [© symbol represents concatenation, $\times$ symbol represents product, X represents input features, $X^{'}$ represents output features.]

3.3 Proposed method

The proposed method EDoRA is inspired by the weight-decomposed low-rank adaptation method (DoRA) [17]. This method enhances efficiency by dissecting pre-trained model weights into magnitude and directional constituents, each of which is subsequently fine-tuned for optimized performance. This decomposition allows for concentrated updates that minimize the quantity of trainable parameters while maximizing the efficacy of the learning process. The adaptation process is defined in the following steps:

Initial Decomposition: The process starts with the division of pre-trained model weights into magnitude and directional components. This separation is crucial for enabling specific updates during fine-tuning. The weight decomposition is represented as:

$$W = m\frac{V}{\|V\|_c} = \|W\|_c\frac{W}{\|W\|_c} \tag{3}$$

where W is the weight matrix, m is a vector representing the magnitude, and V is the matrix representing directional values, with $\|V\|_c$ denoting the column-wise norm [17]. The number of parameters in the directional component is more than the magnitude component; therefore, only the directional component is fine-tuned via LoRA to keep the number of trainable parameters efficient [17].

Fine-tuning of Directional Component: The directional component is fine-tuned using Low-Rank Adaptation (LoRA) [9]. This process involves updates focused primarily on parameters that offer the highest utility, thus enhancing

efficiency. The adapted weight W' after fine-tuning is given by:

$$W' = m\frac{V + \Delta V}{\|V + \Delta V\|_c} = m\frac{W_0 + BA}{\|W_0 + BA\|_c} \tag{4}$$

where V represents the weights of pre-trained model, and ΔV is the change in directional component learned by multiplying two low-rank matrices B and A [17].

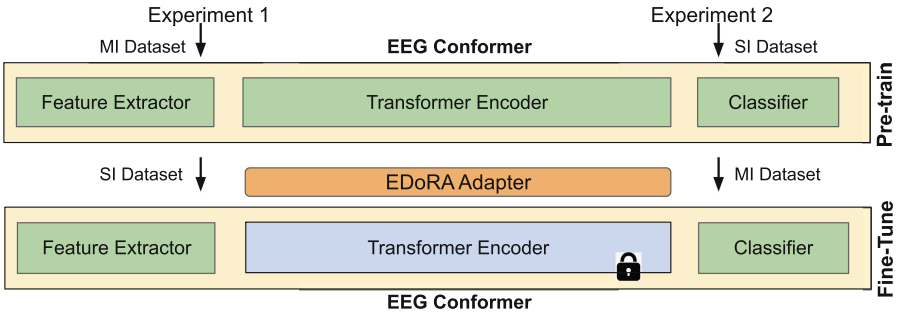


Fig. 2. Framework of the proposed method. Two experiments are performed in this work, and in these experiments EEG Conformer model is pre-trained on one dataset, and then fine-tuned on other dataset with only EDoRA adapter on each operation of transformer encoder of EEG Conformer and vice-versa. [Freezed weights are shown with lock]

Eq. 4 represents the fine-tuned weights of the DoRA method. As EEG signals are highly non-stationary, the features learned through fine-tuning individual DoRA adapters have a high variance in feature learning that may result in overfitting and local optima. Therefore, an ensemble of multiple DoRA adapters reduces the overfitting and local optima issues, which leads to stable and reliable performance [2]. Therefore, the proposed method EDoRA is an ensemble of multiple DoRA adapters. The weight updation in each individual adapter of EDoRA will follow the Eq. 4. The input feature $x \in R^{d\times t}$ is split into n equal parts that gives $x_i \in R^{d\times \frac{t}{n}}$, where $i = 0, 1, 2 \ldots n$. Each x_i is fed to $DoRA_i$, and the output weights of each $DoRA_i$ are concatenated. This process is depicted in Fig. 1 (b). The modified forward pass equation for EDoRA is defined by the following equations:

$$\begin{aligned} h &= concat_{i=0}^{n}\left(m_i \frac{Vx_i + \Delta V_i x_i}{\|Vx_i + \Delta V_i x_i\|_c}\right) \\ &= concat_{i=0}^{n}\left(m_i \frac{W_0 x_i + (B_i A_i)x_i}{\|W_0 x_i + (B_i A_i)x_i\|_c}\right) \end{aligned} \tag{5}$$

where n denotes number of equal parts of the input features x, $A_i \in R^{\frac{r}{n}\times d}$ and $B_i \in R^{k\times \frac{r}{n}}$ are the low rank matrices with rank $\frac{r}{n}$, $x_i \in R^{d\times \frac{t}{n}}$ is the

input feature matrix, $m_i \in 1 \times k$ is the magnitude vector, $W_0 \in R^{d \times k}$ are the pre-trained model weights, and $h \in R^{k \times t}$ represents the updated feature.

Detailed overview of weight updation in the proposed method is depicted in Fig 1 (a), and detailed depiction of feature updation is given in Fig. 1 (b).

3.4 Optimization Procedure

In this study, the EEG Conformer model is firstly pre-trained on d_1 EEG data in $D_{d_1}^S$ and then fine-tuned using d_2 EEG data in $D_{d_2}^T$. The decoding model can be represented as a classifier $m : R^{N_c \times N_t} \rightarrow y_{d_1} \mid y_{d_2}$ which is defined as:

$$m(X_i; \theta) = g(\phi(X_i; \theta_\phi); \phi_g) \tag{6}$$

where ϕ denotes feature extraction and transformer encoder module with parameters θ_ϕ, and g denotes classifier module with learnable parameters ϕ_g. This model learns the classification of data by minimizing the prominently used cross-entropy loss [20]. Fig. 2 illustrates the network architecture and adaptation strategy of our work.

4 Experimental Results

4.1 Datasets

Motor Imagery Dataset The BCI Competition IV 2a is a publicly available motor imagery dataset that comprises EEG data from 9 subjects performing motor imagery tasks of the left hand, right hand, both feet, and tongue [29]. The dataset features EEG signals recorded at a 250 Hz sampling rate from 22 electrodes over two sessions, each containing six runs of 48 trials per motor task, totaling 288 trials per session. Hereon, this dataset will be depicted as '*MI dataset*'.

Speech Imagery Dataset The publicly available Arizona state university (ASU) dataset is used for the classification of the speech imagery tasks [22]. In this study, we used a dataset for short word classification. The dataset consists of three class SI of the English words in, "out" and up. Each class consists of 100 trials, and a single trial lasts for 5 s. The data were acquired from six subjects with 60 EEG channels. The data was pre-processed using a frequency range of 8-70 Hz. Hereon, this dataset will be depicted as '*SI dataset*'.

Data pre-processing MI dataset is kept at 250Hz sampling rate whereas SI dataset is downsampled from 1000Hz to 250Hz. Keeping the data temporal length same, 4 second data is being used for both datasets. Further, 4th-order Butterworth filter is used to eliminate low-frequency noise in the 4-40 Hz range and z-score normalization applied.

4.2 Experimental Details

To implement the proposed and compared models, PyTorch deep learning library [23] was used and executed on a NVIDIA RTX Quadro 5000 GPU system with 16 GB GPU memory and 16 GB of RAM. For pre-training, the model is trained for 2000 epochs with learning rate of 0.0002 and 80-20 split is used for both datasets, i.e., 80% training data and 20% testing data. For fine-tuning, all methods are trained for 500 epoch having batch size of 72. In this work, we have followed the settings of the dataset split most often used in the literature. 80-20 split is used for SI dataset whereas session based split used in the original paper is employed for MI dataset. For both pre-training and fine-tuning, Adam optimizer is used with constant learning rate of 0.0002 with β_1 as 0.5 and β_2 as 0.999. The metric used to evaluate the model performance is accuracy, confusion matrix, AUC-ROC score and kappa measure [15,20].

Table 1. Accuracy comparison of proposed method with full fine-tuning and other parameter efficient adaptation methods for SI and MI EEG signal classification. [Standard deviation is reported in round brackets.]

Dataset	Subject	Methods			
		Fine-Tune	LoRA [9]	DoRA [17]	EDoRA(our)
Speech Imagery	1	51.67	50.00	50.00	53.33
	3	48.33	50.00	48.33	41.67
	5	48.33	45.00	45.00	50.00
	6	51.67	48.33	50.00	53.33
	8	58.33	65.00	65.00	61.67
	12	51.67	53.33	53.33	58.33
	Average	51.67 (3.65)	51.94 (6.95)	51.94 (6.95)	**53.06** (6.94)
Motor Imagery	1	81.60	79.86	79.51	81.94
	2	52.78	51.74	52.08	49.31
	3	85.76	84.38	84.38	86.11
	4	65.63	65.63	66.67	67.71
	5	42.36	44.44	44.44	44.79
	6	51.04	51.04	52.43	56.25
	7	81.25	78.13	78.13	80.90
	8	80.21	78.47	78.82	80.21
	9	79.86	72.92	73.96	78.13
	Average	68.94 (16.36)	67.40 (14.81)	67.82 (14.61)	**69.48** (15.60)

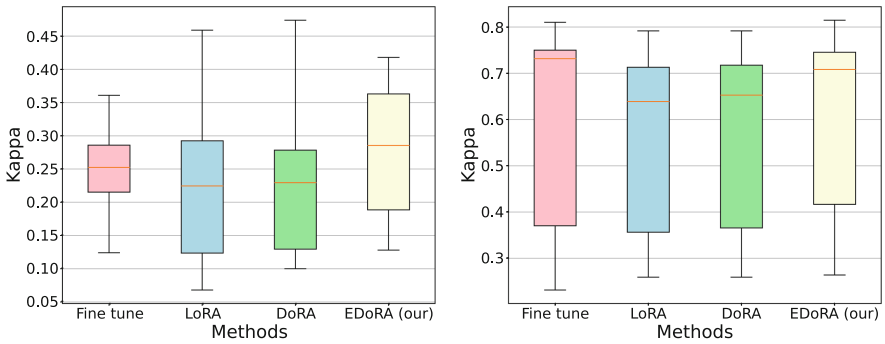
4.3 Performance Evaluation

The performance of the proposed parameter-efficient adaptation methods is validated through two mental imagery task datasets, i.e., the speech imagery dataset

(SI dataset) and the motor imagery dataset (MI dataset). The method is first pre-trained on one dataset and then adapted to the other dataset via fine-tuning and vice-versa. The evaluation performance of the each fine-tuned dataset is discussed in the following sections of this paper.

The performance of the proposed parameter-efficient adaptation method is stated in Table 1. The first method compared with the proposed method is a fully fine-tuned model (depicted as fine-tuned in the table). This method requires training all parameters of the pre-trained model to optimize the target domain data. The other compared methods are state-of-the-art (SOTA) parameter-efficient adaptation methods, LoRA [9] and DoRA [17]. The comparison table of accuracy metrics validates the effectiveness of the proposed adaptation method as our method has performed better than full fine-tuning, which requires a huge number of parameters to train, whereas the proposed method only requires a small number of the total trainable parameters.

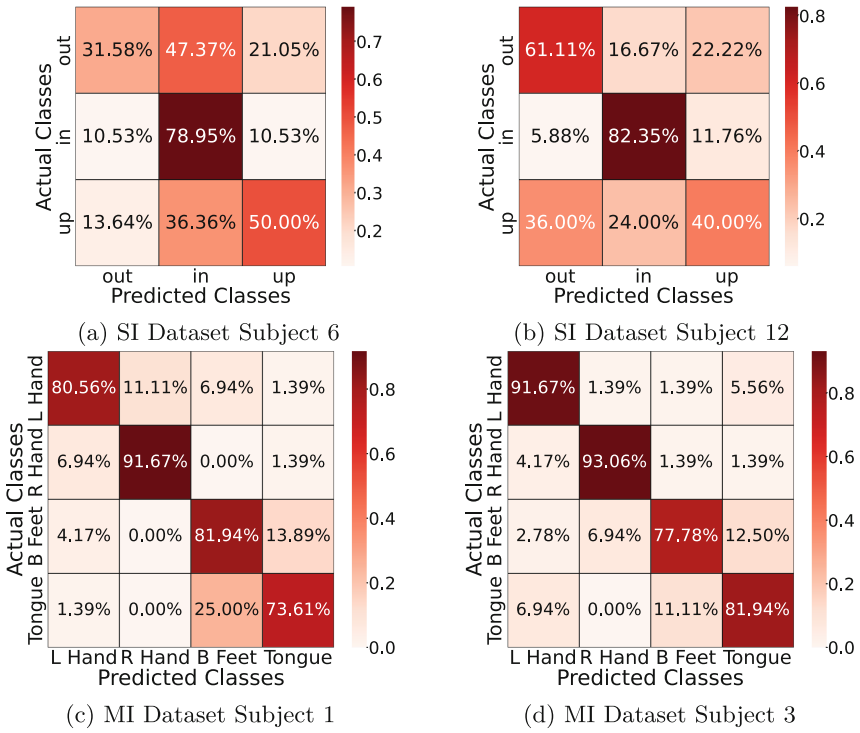
In the SI dataset, the accuracy of the proposed method is 1.39% more than full fine-tuning and 1.12% more than both LoRA and DoRA. Similarly, in the MI dataset, the proposed method is superior to the compared methods with margins of 0.54%, 2.08%, and 1.66% than full fine-tuning, LoRA and DoRA, respectively. Fig. 3 shows the kappa value comparison of the proposed method with compared methods. The box plots show that for the SI dataset, the median is clearly distant from the compared methods, and the minimum value of the kappa measure is also better than all of the compared methods. For the MI dataset, although the median of the proposed method is below the median of the full fine-tune method, the box length is smaller, and the minimum value of the kappa is higher. Thus, the kappa measure of both datasets suggests that the performance of the proposed method is better than the compared methods in both datasets.



(a) SI Dataset (b) MI Dataset

Fig. 3. Mean Kappa measure of proposed method and compared methods.

Fig. 4 and Fig. 5 show the per-class discriminative power of the proposed method. Fig. 4 shows the confusion matrices of two subjects of each dataset.



(a) SI Dataset Subject 6

(b) SI Dataset Subject 12

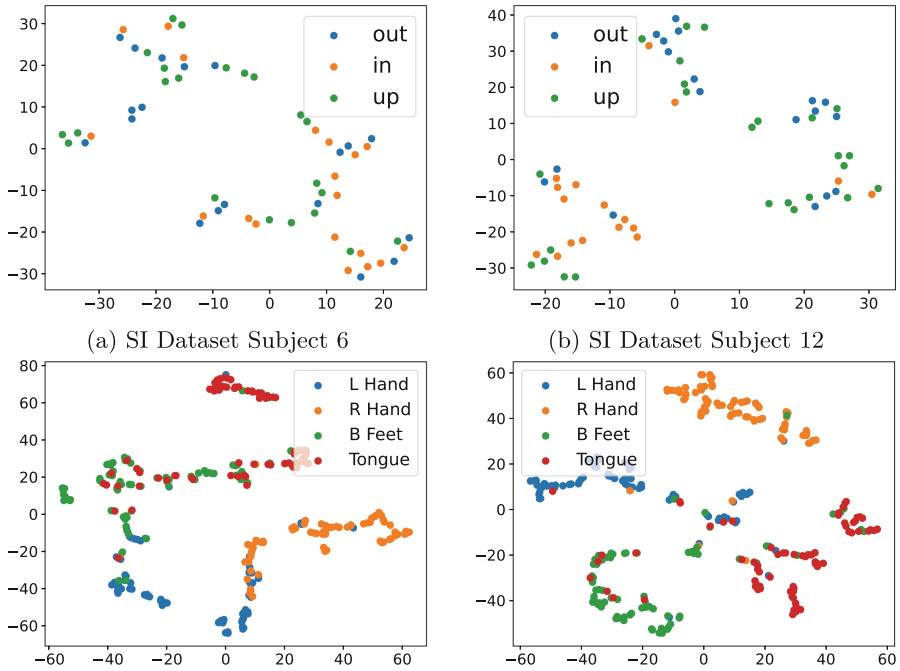
(c) MI Dataset Subject 1

(d) MI Dataset Subject 3

Fig. 4. Confusion matrices of proposed method on two subjects of SI and MI dataset.

Fig. 4a and Fig. 4b show that class “in" is more accurately classified compared to the other two classes. The reason might be the similarity of the “out" and “up" words and the dissimilarity of “in" from these two words. Also, the behavior is consistent on the confusion matrices of both subjects. Fig. 4c and Fig. 4d show that for motor imagery tasks, the per-class classification performance is balanced. However, left-hand imagined movement and right-hand imagined movement are least confused with other classes by the proposed methods. The class-wise clustering of the EEG signals in the feature space is shown with the help of t-SNE plots in Fig. 5. Fig. 5a and Fig. 5b show where the testing data points of the SI dataset lie in the feature space. These figures show the clusters of three classes of the SI dataset, where the cluster of the “in" class is clear and distinct, whereas the “out" and “up" classes show some overlap. This finding is similar to the confusion matrices of the SI dataset. Fig. 5c and Fig. 5d depict the t-SNE plots for two subjects of the MI dataset, and these plots show clear and distinct clusters of all four classes of the MI dataset. Therefore, confusion matrices and t-SNE plots show the classification performance of the proposed parameter-efficient adaptation method.

The parameter-efficient adaptation methods have a hyper-parameter of rank that decides the number of neurons of the adapters. Hence, we present the com-



(a) SI Dataset Subject 6 (b) SI Dataset Subject 12

(c) MI Dataset Subject 1 (d) MI Dataset Subject 3

Fig. 5. t-SNE plots of proposed method (EDoRA) on two subjects of SI and MI dataset.

parison of the performance of the proposed parameter-efficient method EDoRA with different rank values. Table 2 shows the accuracy metric of the proposed EDoRA method and compared methods. In this table, for each rank, the highest accuracy measure among LoRA, DoRA, and EDoRA (our) is shown in bold. For both datasets, EDoRA has performed superior to compared parameter-efficient adaptation methods with the exception of rank 2, where LoRA (68.90%) performed slightly better than EDoRA (68.63%). However, maximum accuracy with the least standard deviation is attained by EDoRA only (rank=8, n=4, accuracy=69.52%). Additionally, the highest accuracy of the proposed method is better than the accuracy of the full fine-tune method. It shows the strength of the low-rank adaptation approach with a very small number of parameters over full fine-tuning. These results demonstrate the ability of EDoRA to adapt from one mental imagery dataset to another mental imagery dataset and vice versa.

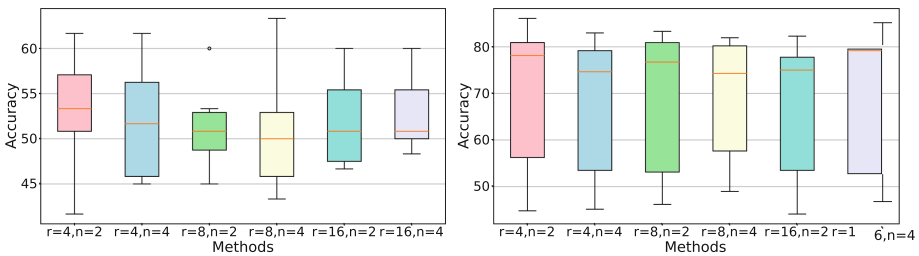
4.4 Ablation study

In our proposed ensemble of the PEFT adapter method, the number of adapters (n) is a hyper-parameter of the EDoRA method. Therefore, we demonstrate the impact of the number of adapters in an ensemble with different ranks (r) on EDoRA in Fig. 6. This ablation study shows that for the SI dataset, the

Table 2. Accuracy comparison of proposed method with full fine-tuning and SOTA parameter-efficient adaptation methods with different ranks (r) for SI and MI datasets. [D: Dataset, Standard deviation is reported in round brackets, EDora† is proposed method with n=4.]

D	Rank	Methods				
		Fine-tune	LoRA	DoRA	EDoRA(our)	EDora†(our)
SI	2	51.67 (3.65)	52.22 (6.47)	52.22 (6.64)	**53.33** (8.63)	–
	4		51.94 (6.95)	51.94 (6.95)	**53.06** (6.94)	51.95 (6.87)
	8		50.56 (6.12)	**51.39** (6.78)	**51.39** (5.10)	50.83 (7.21)
	16		52.50 (7.28)	51.11 (6.38)	51.94 (5.42)	**52.78** (4.55)
MI	2	68.94 (16.36)	**68.90** (15.81)	68.75 (15.58)	68.63 (15.27)	–
	4		67.40 (14.81)	67.82 (14.61)	**69.48** (15.60)	68.09 (14.34)
	8		68.36 (15.80)	67.94 (16.20)	**68.94** (14.50)	69.52 (13.08)
	16		66.63 (14.79)	67.48 (14.66)	67.94 (14.30)	**68.83** (14.82)

performance of all variants is quite variable. This phenomenon can be because the number of samples in the testing set is only 60. It is a small number, and misclassifying only a few samples can result in lower accuracy. Hence, variants of the SI dataset show a variable performance. However, for the MI dataset, the performance is similar for all of the variants of EDoRA as the number of testing samples is 288, which is comparatively larger than the testing samples of the SI dataset. Therefore, this analysis suggests the robustness of the proposed method EDoRA over different numbers of n and rank r in both datasets.



(a) SI Dataset (b) MI Dataset

Fig. 6. Accuracy of proposed method (EDoRA) with different ranks (r) and segment (n).

4.5 Parameter analysis

Table 3 shows the number of trainable parameters required by EDoRA and compared PEFT methods (LoRA and DoRA). According to this table, the difference

Table 3. Trainable parameter requirement and mean accuracy of proposed method and compared state-of-the-art parameter-efficient adaptation methods.

Dataset	Method	LoRA [9]	DoRA [17]	EDoRA(our)
Speech Imagery	#Parameters	17k	19k	21k
	Average Accuracy	51.94 (6.95)	51.94 (6.95)	**53.06** (6.94)
Motor Imagery	#Parameters	17k	19k	21k
	Average Accuracy	67.40 (14.81)	67.82 (14.61)	**69.48** (15.6)

in parameters between EDoRA and DoRA is only $2k$, but the difference in accuracy is 1.12% and 1.66% in SI and MI datasets, respectively. A similar behavior is observed when comparing EDoRA with LoRA. It shows that our method is comparable to the other SOTA methods in the trade-off of parameters and performance.

5 Conclusion

In this work, we explored the parameter-efficient fine-tuning methods for EEG-based mental imagery task adaptation. Our work is the first to explore the performance and efficiency of parameter-efficient adaptation methods that do not require all parameters of the pre-trained model to be trained. Instead, it trains only a small amount of parameters based on the rank-decomposition technique. We proposed EDoRA, a parameter-efficient fine-tuning method that is an ensemble of multiple parameter-efficient adapters. These adapters decompose the pre-trained weights into magnitude and direction components and adapt these components to the target domain for enhanced fine-tuning. This work is the first to investigate transfer learning-based domain adaptation of speech imagery task from motor imagery task and vice-versa. The effectiveness of the proposed method is validated on two publicly available mental imagery datasets, one dataset of speech imagery and the other of motor imagery. For elevated feature extraction, we have adopted a convolutional transformer approach-based model known as EEG Conformer. The adaptation framework comprises pre-training the model on one dataset and then fine-tuning the model on another dataset. The performance evaluation on these two datasets exhibits the effectiveness and robustness of the proposed approach in the domain of EEG signal classification. In the future, the proposed method can be optimized to make it more parameter-efficient while increasing performance.

References

1. Ahn, H.J., Lee, D.H., Jeong, J.H., Lee, S.W.: Multiscale convolutional transformer for EEG classification of mental imagery in different modalities. IEEE Trans. Neural Syst. Rehabil. Eng. **31**, 646–656 (2022)

2. Chakladar, D.D., Roy, P.P., Iwamura, M.: EEG-based cognitive state classification and analysis of brain dynamics using deep ensemble model and graphical brain network. IEEE Transactions on Cognitive and Developmental Systems **14**(4), 1507–1519 (2021)
3. Chen, P., Gao, Z., Yin, M., Wu, J., Ma, K., Grebogi, C.: Multiattention adaptation network for motor imagery recognition. IEEE Transactions on Systems, Man, and Cybernetics: Systems **52**(8), 5127–5139 (2021)
4. Chen, Y., Yang, R., Huang, M., Wang, Z., Liu, X.: Single-source to single-target cross-subject motor imagery classification based on multisubdomain adaptation network. IEEE Trans. Neural Syst. Rehabil. Eng. **30**, 1992–2002 (2022)
5. Edelman, B.J., Meng, J., Suma, D., Zurn, C., Nagarajan, E., Baxter, B.S., Cline, C.C., He, B.: Noninvasive neuroimaging enhances continuous neural tracking for robotic device control. Science robotics **4**(31), eaaw6844 (2019)
6. Frolov, A.A., Mokienko, O., Lyukmanov, R., Biryukova, E., Kotov, S., Turbina, L., Nadareyshvily, G., Bushkova, Y.: Post-stroke rehabilitation training with a motor-imagery-based brain-computer interface (BCI)-controlled hand exoskeleton: a randomized controlled multicenter trial. Front. Neurosci. **11**, 400 (2017)
7. Hang, W., Feng, W., Du, R., Liang, S., Chen, Y., Wang, Q., Liu, X.: Cross-subject EEG signal recognition using deep domain adaptation network. IEEE Access **7**, 128273–128282 (2019)
8. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for NLP. In: International conference on machine learning. pp. 2790–2799. PMLR (2019)
9. Hu, E.J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2021)
10. Hu, L., Hong, W., Liu, L.: Msatnet: multi-scale adaptive transformer network for motor imagery classification. Front. Neurosci. **17**, 1173778 (2023)
11. Huang, X., Choi, K.S., Zhou, N., Zhang, Y., Chen, B., Pedrycz, W.: Shallow inception domain adaptation network for EEG-based motor imagery classification. IEEE Transactions on Cognitive and Developmental Systems **16**(2), 521–533 (2023)
12. Huang, Y.M., Hung, H.N., Tseng, V.S.: Privacy-preserving attention-weighted multi-source domain adaptation for EEG motor imagery. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7470–7474. IEEE (2024)
13. Jeon, E., Ko, W., Suk, H.I.: Domain adaptation with source selection for motor-imagery based BCI. In: 2019 7th International Winter Conference on Brain-Computer Interface (BCI). pp. 1–4. IEEE (2019)
14. Jiménez-Guarneros, M., Gómez-Gil, P.: Standardization-refinement domain adaptation method for cross-subject EEG-based classification in imagined speech recognition. Pattern Recogn. Lett. **141**, 54–60 (2021)
15. Kaushik, P., Tripathi, I., Roy, P.P.: Motor activity recognition using eeg data and ensemble of stacked BLSTM-LSTM network and transformer model. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
16. Lee, J., Choi, J.W., Jo, S.: Selective multi-source domain adaptation network for cross-subject motor imagery discrimination. IEEE Transactions on Cognitive and Developmental Systems (2023)
17. Liu, S.Y., Wang, C.Y., Yin, H., Molchanov, P., Wang, Y.C.F., Cheng, K.T., Chen, M.H.: DoRA: Weight-decomposed low-rank adaptation. In: Forty-first International Conference on Machine Learning (2024)

18. Liu, S., Zhang, J., Wang, A., Wu, H., Zhao, Q., Long, J.: Subject adaptation convolutional neural network for EEG-based motor imagery classification. J. Neural Eng. **19**(6), 066003 (2022)
19. Lotey, T., Keserwani, P., Dogra, D.P., Roy, P.P.: Feature reweighting for EEG-based motor imagery classification. arXiv preprint arXiv:2308.02515 (2023)
20. Lotey, T., Keserwani, P., Wasnik, G., Roy, P.P.: Cross-session motor imagery EEG classification using self-supervised contrastive learning. In: 26th International Conference on Pattern Recognition. pp. 975–981. IEEE (2022)
21. Miao, M., Yang, Z., Zeng, H., Zhang, W., Xu, B., Hu, W.: Explainable cross-task adaptive transfer learning for motor imagery EEG classification. J. Neural Eng. **20**(6), 066021 (2023)
22. Nguyen, C.H., Karavas, G.K., Artemiadis, P.: Inferring imagined speech using EEG signals: a new approach using riemannian manifold features. J. Neural Eng. **15**(1), 016002 (2017)
23. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems **32** (2019)
24. Phunruangsakao, C., Achanccaray, D., Hayashibe, M.: Deep adversarial domain adaptation with few-shot learning for motor-imagery brain-computer interface. IEEE Access **10**, 57255–57265 (2022)
25. Saha, S., Baumert, M.: Intra-and inter-subject variability in EEG-based sensorimotor brain computer interface: a review. Front. Comput. Neurosci. **13**, 87 (2020)
26. Simistira Liwicki, F., Gupta, V., Saini, R., De, K., Abid, N., Rakesh, S., Wellington, S., Wilson, H., Liwicki, M., Eriksson, J.: Bimodal electroencephalography-functional magnetic resonance imaging dataset for inner-speech recognition. Scientific Data **10**(1), 378 (2023)
27. Song, Y., Zheng, Q., Liu, B., Gao, X.: Eeg conformer: Convolutional transformer for EEG decoding and visualization. IEEE Trans. Neural Syst. Rehabil. Eng. **31**, 710–719 (2022)
28. Song, Y., Zheng, Q., Wang, Q., Gao, X., Heng, P.A.: Global adaptive transformer for cross-subject enhanced EEG classification. IEEE Trans. Neural Syst. Rehabil. Eng. **31**, 2767–2777 (2023)
29. Tangermann, M., Müller, K.R., Aertsen, A., Birbaumer, N., Braun, C., Brunner, C., Leeb, R., Mehring, C., Miller, K.J., Müller-Putz, G.R., et al.: Review of the BCI competition IV. Front. Neurosci. **6**, 55 (2012)
30. Xia, K., Deng, L., Duch, W., Wu, D.: Privacy-preserving domain adaptation for motor imagery-based brain-computer interfaces. IEEE Trans. Biomed. Eng. **69**(11), 3365–3376 (2022)
31. Xu, D.q., Li, M.a.: A dual alignment-based multi-source domain adaptation framework for motor imagery EEG classification. Applied Intelligence **53**(9), 10766–10788 (2023)
32. Young, A.J., Ferris, D.P.: State of the art and future directions for lower limb robotic exoskeletons. IEEE Trans. Neural Syst. Rehabil. Eng. **25**(2), 171–182 (2016)
33. Zhang, Y., Qiu, S., Wei, W., Ma, X., He, H.: Dynamic weighted filter bank domain adaptation for motor imagery brain-computer interfaces. IEEE Transactions on Cognitive and Developmental Systems **15**(3), 1348–1359 (2022)
34. Zhong, X.C., Wang, Q., Liu, D., Liao, J.X., Yang, R., Duan, S., Ding, G., Sun, J.: A deep domain adaptation framework with correlation alignment for EEG-based motor imagery classification. Comput. Biol. Med. **163**, 107235 (2023)

Neural Networks Meet Neural Activity: Utilizing EEG for Mental Workload Estimation

Gourav Siddhad[1(✉)], Partha Pratim Roy[1], and Byung-Gyu Kim[2]

[1] Department of Computer Science and Engineering, Indian Institute of Technology, Roorkee, Roorkee 247667, Uttarakhand, India
{g_siddhad,partha}@cs.iitr.ac.in

[2] Division of AI Engineering, Sookmyung Women's University, Seoul 04310, Republic of Korea
bg.kim@sookmyung.ac.kr

Abstract. Electroencephalography (EEG) offers non-invasive, real-time mental workload assessment, which is crucial in high-stakes domains like aviation and medicine and for advancing brain-computer interface (BCI) technologies. This study introduces a customized ConvNeXt architecture, a powerful convolutional neural network, specifically adapted for EEG analysis. ConvNeXt addresses traditional EEG challenges like high dimensionality, noise, and variability, enhancing the precision of mental workload classification. Using the STEW dataset, the proposed ConvNeXt model is evaluated alongside SVM, EEGNet, and TSception on binary (No vs SIMKAP task) and ternary (SIMKAP multitask) class mental workload tasks. Results demonstrated that ConvNeXt significantly outperformed the other models, achieving accuracies of 95.76% for binary and 95.11% for multi-class classification. This demonstrates ConvNeXt's resilience and efficiency for EEG data analysis, establishing new standards for mental workload evaluation. These findings represent a considerable advancement in EEG-based mental workload estimation, laying the foundation for future improvements in cognitive state measurements. This has broad implications for safety, efficiency, and user experience across various scenarios. Integrating powerful neural networks such as ConvNeXt is a critical step forward in non-invasive cognitive monitoring.

Keywords: Brain-Computer Interface · ConvNeXt · Electroencephalography · Mental Workload · STEW Dataset

1 Introduction

With its capacity to directly record the complex electrical activity of the brain, electroencephalography (EEG) is a potent technique for non-invasively assessing mental workload in real-time. Evaluating mental workload is crucial in high-stakes fields like medicine and aviation, but it's also important in regular workplaces and educational settings [25]. This non-invasive method is essential for

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15311, pp. 325–339, 2025.
https://doi.org/10.1007/978-3-031-78195-7_22

improving brain-computer interface (BCI) applications and understanding neural processes. Its applications range from ensuring safety in critical operations to optimizing cognitive performance and tailoring learning experiences. Within the quickly developing field of cognitive neuroscience, EEG is essential for bridging the knowledge gap between theoretical concepts and real-world applications related to mental stress [8]. By examining the electrical patterns obtained from EEG, researchers can gain crucial knowledge about healthy and pathological brain functions [2]. This improves the capacity to control and interpret cognitive loads in various scenarios. This finding highlights EEG's dual potential to advance scientific understanding and practical application. These capabilities underline the EEG's critical role in the realms of current neuroscience and BCI research.

EEG, with its high temporal resolution, is a valuable tool for capturing rapid changes in mental workload. However, its high dimensionality, intrinsic noise, and non-stationarity [3] make it challenging to extract meaningful information. Additionally, the complexity of the human brain and limitations of EEG technology, such as variations in cognitive abilities, low signal-to-noise ratio, and poor spatial resolution [14], complicate the accurate localization of neural sources. Individual differences in brain responses and EEG acquisition methods further contribute to signal variability. To address these challenges and fully leverage EEG's potential in high-stakes situations, advanced machine learning models are crucial [26]. These models can improve the accuracy and consistency of EEG data processing, facilitating the development of user-friendly systems for estimating mental workload. By incorporating subject-specific information, such as task complexity and individual cognitive characteristics, these models can enhance the classification accuracy of mental workload states [11]. This multidisciplinary approach is expected to drive significant advancements in the field, enabling more effective monitoring and analysis of cognitive processes.

Enter ConvNeXt [19], a cutting-edge convolutional neural network (CNN) architecture initially designed to address computer vision problems. ConvNeXt is modified for EEG analysis in this work, which is a significant innovation. ConvNeXt has the potential to revolutionise EEG data processing by extracting subtle patterns from EEG signals more precisely and effectively than conventional models because of its improved convolutional operations, optimised layer structures, and effective training methodologies. This modification holds great potential for precisely detecting the complex patterns in brain signals that correlate to different mental workloads. It represents a shift from conventional EEG analysis techniques and could result in workload estimates that are more accurate and insightful. Because of ConvNeXt's improved feature extraction capabilities, neural signatures associated with different brain states, clinical diseases, and cognitive tasks may be precisely identified. This expands the field of neuroscience research and advances BCI technology, with broad implications ranging from better seizure detection to enhanced sleep stage categorization and cognitive load evaluation.

This study investigates the integration of the ConvNeXt architecture into EEG data analysis. It highlights the potential of ConvNeXt to revolutionize the precision of computational neuroscience and BCI research. Researchers can gain profound insights into the brain's electrical activity across various cognitive states by analysing EEG data. ConvNeXt's advanced design efficiently learns complex patterns within EEG data, outperforming traditional models in feature extraction. The use of ConvNeXt tackles the challenges associated with EEG signals, offering significant improvements in cognitive performance and ergonomics across diverse settings. Its potential to differentiate between types of cognitive loads sets a new standard for EEG analysis, marking a substantial advancement in the field. As research continues, the application of ConvNeXt in EEG analysis is expected to drive breakthroughs in understanding the brain's complexities and advancing cognitive ergonomics.

Contributions of this work are:

- While CNNs have shown promise in EEG classification, ConvNeXt's advanced architecture ushers in a new era for efficiently extracting complex patterns from EEG data, potentially surpassing traditional models by a substantial margin.
- Beyond classification accuracy, techniques to interpret the learned features within the ConvNeXt model will be explored. This will provide valuable insights into the specific neural activity patterns associated with different cognitive states, offering a deeper understanding of brain processes during controlled experiments.
- This work aims to significantly improve the current standards of EEG analysis, offering deeper insights into brain activity and paving the way for advancements in cognitive research.

The paper is structured as follows: Section 2 reviews research on mental workload and EEG analysis, highlighting the challenges and opportunities. Section 3 delves into the ConvNeXt model and its adaptation for EEG, including the description of the dataset and classifiers. Section 5 presents experimental findings, and Section 6 concludes this work and discusses future directions.

2 Related Work

Recent research on estimating mental workload using EEG signals has produced various innovative methods and insights. Early studies, such as those by Hernandez et al. [12], explored the evaluation of pilots' mental workloads in high-risk cockpit environments through multitasking. The work by Di et al. [9] extended EEG analysis to the driving context, integrating EEG data with subjective assessments and vehicle dynamics to study the effects of traffic and road conditions on driver workload. Kartali et al. [13] contributed by focusing on real-time mental workload estimation using EEG. Singh et al. [32] applied a combination of 1D-CNN and Synthetic Minority Oversampling Technique (SMOTE) to enhance the classification accuracy of mental workload levels.

Advancements in machine learning have substantially improved the application and understanding of EEG technology in environments characterized by high stress and multitasking. Qu et al. [24] addressed the challenge of EEG signal non-stationarity through cross-session subspace alignment, significantly improving signal classification across sessions. Mastropietro et al. [20] highlighted the importance of electrode configurations and signal processing techniques in enhancing the sensitivity and reproducibility of EEG-based mental workload measurements. In their studies, So et al. [33] employed Support Vector Machines (SVM) to accurately estimate workload levels, demonstrating the efficiency of a single-channel EEG device in monitoring dynamic changes with notable accuracy. Similarly, Chin et al. [5] provided evidence that EEG can effectively differentiate between various levels of cognitive workload during mental arithmetic tasks, illustrating the method's versatility and effectiveness.

Furthermore, various machine learning algorithms, including SVM, Random Forest, and k-Nearest Neighbors (k-NN), have been applied to analyze EEG signals. Research by Singh et al. [31] and Pandey et al. [21] has demonstrated varied success rates, particularly in multitasking scenarios where the analysis of statistical and fractal dimension (FD) features is crucial, as shown by Lim et al. [16]. These techniques have proven particularly effective in enhancing the practical applications of EEG in diverse cognitive studies.

Innovative computational approaches have greatly enhanced EEG-based classification tasks by integrating advanced deep-learning models [30]. Cheng et al. [4] introduced a novel combination of 3D-CNN with LSTM and attention mechanisms, significantly improving spatial-temporal feature learning. Similarly, Yao et al. [35] demonstrated the efficacy of combining CNN and transformer models, which has significantly advanced EEG-based classification tasks. Further contributions in this field include the work of Aldawsari et al. [1], who optimized a 1D-CNN model, showcasing the potential of lightweight deep learning methods for real-time EEG-based emotion recognition. Saleh et al. [27] further leveraged transformer networks to enhance classification accuracy on eye direction, PPG, and EEG data, demonstrating the versatility of transformers in handling various types of physiological data.

Additionally, Siddhad et al. [29] utilized transformer networks for classifying raw EEG data, illustrating their adaptability to tasks beyond natural language processing, such as mental workload classification. This adaptation addresses common challenges in EEG data classification, including the dependency on preprocessing and the need for hand-crafted feature extraction, by leveraging deep learning to potentially automate these processes. Moreover, Parveen et al. [22] introduced an attention-based 1D-CNN for mental workload classification. This model enables the identification of specific patterns of brain activity associated with various workload levels, highlighting the precise capabilities of modern computational models in interpreting complex neural signals. These developments collectively push the boundaries of EEG data analysis, paving the way for more accurate and efficient applications in various fields.

In addition to machine learning, using artificial neural networks (ANN) in EEG analysis has been prominently featured, with research by Samima et al. [28] demonstrating notably high accuracy in estimating mental workload in operators. This achievement illustrates the increasing integration of complex neural network architectures in EEG data analysis, highlighting a significant trend in the field. This trend is further contextualized within the broader scope of advancements discussed in the 'Brain Informatics' collection by Liu et al. [18]. This collection covers a vast range of studies in brain science, human information processing systems, and brain big data analytics. It points to integrating advanced computational models like transformers in EEG analysis as part of a broader exploration of brain informatics technologies. Such integration is pivotal in advancing a comprehensive understanding of mental health through informatics paradigms, demonstrating the interconnection between innovative technological applications and fundamental brain science research.

Measurement, classification, and understanding of cognitive load have advanced significantly due to studies on EEG signals for mental workload estimation, especially in high-stress, multitasking situations. These researches have played a pivotal role in developing systems that evaluate cognitive load in real-time across various areas, including workplace safety, aviation, education, and the automotive industry. This has led to a deeper understanding of EEG and its practical applications. To fully utilize EEG-based workload estimation, several obstacles remain. These include the need for more in-depth task analysis that considers the complexities of real-world scenarios, advancements in real-time processing systems for increased accuracy and reduced latency, and the requirement for personalized models to address significant inter-subject variability and enhance accuracy.

Additionally, it is imperative to identify and optimize predictive EEG features using advanced machine-learning techniques. Integrating EEG data with other physiological indicators can provide a more thorough workload assessment. Models must also adapt to dynamic workload levels over prolonged tasks and consider individual cognitive differences and mental states for more precise estimations. Lastly, improving the usability and wearability of EEG devices is essential to facilitate their broader adoption, especially in workplace settings where practicality and comfort are critical. These collective efforts highlight the crucial role of advanced computational models and machine learning techniques in refining EEG applications and addressing their challenges to maximize their effectiveness across various fields.

Improving EEG-based mental workload estimation requires addressing the intrinsic challenges of EEG data, namely, high dimensionality, noise susceptibility, and non-stationary nature. Recurrent neural networks (RNNs) and convolutional neural networks (CNNs) are examples of advanced computational models necessary for precisely interpreting EEG data in precision-critical applications like brain-computer interfaces and medical diagnosis. These models effectively manage data from multiple scalp electrodes through dimensionality reduction and automatic feature extraction, increasing analytical accuracy and effi-

ciency [10]. Moreover, EEG readings can be distorted by noise and artefacts from electrical interference, muscles, and other sources, which can mask brain activity. To overcome these problems, models like EEGNet use advanced noise-reduction techniques, which produce more reliable analyses [7]. The analysis is further complicated by the non-stationary nature of EEG data, which can be impacted by shifts in the subject's physiological or cognitive state, task involvement levels, or environmental variables. Adaptive filtering or time-frequency analysis techniques are essential for capturing the dynamic elements of EEG signals because traditional signal processing approaches, which presume stationarity, are insufficient [34]. Improvements in these areas can greatly increase the responsiveness, accuracy, and ease of use of EEG-based mental workload estimation, hence increasing its applications.

Incorporating advanced computational models is critical to effectively handle the intricacies of EEG data and improve applications in domains where accuracy is paramount, such as neurofeedback therapy. These models enable a robust and reliable method of EEG analysis by combining insights from computer science, cognitive science, and neuroscience. Essentially, using these models to read EEG signals reliably improves the interpretation of complex data, leading to more dependable and efficient applications in various fields.

3 ConvNeXt

This study incorporates a cutting-edge CNN, the ConvNeXt architecture [19], into EEG-based measures of mental workload. ConvNeXt was originally developed for computer vision, but its ability to effectively learn complicated patterns within noisy and high-dimensional data makes it a promising tool for EEG analysis. The success of this hybrid architecture in image-related tasks can be attributed to its combination of the capabilities from both attention-based mechanisms of Transformer models and classic CNNs. The well-known ResNet design, which is renowned for its residual connections that allow the training of extremely deep networks, serves as an inspiration for ConvNeXt. ConvNeXt, however, differs greatly due to key modifications. It uses larger kernel sizes in its depthwise convolutions for wider receptive fields, layer normalisation for increased stability, and inverted residual structures to optimize computational efficiency. These modifications make ConvNeXt very well-suited to handle the complexities of EEG signals.

The ConvNeXt architecture is redesigned using convolutional blocks that gradually downsample input images while boosting channel capacity, structuring it into multiple stages suitable for EEG analysis. This modification improves the model's capacity to extract workload-related patterns by addressing the unique challenges of EEG data, such as high dimensionality, noise levels, and subtle signal fluctuations. Updated convolutional layers, stochastic depth, data augmentation, and LayerNorm, which stabilise the learning process, are some major modifications. Due to its scalability and processing efficiency, ConvNeXt is suited

for real-time applications such as mental workload evaluation, sleep stage classification, and seizure detection. Through techniques like extensive data augmentation and a revised learning rate schedule to prevent overfitting and increase convergence, the combination of this advanced architecture with EEG data has the potential to revolutionise analysis, enhancing precision and extending the spectrum of applications.

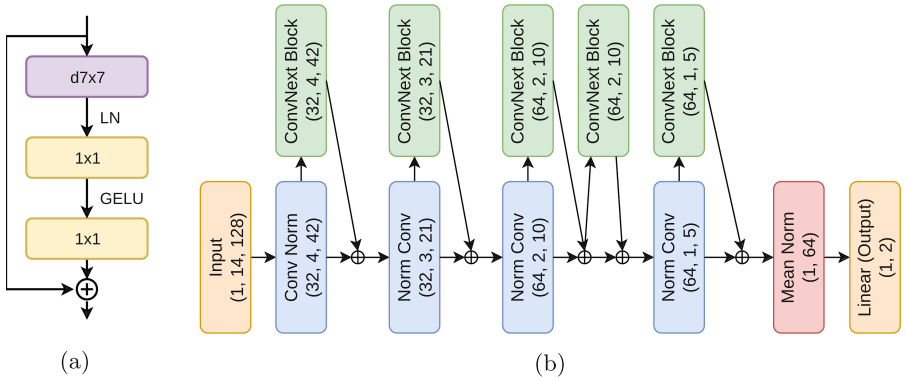


Fig. 1. (a) ConvNeXt Block and (b) ConvNeXt Model. Conv, Norm, and Mean in the ConvNeXt model represent 2D Convolution, Normalization, and Mean layer, respectively. The shape in each block represents the output shape of that block.

The ConvNeXt architecture introduces several modifications over traditional CNNs to enhance its adaptability. Key to these adaptations is the replacement of all batch normalization with layer normalization. Unlike batch normalization, which standardizes inputs using batch mean and variance, layer normalization standardizes inputs across each feature for every data point, providing consistency regardless of batch size variations. In ConvNeXts, the AdamW optimizer is employed, which refines the standard Adam approach by applying accurate weight decay, thus improving regularization and generalization capabilities. It is further distinguished by its use of 1×1 conv layers, with the depth-wise conv layer repositioned at the top of the stack, unlike in ResNeXt blocks. Additionally, the Gaussian Error Linear Unit (GELU), a smoother variant of the ReLU, is utilized, enhancing the network's non-linear processing capabilities.

Significant architectural scaling was undertaken for ConvNeXt, as illustrated in Fig 1b, for its application to EEG, where the original ConvNeXt dimensions were tailored for 224×224 image sizes. Specifically, the arrangement of ConvNeXt blocks was altered from $(3, 3, 9, 3)$ to $(1, 1, 2, 1)$, and the number of channels in convolutional layers was decreased from $(96, 192, 384, 768)$ to $(32, 32, 64, 64)$. These changes reduced the model's complexity and the number of parameters and decreased training time, enhancing efficiency. The adapted model was subsequently trained for 100 epochs, optimizing it for the specific challenges and requirements of EEG data analysis as applied in this context.

The success of this ConvNeXt adaptation to EEG has the potential for broader applications in biomedical engineering and cognitive neuroscience, where precise pattern recognition is crucial. This approach significantly advances deep learning applications in neurological assessment, bridging the gap between high-level computer vision techniques and EEG signal analysis.

4 Experiments

4.1 Experimental Data

To validate the performance of the proposed methodology, it was tuned and applied to the open-access mental workload dataset known as the simultaneous task EEG workload (STEW) dataset [17]. It consists of raw EEG data from 48 subjects who participated in a multitasking workload experiment that utilized the simultaneous capacity (SIMKAP) multitasking test. The signals were captured using the Emotiv EPOC EEG headset, with 16-bit A/D resolution, 128 Hz sampling frequency, and 14 channels, namely AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, AF4 according to the 10-20 international system with two reference channels (CMS, DRL). There are two parts to the experiment. First, the data was acquired for 2.5 minutes with subjects at rest or "No task". Next, subjects performed the SIMKAP test with EEG being recorded and the final 2.5 minutes were used as the workload condition. Subjects rated their perceived mental workload on a rating scale of 1-9 after each experiment segment.

EEG signals in their raw form (captured from a device) contain noise and artefacts and must be cleaned before use. EEG data is imported and bandpass filtering is done to remove environmental/muscle noise from scalp EEG. After epoching and removing bad epochs from the data, independent component analysis (ICA) is applied and bad channels are manually removed. The dataset is used with a sampling rate of 128 Hz, the same as during acquisition. After min-max scaling, the data was epoched into one-second intervals with a half-second overlap, resulting in data shaped as (1, channel count, EEG length), i.e., (1, 14, 128), yielding a total of 26,910 samples. The dataset is split into 70:15:15 ratios for train, validation, and test sets.

4.2 Experimental Setup

The experimental setup involved a DELL Precision 7820 Tower Workstation, with Ubuntu 22.04 OS, Intel Core(TM) Xeon Silver 4216 CPU, and an NVIDIA RTX A2000 12GB GPU. This hardware facilitated the implementation of DL models using Python 3.10 and the PyTorch library. The Adam optimizer, known for its computational efficiency, was used with default parameters ($\eta = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$). EEGNet and TSception were trained for 100 epochs, with batches of 16 and a learning rate of $1e-4$. The Radial Basis Function (RBF) kernel from scikit-learn [23] was used with default settings for SVM. Classification accuracy was determined through stratified five-fold cross-validation, averaging the results for comprehensive assessment.

4.3 Classifiers

This study uses four popular models for EEG analysis, namely EEGNet, TSception, Transformer, and SVM. EEGNet [15] is a compact convolutional neural network specifically designed for EEG data. Its success in a range of EEG tasks, efficiency, and reduced complexity make it appropriate for smaller datasets, which has led to a rise in its use in both research and practical applications. A deep learning model designed for time-series EEG data, TSception [10] highlights the temporal dynamics in the data. It is skilled at capturing the intricacies found in EEG data because it can effectively extract temporal information at different scales. TSception has demonstrated efficacy in identifying emotions and evaluating cognitive burden. The Transformer [29] is a neural network architecture that relies on self-attention mechanisms to compute representations of input sequences. By avoiding recurrence and convolution, the Transformer is highly parallelizable and efficient, capturing long-range dependencies in the data. Its encoder-decoder structure allows for sophisticated processing and understanding of complex EEG patterns. A popular machine learning algorithm used in EEG analysis is SVM [6]. By determining the ideal distance in high-dimensional space between classes, SVMs are excellent at classification. Robustness is one of SVM's strongest points; it performs particularly well with high-dimensional EEG data and can handle non-linear correlations using kernels. Each method, TSception, SVM, and EEGNet, offers unique advantages for EEG analysis. Neural network-based, EEGNet and TSception are especially good at processing raw EEG data and automatically extracting useful features. At the same time, SVM offers a robust classification method with well-understood theoretical foundations.

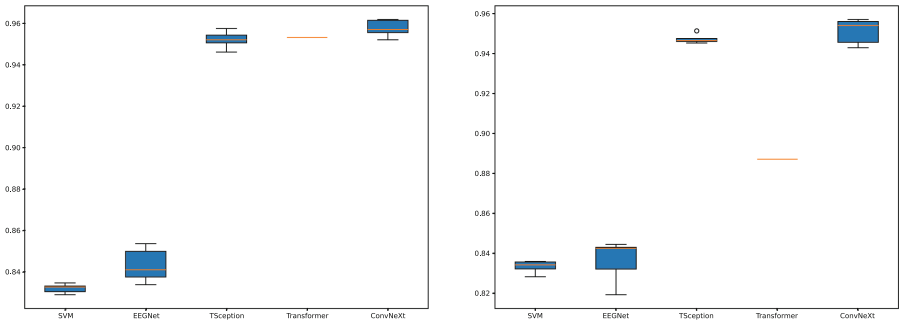
5 Performance Evaluation

This research aims to improve understanding of cognitive states by estimating mental workload through EEG, an essential tool in cognitive neuroscience. The effectiveness of several advanced machine learning classifiers was evaluated on the STEW dataset, a benchmark in cognitive load research. The study involved two primary analyses: a binary classification task (No vs. SIMKAP) and a more complex multi-class task (SIMKAP multi-task), designed to test the classifiers under varying complexities of mental workload representation. Five classifiers were assessed: SVM, EEGNet, TSception, Transformer, and a novel classifier introduced in this study. Their performance was measured in terms of accuracy and reliability. Results were substantiated by 95% confidence intervals to confirm the robustness of the findings.

The results demonstrate significant contributions to the discourse on neural network-based approaches and machine learning techniques in EEG analysis. The ConvNeXt model, in particular, showcased exemplary performance in accurately classifying mental workload levels under diverse conditions. This achievement sets new standards for accuracy and reliability in mental workload

Table 1. Performance Comparison on STEW dataset with 95% Confidence Interval

Classifier	No vs SIMKAP task 2 class	SIMKAP multi-task 3 class
SVM [6]	83.34 ± 0.39	83.21 ± 0.28
EEGNet [15]	84.33 ± 1.03	83.64 ± 1.33
TSception [10]	95.21 ± 0.53	94.73 ± 0.30
Transformer [29]	95.32 ± 0.00	88.72 ± 0.00
Proposed ConvNeXt	$\mathbf{95.76 \pm 0.51}$	$\mathbf{95.11 \pm 0.80}$



(a) (b)

Fig. 2. Box plot for (a) 'No vs SIMKAP task' and (b) 'SIMKAP multi-task'.

estimation. The analysis methodically presents classification accuracies, highlighting the technical strengths of this approach and its broader implications for cognitive neuroscience and human-computer interaction.

The evaluation methodology utilizes the STEW dataset, organized into two scenarios, to assess the model's classification capabilities thoroughly. In the binary classification scenario, the model distinguishes between a no-task baseline and a defined cognitive task, known as the 'No vs SIMKAP task.' This scenario is critical for evaluating the model's accuracy in differentiating resting cognitive states from those involved in a SIMKAP task. The more complex ternary classification scenario introduces a gradient of mental workload levels-classified as low, medium, and high, together referred to as 'SIMKAP multitask.' This detailed evaluation demonstrates the model's proficiency in differentiating among varied cognitive loads, mirroring real-world conditions as depicted by the multidimensional data of the STEW dataset.

Table 1 presents a comparative analysis of five classifiers: SVM, EEGNet, TSception, Transformer, and the proposed ConvNeXt model, across the two classification scenarios. The binary task involves two classes, while the multi-class task includes three. This table not only highlights the performance metrics of each classifier, presumably in accuracy percentages with 95% confidence interval,

underscoring variability in performance across different trials or datasets. This variability is crucial for evaluating the robustness and reliability of the models in real-world applications. Moreover, Table 1 emphasizes the effectiveness of advanced machine learning and neural network models in interpreting EEG data for mental workload classification. Notably, TSception and the proposed method exhibit exceptional promise, as indicated by their strong performance metrics.

The results demonstrate the variability in performance metrics of different classifiers evaluated on the STEW dataset. The SVM classifier shows moderate effectiveness, achieving 83.34% (±0.39) accuracy in the binary classification task and 83.21% (±0.28) in the multi-class task. EEGNet exhibits slightly better performance, with 84.33% (±1.03) accuracy for the binary task and 83.64% (±1.33) for the multi-class task. TSception, a more advanced classifier, significantly outperforms the other models with impressive accuracies of 95.21% (±0.53) in the binary classification and 94.73% (±0.30) in the multi-class scenario. The Transformer model achieves high performance in the binary task with an accuracy of 95.32% (±0.00) but shows a drop in performance for the multi-class task with an accuracy of 88.72% (±0.00). The proposed model records the highest accuracies of 95.76% (±0.51) in the binary task and 95.11% (±0.80) in the multi-class task, demonstrating superior capability in managing the classification challenges of the STEW dataset.

These results underscore the comparative strengths and potential real-world applicability of advanced neural network-based classifiers, particularly TSception and the proposed model, in complex EEG data classification tasks. A detailed boxplot presented in Figure 2 illustrates the distribution of classifier accuracies across multiple trials, providing a visual performance comparison. While TSception closely rivals the proposed model's effectiveness, it still performs exceptionally well, especially in EEG-based workload classification. Although outperformed by the more advanced classifiers, SVM maintains strong and reliable performance across both tasks. Similarly, EEGNet, despite being the least accurate of the tested classifiers, still holds its ground, particularly in the multi-class scenario. This visual evidence, as depicted in the boxplot, highlights the importance of choosing an appropriate classifier based on the specific requirements and complexities of the task, demonstrating that newer or more sophisticated models like TSception or the proposed model can offer substantial benefits in certain scenarios. The accuracy of the Transformer model is derived from the respective research paper that utilized the same dataset. However, it did not include the confidence interval, showing it as ±0.00.

Overall, the proposed model achieves the highest accuracy and robustness across both classification tasks, affirming its superiority in addressing the challenges presented by the 'No vs SIMKAP' and 'SIMKAP multi-task' scenarios. These findings represent a pivotal advancement in non-invasive methods for quantifying mental workload, providing insights that could improve real-time cognitive state assessment and enhance safety, efficiency, and user experience in high-demand settings. The efficacy of the ConvNeXt model, tailored for EEG signal analysis, underscores the potential of advanced machine learning classifiers

in this field. Moving forward, this study sets the stage for further innovations in mental workload assessment using EEG data, aiming to refine and expand the applicability of these techniques in both theoretical and practical contexts.

By elucidating the strengths and limitations of current methodologies, this research also suggests directions for future efforts to improve and innovate in the area of mental workload estimation. The comparative analysis across different models aids in understanding the potential of neural network-based approaches for real-world applications, inspiring continued exploration and development in cognitive neuroscience and related fields.

6 Conclusion

This study assessed the classification capabilities of advanced neural network-based classifiers using the STEW dataset for binary and multi-class scenarios, aiming to evaluate each model's precision in distinguishing between a no-task baseline and SIMKAP tasks for neuroscientific applications. The proposed model displayed exceptional accuracy, demonstrating its potential for real-world neuroscientific settings, such as real-time mental workload assessments in critical sectors like air traffic management and healthcare. This could significantly enhance safety and efficiency.

The findings mark a significant advancement in mental workload estimation through EEG analysis, achieved using a customized ConvNeXt model. This model accomplished classification accuracies of 95.76% for the binary (No vs SIMKAP task) and 95.11% for the ternary (SIMKAP multitask) classes, showcasing the potential of deep learning in the sophisticated interpretation of EEG data. These results contribute substantially to fields like cognitive neuroscience and human-computer interaction.

Looking ahead, this study encourages further research to address current limitations by expanding dataset diversity and exploring broader practical applications. By reinforcing the importance of selecting appropriate classifiers and advancing machine learning in neuroscientific research, this work sets the stage for future studies to create more intuitive and adaptive interfaces, meeting the complex needs of diverse sectors. Further investigations should validate and potentially broaden the applicability of these models in practical scenarios.

References

1. Aldawsari, H., Al-Ahmadi, S., Muhammad, F.: Optimizing 1d-cnn-based emotion recognition process through channel and feature selection from eeg signals. Diagnostics **13**(16), 2624 (2023)

2. Aznan, N.K.N., Atapour-Abarghouei, A., Bonner, S., Connolly, J.D., Breckon, T.P.: Leveraging synthetic subject invariant eeg signals for zero calibration bci. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 10418–10425. IEEE (2021)
3. Chakladar, D.D., Dey, S., Roy, P.P., Iwamura, M.: Eeg-based cognitive state assessment using deep ensemble model and filter bank common spatial pattern. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 4107–4114. IEEE (2021)
4. Cheng, S., Hao, Y.: 3d-clmi: A motor imagery eeg classification model via fusion of 3d-cnn and lstm with attention. arXiv preprint arXiv:2312.12744 (2023)
5. Chin, Z.Y., Zhang, X., Wang, C., Ang, K.K.: Eeg-based discrimination of different cognitive workload levels from mental arithmetic. In: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). pp. 1984–1987. IEEE (2018)
6. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**, 273–297 (1995)
7. Craik, A., He, Y., Contreras-Vidal, J.L.: Deep learning for electroencephalogram (eeg) classification tasks: a review. J. Neural Eng. **16**(3), 031001 (2019)
8. Das Chakladar, D., Roy, P.P.: Cognitive workload estimation using physiological measures: A review. Cogn. Neurodyn. **18**(4), 1445–1465 (2024)
9. Di Flumeri, G., Borghini, G., Aricò, P., Sciaraffa, N., Lanzi, P., Pozzi, S., Vignali, V., Lantieri, C., Bichicchi, A., Simone, A., et al.: Eeg-based mental workload neurometric to evaluate the impact of different traffic and road conditions in real driving settings. Front. Hum. Neurosci. **12**, 509 (2018)
10. Ding, Y., Robinson, N., Zhang, S., Zeng, Q., Guan, C.: TSception: Capturing temporal dynamics and spatial asymmetry from EEG for emotion recognition. IEEE Transactions on Affective Computing (2022)
11. Gupta, A., Siddhad, G., Pandey, V., Roy, P.P., Kim, B.G.: Subject-specific cognitive workload classification using eeg-based functional connectivity and deep learning. Sensors **21**(20), 6710 (2021)
12. Hernández-Sabaté, A., Yauri, J., Folch, P., Piera, M.À., Gil, D.: Recognition of the mental workloads of pilots in the cockpit using eeg signals. Appl. Sci. **12**(5), 2298 (2022)
13. Kartali, A., Janković, M.M., Gligorijević, I., Mijović, P., Mijović, B., Leva, M.C.: Real-time mental workload estimation using eeg. In: Human Mental Workload: Models and Applications: Third International Symposium, H-WORKLOAD 2019, Rome, Italy, November 14–15, 2019, Proceedings 3. pp. 20–34. Springer (2019)
14. Kostas, D., Aroca-Ouellette, S., Rudzicz, F.: Bendr: using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data. Front. Hum. Neurosci. **15**, 653659 (2021)
15. Lawhern, V.J., Solon, A.J., Waytowich, N.R., Gordon, S.M., Hung, C.P., Lance, B.J.: EEGNet: a compact convolutional neural network for EEG-based brain-computer interfaces. J. Neural Eng. **15**(5), 056013 (2018)
16. Lim, W.L., Sourina, O., Liu, Y., Wang, L.: Eeg-based mental workload recognition related to multitasking. In: 2015 10th International Conference on Information, Communications and Signal Processing (ICICS). pp. 1–4. IEEE (2015)
17. Lim, W.L., Sourina, O., Wang, L.P.: Stew: Simultaneous task eeg workload data set. IEEE Trans. Neural Syst. Rehabil. Eng. **26**(11), 2106–2114 (2018)
18. Liu, F., Zhang, Y., Kuai, H., Stephen, E.P., Wang, H.: Brain Informatics: 16th International Conference, BI 2023, Hoboken, NJ, USA, August 1–3, 2023, Proceedings, vol. 13974. Springer Nature (2023)

19. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
20. Mastropietro, A., Pirovano, I., Marciano, A., Porcelli, S., Rizzo, G.: Reliability of mental workload index assessed by eeg with different electrode configurations and signal pre-processing pipelines. Sensors **23**(3), 1367 (2023)
21. Pandey, V., Choudhary, D.K., Verma, V., Sharma, G., Singh, R., Chandra, S.: Mental workload estimation using eeg. In: 2020 Fifth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN). pp. 83–86. IEEE (2020)
22. Parveen, F., Bhavsar, A.: Attention based 1d-cnn for mental workload classification using eeg. In: Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments. pp. 739–745 (2023)
23. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
24. Qu, H., Zhang, M., Pang, L.: Mental workload classification method based on eeg cross-session subspace alignment. Mathematics **10**(11), 1875 (2022)
25. Radüntz, T., Fürstenau, N., Mühlhausen, T., Meffert, B.: Indexing mental workload during simulated air traffic control tasks by means of dual frequency head maps. Front. Physiol. **11**, 300 (2020)
26. Rojas, A., Kroupi, E., Martens, G., Thibaut, A., Barra, A., Laureys, S., Ruffini, G., Soria-Frisch, A.: Prediction of minimally conscious state responder patients to non-invasive brain stimulation using machine learning algorithms. In: Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part I. pp. 515–525. Springer (2021)
27. Saleh, M., Paquelet, S., Castel, P., Hoarau, M., Pallamin, N., Lewkowicz, D.: An efficient deep-learning-based solution for the recognition of relative changes in mental workload using wearable sensors. In: 2023 IEEE SENSORS. pp. 1–4. IEEE (2023)
28. Samima, S., Sarma, M.: Eeg-based mental workload estimation. In: 2019 41st Annual international conference of the IEEE engineering in medicine and biology society (EMBC). pp. 5605–5608. IEEE (2019)
29. Siddhad, G., Gupta, A., Dogra, D.P., Roy, P.P.: Efficacy of transformer networks for classification of eeg data. Biomed. Signal Process. Control **87**, 105488 (2024)
30. Siddhad, G., Iwamura, M., Roy, P.P.: Enhancing eeg signal-based emotion recognition with synthetic data: Diffusion modeel approach. arXiv preprint arXiv:2401.16878 (2024)
31. Singh, U., Ahirwal, M.K.: Mental workload classification for multitasking test using electroencephalogram signal. In: 2021 IEEE International Conference on Technology, Research, and Innovation for Betterment of Society (TRIBES). pp. 1–6. IEEE (2021)
32. Singh, U., Ahirwal, M.K.: Improved classification of mental workload using one dimensional convolutional neural network and smote technique. In: Proceedings of the 2023 7th International Conference on Graphics and Signal Processing. pp. 50–55 (2023)
33. So, W.K., Wong, S.W., Mak, J.N., Chan, R.H.: An evaluation of mental workload with frontal eeg. PLoS ONE **12**(4), e0174949 (2017)

34. Vidaurre, C., Sander, T.H., Schlögl, A.: Biosig: the free and open source software library for biomedical signal processing. Computational intelligence and neuroscience **2011** (2011)
35. Yao, X., Li, T., Ding, P., Wang, F., Zhao, L., Gong, A., Nan, W., Fu, Y.: Emotion classification based on transformer and cnn for eeg spatial-temporal feature learning. Brain Sci. **14**(3), 268 (2024)

Awake at the Wheel: Enhancing Automotive Safety Through EEG-Based Fatigue Detection

Gourav Siddhad[1(✉)], Sayantan Dey[1], Partha Pratim Roy[1], and Masakazu Iwamura[2]

[1] Department of Computer Science and Engineering, Indian Institute of Technology, Roorkee, Roorkee 247667, Uttarakhand, India
{g_siddhad,partha}@cs.iitr.ac.in, sayantan.cs@sric.iitr.ac.in
[2] Department of Core Informatics, Graduate School of Informatics, Osaka Metropolitan University, Sakai, Osaka 599-8531, Japan
masa.i@omu.ac.jp

Abstract. Driver fatigue detection is increasingly recognized as critical for enhancing road safety. This study introduces a method for detecting driver fatigue using the SEED-VIG dataset, a well-established benchmark in EEG-based vigilance analysis. By employing advanced pattern recognition technologies, including machine learning and deep neural networks, EEG signals are meticulously analyzed to discern patterns indicative of fatigue. This methodology combines feature extraction with a classification framework to improve the accuracy of fatigue detection. The proposed NLMDA-Net reached an impressive accuracy of 83.71% in detecting fatigue from EEG signals by incorporating two novel attention modules designed specifically for EEG signals, the channel and depth attention modules. NLMDA-Net effectively integrate features from multiple dimensions, resulting in improved classification performance. This success stems from integrating temporal convolutions and attention mechanisms, which effectively interpret EEG data. Designed to capture both temporal and spatial characteristics of EEG signals, deep learning classifiers have proven superior to traditional methods. The results of this study reveal a substantial enhancement in detection rates over existing models, highlighting the efficacy of the proposed approach for practical applications. The implications of this research are profound, extending beyond academic realms to inform the development of more sophisticated driver assistance systems. Incorporating this fatigue detection algorithm into these systems could significantly reduce fatigue-related incidents on the road, thus fostering safer driving conditions. This paper provides an exhaustive analysis of the dataset, methods employed, results obtained, and the potential real-world applications of the findings, aiming to contribute significantly to advancements in automotive safety.

Keywords: Driver Fatigue · EEG · Safety · SEED-VIG · Vigilance

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15311, pp. 340–353, 2025.
https://doi.org/10.1007/978-3-031-78195-7_23

1 Introduction

Enhancing road safety through effectively managing driver fatigue is paramount in the automotive industry, given its significant role in global road accidents. This prevalent issue impairs cognitive and motor functions, diminishing a driver's alertness and responsiveness to changing road conditions [8]. In light of these concerns, this study focuses on fatigue detection using advanced computational techniques applied to electroencephalogram (EEG) signals, a direct method has shown promise over traditional indirect methods such as monitoring steering wheel movements or analyzing eyelid closures.

Recent statistics indicate that driver fatigue is implicated in about 20% of road accidents, underscoring the essential need for effective detection systems in modern vehicles [26]. Unlike traditional approaches, which often result in delayed fatigue detection, EEG-based methods allow for real-time, accurate assessments by directly measuring neurological activity. These techniques utilize the distinct capabilities of EEG signals to mirror neurophysiological changes linked to fatigue, capturing specific brain wave patterns such as theta and alpha waves. This enables a precise evaluation of a driver's vigilance levels, which is unachievable through other methods [44]. Moreover, recent research has demonstrated the feasibility of decoding cognitive states such as attention and distraction in a real-life setting using EEG [19]. This suggests that EEG-based systems could potentially be used to identify a wider range of driver states, including those that may contribute to accidents beyond fatigue.

This study employs the SEED-VIG dataset [42], renowned for its application in EEG-based vigilance estimation, facilitating the accurate examination of signals pertinent to real-world driving situations. The research enhances pattern recognition methods for robust feature extraction and effective classification of fatigue states by integrating traditional machine learning algorithms with deep neural networks. This dual approach significantly improves the accuracy and reliability of fatigue detection systems, effectively overcoming the constraints of existing models.

This paper introduces a unified lightweight NLMDA-Net to facilitate relevant feature extraction from complex EEG signals with the help of multi-dimensional attention modules. The contributions of this paper are as follows:

- A lightweight network, NLMDA-Net, is proposed for driver fatigue detection using EEG data. It comprises the feature extraction capabilities of ConvNet and EEGNet.
- Channel Attention Module: The module leverages a tensor product to expand channel information into the depth dimension, enhancing the network's ability to process and analyze spatial features in EEG signals. This innovation increases sensitivity to spatial variations. Furthermore, the tanh function, a non-linear activation mechanism, stabilizes the learning process by normalizing amplitude variability. Its properties prevent the dying gradient problem and facilitate the capture of bi-directional relationships, which is essential for focusing the attention mechanism on the most informative EEG features.

- Parameter Efficiency: By reducing the number of convolution kernels as the network depth increases, NLMDA-Net tailors its architecture better to suit the predominant information-rich time domain of EEG signals, enhancing the network's efficiency and effectiveness.
- Adaptation to Data Scarcity: The network design is particularly suited for scenarios with limited EEG data, preventing over-fitting and accommodating EEG's low spatial resolution characteristics.

The structure of this paper is designed to methodically explore EEG-based fatigue detection and its implications for enhancing road safety technologies. Section 2 reviews recent literature on driver drowsiness and vigilance. Section 3 explains the methodology employed. Section 4 presents the empirical findings. The paper is concluded in Section 5, where the discussion extends to this research's implications and future directions.

2 Related Work

Early research in EEG-based fatigue detection has primarily focused on identifying fatigue-associated biomarkers, such as the theta and alpha EEG frequency bands, such as variations in theta and alpha EEG frequency bands [14,16]. Subsequent advancements have introduced sophisticated signal processing techniques to improve detection accuracy, incorporating wavelet transforms and power spectral density analysis [1,5]. The advent of deep learning has significantly transformed EEG analysis. In particular, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have become increasingly prevalent, appreciated for their adept handling of spatial and temporal data, respectively [3,32].

Recent efforts have seen the development of hybrid models that combine CNNs with RNNs or other machine learning techniques to capitalize on their spatial and temporal feature extraction capabilities [2,40]. Comparative studies of deep learning architectures indicate that CNNs provide superior accuracy and enhance computational efficiency, rendering them ideal for real-time applications [38]. Deep learning models generally surpass traditional machine learning methods due to their enhanced capacity to manage large, complex datasets without extensive feature engineering [10,37].

Furthermore, the application of transfer learning with pre-trained models on EEG data has demonstrated potential in mitigating the challenges posed by the need for large labelled datasets, which are often a limiting factor in EEG research [31]. Additionally, recent advancements in synthetic data generation [34] also offer potential to augment real-world datasets and improve model performance. Transformers and attention mechanisms have emerged as powerful tools for EEG analysis [20,33]. Furthermore, recent studies have explored the integration of attention mechanisms into deep neural networks to enhance the identification of fatigue-related EEG features [11,25]. However, challenges remain, such as the variability in EEG signals across individuals, which can affect model generalization [18]. Additionally, the presence of artifacts in EEG data due to head movements or external electrical interference continues to be a significant issue, potentially compromising the effectiveness of fatigue detection systems [21].

3 Proposed Methodology: NLMDA-Net

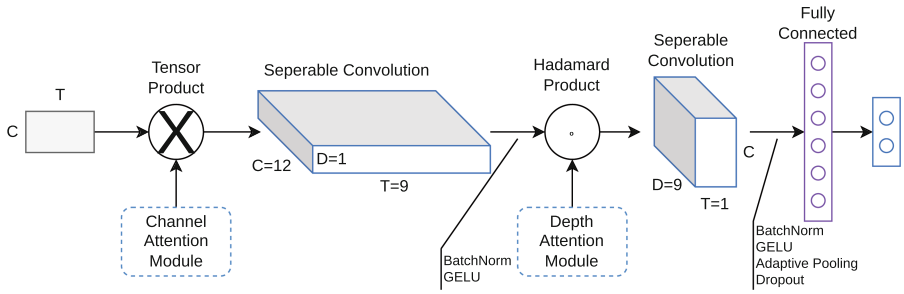


Fig. 1. Architecture of NLMDA-Net: It comprises the benchmark network, the channel attention module, and the depth attention module. Channel attention is facilitated through a tensor product, and depth attention is achieved through a Hadamard product of tensors. NLMDA-Net's output varies with the number of task categories. Notably, C represents the number of EEG channels, and T signifies the time samples in a trial.

The NLMDA-Net architecture shown in Figure 1 represents a novel integration of benchmark network capabilities enhanced by two specialized attention modules: the channel attention and the depth attention modules. The channel attention module is designed to strengthen the network's ability to discern relevant information within the spatial dimensions of EEG signals. Complementarily, the depth attention module aims to refine the representation of high-dimensional EEG features, ensuring a deeper and more targeted analysis. These modules are strategically developed to be compatible with any existing convolutional neural network structures.

As illustrated in Figure 1, the conceptual foundation of NLMDA-Net is intricately tied to the fundamental properties of EEG signals. Contrary to the prevailing trend in deep learning, which favours increasingly complex architectures [12,22,35,36], insights from neuroscience [17,28,29] advocate for the inherent simplicity of EEG characteristics. This simplicity suggests that even a shallower network architecture might suffice for effective EEG signal decoding.

Moreover, the typical scarcity of EEG data combined with the high data requirements of neural network models highlights the necessity for a more streamlined and lightweight network design. Such a design is essential to prevent over-fitting and accommodate the low spatial resolution of EEG and the diverse informational content across its temporal and spatial dimensions. The tailored approach provided by NLMDA-Net, depicted in Figure 1, is particularly suited to these unique challenges.

The design considerations of NLMDA-Net and its compatibility with existing architectures are further detailed in the subsequent sections, providing a comprehensive overview of each component within the NLMDA-Net framework. This structured exposition underscores the architecture's potential to improve

EEG-based applications through focused attention mechanisms and simplified network design.

3.1 Benchmark Network

The NLMDA-Net architecture merges the foundational feature extraction capabilities of ConvNet [30] with the advanced separable convolution technique from EEGNet [23], optimizing the extraction of temporal and spatial features from EEG signals. This architecture strategically employs a two-layer convolutional setup comprising a temporal convolutional layer and a spatial convolutional layer, utilizing separable convolutions to reduce the network's parameter count effectively.

In this approach, the temporal convolutional layer is characterized by a kernel size of (12, 1, 9), where 12 denotes the number of kernels, and the kernels' spatial and temporal dimensions are one and nine, respectively. Conversely, the spatial convolutional layer employs a kernel size of (7, C, 1), with C representing the number of EEG channels. These consistent kernel dimensions are upheld throughout all experiments conducted within the NLMDA-Net framework, ensuring a standardized assessment of the network's efficacy in extracting features from EEG data.

Moreover, NLMDA-Net employs the Gaussian Error Linear Unit (GELU) activation function, as introduced by Hendrycks et al. [13], offering improved smoothness compared to the Exponential Linear Unit (ELU) [6] utilized in previous models such as EEGNet and ConvNet. To effectively handle the typically substantial number of parameters required by fully connected layers, NLMDA-Net integrates adaptive average pooling. This technique dynamically adjusts the pooling kernel size to $(1, k_{pooling})$, where $k_{pooling}$ is detailed in Equation (1). This adaptive approach ensures efficient parameter management while preserving the network's feature extraction capabilities.

$$k_{pooling} = max(1, \lfloor f/10/N \rfloor) \tag{1}$$

where, the operator $\lfloor \rfloor$ represents the mathematical operation of rounding down to the nearest integer. It's applied concerning the input signal frequency, denoted by f, where N signifies the number of training samples. The determination of N is intricately tailored to accommodate the intricacies of EEG data collection, as elucidated by Equation (2). This method ensures the parameters are finely tuned to align with the dataset's unique characteristics.

$$N = max(1, \lfloor N_t/200 \rfloor) \tag{2}$$

where, N_t denotes the number of training samples, a pivotal parameter crucial for quantifying the dataset size utilized in model training. Its significance lies in its direct impact on the robustness and generalizability of the learned features.

NLMDA-Net distinctively adapts its architecture by reducing the number of convolution kernels from 12 to 1 as the network depth increases, a decision driven by two primary considerations. Firstly, employing a higher number of kernels in

the spatial layers risks rapid over-fitting, leading to an exponential increase in the parameters of the fully connected layers, potentially compromising the network's capability to extract meaningful features. Secondly, considering that EEG signals predominantly contain richer information in the time domain than the spatial domain, it is pragmatic to allocate more kernels for extracting time domain features. This approach enhances the network's ability to effectively capture the most relevant data.

3.2 Channel Attention Module

In EEG data acquisition, the signal captured by a single electrode channel is a composite of various neuronal activities influenced by volume conduction effects. Certain studies have employed source reconstruction techniques to enhance the spatial resolution of EEG signals to identify neuronal regions associated with specific EEG activities for in-depth analysis [4,15]. However, these techniques often require extensive prior knowledge and face integration challenges with end-to-end neural network models, complicating the decoding of EEG signals across different paradigms.

In the context of neural network architecture, models tailored explicitly for EEG decoding, such as EEGNet [23], ConvNet [30], and DRDA [43], typically prioritize temporal convolutions over spatial ones. This approach can result in a relative neglect of spatial dimensions within EEG signals. To bridge this gap, a novel channel attention module is introduced that enhances the neural network's ability to assimilate spatial information from EEG data. This module draws conceptual parallels with source reconstruction techniques, acting on the input data to expand its spatial dimensions into the depth dimension through a Tensor product. This adaptation aims to improve the processing and analysis of spatial features within EEG signals, enhancing the overall efficacy of the neural network in decoding complex EEG data.

In this model, consider an EEG input sample denoted by x, where $x \in \mathbb{R}^{1*C*T}$. Here, C represents the number of channels, and T denotes the time samples. Additionally, a tensor c is introduced, which follows a normal distribution, defined as $c \in \mathbb{R}^{D*1*C}$, where D corresponds to both the number of instances and convolutional kernels.

The channel attention module in the architecture utilizes a tensor product to project channel information from x into the depth dimension, D. This operation preserves the spatial information inherent in the channel data and effectively integrates it with the following temporal convolution. The mathematical representation of this operation is outlined in the subsequent formula, illustrating how the module enhances the feature extraction capabilities of the neural network by augmenting the depth dimension with significant channel-specific information. This approach facilitates better analysis of EEG signals by leveraging both spatial and temporal dimensions efficiently. The operation is mathematically represented as:

$$X'_{hct} = \sum_{d} \mathbf{X}_{dct} \mathbf{C}_{hdc} \tag{3}$$

In the employed channel attention module, subscripts in the formula denote the respective dimensions, with matching subscript letters indicating that the two tensors share the same dimensionality in that specified dimension. This module introduces $D \times C$ trainable parameters, where D is a hyper-parameter that can be optimized for specific tasks to enhance performance potentially. However, to maintain a consistent basis for comparison with NLMDA-Net and other benchmark models, D is fixed at 9. This setting introduces significantly fewer parameters than traditional models and effectively maps spatial information into the depth dimension. This strategic approach offers a novel perspective on the attention mechanisms utilized for EEG signals.

The transformation of the input X and its subsequent processing through the neural network's sequential components involves reshaping and applying linear layers. The initial step in this transformation process features a linear transformation followed by a non-linear activation, specifically using the *tanh* function. This methodological choice facilitates the efficient integration and processing of EEG data, emphasizing the innovative use of attention mechanisms to enhance the depth dimension's role in spatial feature representation. The first linear transformation with a non-linear activation *tanh* is represented as

$$E = W_2 * tanh(W_1 * X + b_1) \tag{4}$$

$$\alpha = softmax(E) = \frac{\exp(E)}{\sum \exp(E)} \tag{5}$$

$$\text{Context}_{b,c,n,t} = \alpha_{b,c,n} \cdot X_{b,c,t} \tag{6}$$

3.3 Effect of Non-Linearity

The hyperbolic tangent (tanh) function, renowned for its output range of [-1, 1], emerges as a potent normalization tool for processing EEG signals. These signals exhibit substantial amplitude fluctuations across diverse recording conditions and subjects. Normalizing such variations with tanh fosters stable learning dynamics, ensuring consistent neural network performance across heterogeneous datasets. Tanh's hallmark smoothness and continuous nature, coupled with a non-zero derivative across its operational span, are pivotal in facilitating gradient flow during backpropagation. This characteristic mitigates the risk of encountering the 'dying gradient problem' prevalent in rectified linear units (ReLU), where gradients may diminish to zero, impeding further learning.

Furthermore, tanh's saturation at the extremities of its range offers resilience against outliers and extreme values in the data, facilitating more robust convergence during training. In contrast to linear activation functions like ReLU, tanh's capacity to yield positive and negative outputs enables the model to capture bi-directional data relationships effectively. This bi-directionality proves particularly advantageous in the context of the channel attention module within neural

networks. Here, tanh synergizes with the softmax function, a staple in attention mechanisms, enhancing the latter's efficacy in spotlighting the most salient features in EEG data. The diverse output range of tanh empowers softmax to operate across a broad spectrum of values, thereby augmenting the attention mechanism's ability to emphasize informative data attributes.

3.4 Depth Attention Module

In computer vision, feature maps in the depth dimension are often regarded as detectors of specific features within an input, identifying 'what' is meaningful ([39,41]. This principle is crucial in models such as the Convolutional Block Attention Module (CBAM), where depth attention aggregates depth information through global pooling and dense layers to refine feature focus[39]. However, this methodology proves less effective when decoding EEG signals due to the distinct nature of spatial and temporal dimensions in EEG data. In EEG, the global pooling and subsequent fully connected layers tend to oversimplify the depth information and drastically increase the model parameters, potentially leading to overfitting and degraded performance of the base network. A specialized depth attention module is proposed, tailored for EEG decoding to address these challenges. This module integrates concepts from local cross-depth interaction techniques, effectively balancing parameter efficiency and depth feature utilization.

The depth attention module is strategically positioned between the temporal and spatial convolution layers, encompassing three main components: Semi-Global Pooling, Local Cross-Depth Interaction, and Adaptive Weighting. In contrast to conventional global pooling methods, Semi-Global Pooling averages the spatial dimensions while retaining temporal details, thus preserving a more comprehensive representation of depth features. Following this pooling, a convolutional layer is employed to encourage local interactions among features, substantially lowering the trainable parameters' count relative to fully connected layers. Subsequently, features undergo adaptive weighting and are transformed into probabilistic values via a softmax function. To maintain the amplitude sensitivity crucial for EEG signals, these softmax outputs are amplified to the level of the original inputs using a Hadamard product.

$$M(F) = (\text{Softmax}\left(\text{Conv}\left(\text{Pooling} * (F)^T\right)\right)^* D^{'})^T \tag{7}$$

where, $F \in \mathbb{R}^{D_o * C_o * T_o}$ represents the input feature tensor, capturing the dimensions of depth (D_o), channels (C_o), and temporal sequence (T_o) and $M(F) \in \mathbb{R}^{D_o * 1 * T_o}$ denotes the output feature map, simplifying the channels to one while maintaining depth and time dimensions. Pooling refers to the semi-global pooling operation, Conv indicates the convolution layer, and T represents the transpose operation applied to the spatial and depth dimensions of the tensor.

4 Results and Discussion

4.1 Experimental Data

The SEED-VIG dataset [42] is an open-source resource for investigating vigilance and driver drowsiness through EEG data collected from 23 participants to ensure diverse subject representation. Participants underwent a driving simulation resembling real-world conditions, enhancing the dataset's applicability for drowsiness studies. EEG recordings utilized 17 channels based on the 1020 system, covering key temporal and posterior regions (FT7, FT8, T7, T8, TP7, TP8 for temporal; CP1, CP2, P1, PZ, P2, PO3, POZ, PO4, O1, OZ, O2 for posterior), ensuring comprehensive brain activity capture. Recorded at 1000Hz, the dataset offers high temporal resolution for detailed vigilance and drowsiness analysis. Fatigue induction was optimized by scheduling sessions post-lunch.

The drowsiness states are calculated as a percentage of eye closure time per unit time (PERCLOS). PERCLOS were categorized into 'awake' and 'drowsy' states at a 0.5 threshold. This binary classification enabled precise evaluation of this method's ability to detect driver fatigue. EEG signals are band-pass filtered between 1-75 Hz to reduce artifacts and down-sampled with a sampling frequency of 200 Hz. The dataset was epoched into one-second intervals, resulting in the shape of (1, channel count, EEG length), i.e., (1, 17, 200), yielding 40710 samples. The dataset is split into 70:15:15 ratios for train, validation, and test sets.

4.2 Implementation Details

The experimental setup involved a DELL Precision 7820 Tower Workstation with Ubuntu 22.04 OS, Intel Core(TM) Xeon Silver 4216 CPU, and an NVIDIA RTX A2000 12GB GPU. This hardware facilitated the implementation of DL models using Python 3.10 and the PyTorch library. The Adam optimizer, known for its computational efficiency, was used with default parameters ($\eta = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$). EEGNet and TSception were trained for 100 epochs, with batches of 16 and a learning rate of $1e-4$. The Radial Basis Function (RBF) kernel from scikit-learn [27] was used with default settings for SVM. Classification accuracy was determined through stratified five-fold cross-validation, averaging the results for comprehensive assessment.

4.3 Evaluation

The data in Table 1 compares various machine learning classifiers, analyzing their accuracy in detecting driver fatigue using the SEED-VIG dataset. This study encompasses a range of classifiers, each employing unique approaches and architectures designed to process and predict based on EEG data.

The SVM classifier exhibits the lowest accuracy at 65.52%, suggesting its linear operational nature may be less effective at interpreting the complex patterns present in EEG signals, in contrast to more sophisticated, non-linear models. On

Table 1. Comparison of Classifier Performance for Detecting Driver Drowsiness Using SEED-VIG Dataset, Shown with 95% Confidence Intervals

Classifier	Accuracy
SVM [7]	65.52 ± 0.02
EEGNet [23]	80.74 ± 0.75
TSception [9]	83.15 ± 0.36
ConvNext [24]	81.95 ± 0.61
LMDA [25]	81.06 ± 0.99
Proposed NLMDA-Net	$\mathbf{83.71 \pm 0.30}$

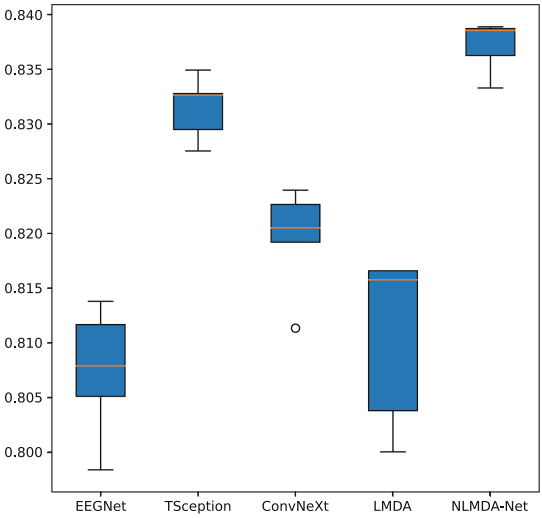


Fig. 2. Boxplot Illustrating the Distribution of Classifier Accuracies for Driver Drowsiness Detection on the SEED-VIG Dataset

the other hand, EEGNet, a neural network specifically optimized for EEG data processing, achieves an improved accuracy of 80.74%. Its architecture, which adeptly handles both spatial and temporal dynamics of EEG signals, significantly outperforms traditional machine learning models like SVM.

Further analysis reveals that TSception and NLMDA-Net demonstrate the highest accuracies, with scores of 83.15% and 83.71%, respectively. These models incorporate advanced features such as temporal convolutions and attention mechanisms, enhancing their capability to capture subtle EEG signal changes associated with fatigue. ConvNext and LMDA also perform commendably, with accuracies of 81.95% and 81.06%, respectively. These classifiers benefit from recent advancements in convolutional network design and machine learning techniques tailored to handle large-scale, complex data structures typical of EEG datasets.

The results from Table 1 and boxplot from Figure 2 consistently indicate that classifiers utilizing deep learning architectures, particularly those incorporating specialized mechanisms for extracting temporal and spatial features, surpass traditional machine learning methods in performance. This trend underscores the advantages of models that can adaptively learn from the intrinsic characteristics of EEG data related to drowsiness, suggesting a strategic direction for future development in this area. The confidence intervals reported also provide valuable insight into the consistency of each model's performance across different experimental setups, reinforcing the reliability of these findings.

5 Conclusion

This study addresses the critical issue of driver fatigue by applying advanced computational techniques to EEG signals. This provides a real-time, direct method for fatigue detection, surpassing traditional methods like monitoring steering movements or eyelid activity. Utilizing the SEED-VIG dataset, the research combines traditional machine learning and deep neural networks to refine pattern recognition techniques, enhancing the detection systems' accuracy and reliability. The results, methodology, and relevant literature are thoroughly explored, leading to discussions on the implications and future directions for enhancing road safety through improved fatigue detection technologies. Adopting deep learning, particularly CNNs and RNNs, has revolutionized EEG analysis by effectively handling spatial and temporal data, with hybrid models enhancing real-time feature extraction. Deep learning excels over traditional methods by managing large datasets with minimal feature engineering. Transfer learning and attention mechanisms have also emerged as solutions to challenges such as data variability and artifacts, improving signal quality and model generalizability. However, refining fatigue detection systems' accuracy and broad applicability remains a challenge.

This study uses the SEED-VIG dataset to evaluate the efficacy of various machine-learning classifiers in detecting driver drowsiness. The results reveal that deep learning models, especially NLMDA-Net, show superior performance, achieving accuracy up to 83.71%. These models excel due to their advanced features, such as temporal convolutions and attention mechanisms, effectively capturing EEG signal differences associated with fatigue. This suggests a significant potential for deep learning approaches to enhance fatigue detection systems, advocating for a strategic pivot towards these technologies to improve the accuracy and generalizability of drowsiness detection methods. The consistency of model performances, supported by confidence intervals, reinforces the reliability of these findings. Future research in EEG-based fatigue detection should prioritize advancements that bolster accuracy and usability. Integrating multimodal data, including heart rate variability, eye tracking, and contextual driving information, can enrich the understanding of the driver's state, facilitating a more comprehensive analysis.

References

1. Ahmadi, A., Bazregarzadeh, H., Kazemi, K.: Automated detection of driver fatigue from electroencephalography through wavelet-based connectivity. Biocybernetics and Biomedical Engineering **41**(1), 316–332 (2021)
2. Ardabili, S.Z., Bahmani, S., Lahijan, L.Z., Khaleghi, N., Sheykhivand, S., Danishvar, S.: A novel approach for automatic detection of driver fatigue using eeg signals based on graph convolutional networks. Sensors **24**(2), 364 (2024)
3. Balam, V.P., Sameer, V.U., Chinara, S.: Automated classification system for drowsiness detection using convolutional neural network and electroencephalogram. IET Intel. Transport Syst. **15**(4), 514–524 (2021)
4. Cao, W., Fang, Z., Hou, G., Han, M., Xu, X., Dong, J., Zheng, J.: The psychological impact of the covid-19 epidemic on college students in china. Psychiatry Res. **287**, 112934 (2020)
5. Chinara, S., et al.: Automatic classification methods for detecting drowsiness using wavelet packet transform extracted time-domain features from single-channel eeg signal. J. Neurosci. Methods **347**, 108927 (2021)
6. Clevert, D.A., Unterthiner, T., Hochreiter, S.: Fast and accurate deep network learning by exponential linear units (elus). arXiv preprint arXiv:1511.07289 (2015)
7. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**, 273–297 (1995)
8. Delvigne, V., Wannous, H., Vandeborre, J.P., Ris, L., Dutoit, T.: Spatio-temporal analysis of transformer based architecture for attention estimation from eeg. In: 2022 26th International Conference on Pattern Recognition (ICPR). pp. 1076–1082. IEEE (2022)
9. Ding, Y., Robinson, N., Zhang, S., Zeng, Q., Guan, C.: TSception: Capturing temporal dynamics and spatial asymmetry from EEG for emotion recognition. IEEE Transactions on Affective Computing (2022)
10. Fouad, I.A.: A robust and efficient eeg-based drowsiness detection system using different machine learning algorithms. Ain Shams engineering journal **14**(3), 101895 (2023)
11. Gao, K., Jia, W., Zhou, Y., Du, R.: Multi-head self-attention enhanced convolutional neural network for driver fatigue detection using eeg signals. In: 2023 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom). pp. 817–823. IEEE (2023)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
13. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)
14. Huang, R.S., Jung, T.P., Makeig, S.: Tonic changes in eeg power spectra during simulated driving. In: Foundations of Augmented Cognition. Neuroergonomics and Operational Neuroscience: 5th International Conference, FAC 2009 Held as Part of HCI International 2009 San Diego, CA, USA, July 19-24, 2009 Proceedings 5. pp. 394–403. Springer (2009)
15. Ieracitano, C., Mammone, N., Hussain, A., Morabito, F.C.: A novel multi-modal machine learning based approach for automatic classification of eeg recordings in dementia. Neural Netw. **123**, 176–190 (2020)

16. Jap, B.T., Lal, S., Fischer, P., Bekiaris, E.: Using eeg spectral components to assess algorithms for detecting fatigue. Expert Syst. Appl. **36**(2), 2352–2359 (2009)
17. Jensen, O., Mazaheri, A.: Shaping functional architecture by oscillatory alpha activity: gating by inhibition. Front. Hum. Neurosci. **4**, 186 (2010)
18. Kar, S., Bhagat, M., Routray, A.: Eeg signal analysis for the assessment and quantification of drivers fatigue. Transport. Res. F: Traffic Psychol. Behav. **13**(5), 297–306 (2010)
19. Kaushik, P., Moye, A., Vugt, M.v., Roy, P.P.: Decoding the cognitive states of attention and distraction in a real-life setting using eeg. Scientific Reports **12**(1), 20649 (2022)
20. Kaushik, P., Tripathi, I., Roy, P.P.: Motor activity recognition using eeg data and ensemble of stacked blstm-lstm network and transformer model. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
21. Kilicarslan, A., Vidal, J.L.C.: Characterization and real-time removal of motion artifacts from eeg signals. J. Neural Eng. **16**(5), 056027 (2019)
22. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Commun. ACM **60**(6), 84–90 (2017)
23. Lawhern, V.J., Solon, A.J., Waytowich, N.R., Gordon, S.M., Hung, C.P., Lance, B.J.: EEGNet: a compact convolutional neural network for EEG-based brain-computer interfaces. J. Neural Eng. **15**(5), 056013 (2018)
24. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022)
25. Miao, Z., Zhao, M., Zhang, X., Ming, D.: Lmda-net: A lightweight multi-dimensional attention network for general eeg-based brain-computer interfaces and interpretability. Neuroimage **276**, 120209 (2023)
26. Orrù, G., Micheletto, M., Terranova, F., Marcialis, G.L.: Electroencephalography signal processing based on textural features for monitoring the driver's state by a brain-computer interface. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 2853–2860. IEEE (2021)
27. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
28. Pfurtscheller, G., Da Silva, F.L.: Event-related eeg/meg synchronization and desynchronization: basic principles. Clin. Neurophysiol. **110**(11), 1842–1857 (1999)
29. Polich, J.: Updating p300: an integrative theory of p3a and p3b. Clin. Neurophysiol. **118**(10), 2128–2148 (2007)
30. Schirrmeister, R.T., Springenberg, J.T., Fiederer, L.D.J., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W., Ball, T.: Deep learning with convolutional neural networks for eeg decoding and visualization. Hum. Brain Mapp. **38**(11), 5391–5420 (2017)
31. Shalash, W.M.: Driver fatigue detection with single eeg channel using transfer learning. In: 2019 IEEE International Conference on Imaging Systems and Techniques (IST). pp. 1–6. IEEE (2019)
32. Sheykhivand, S., Rezaii, T.Y., Meshgini, S., Makoui, S., Farzamnia, A.: Developing a deep neural network for driver fatigue detection using eeg signals based on compressed sensing. Sustainability **14**(5), 2941 (2022)
33. Siddhad, G., Gupta, A., Dogra, D.P., Roy, P.P.: Efficacy of transformer networks for classification of eeg data. Biomed. Signal Process. Control **87**, 105488 (2024)

34. Siddhad, G., Iwamura, M., Roy, P.P.: Enhancing eeg signal-based emotion recognition with synthetic data: Diffusion modeel approach. arXiv preprint arXiv:2401.16878 (2024)
35. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
36. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)
37. Wang, F., Wan, Y., Li, M., Huang, H., Li, L., Hou, X., Pan, J., Wen, Z., Li, J.: Recent advances in fatigue detection algorithm based on eeg. Intelligent Automation & Soft Computing **35**(3) (2023)
38. Wang, H., Zhu, X., Chen, P., Yang, Y., Ma, C., Gao, Z.: A gradient-based automatic optimization cnn framework for eeg state recognition. J. Neural Eng. **19**(1), 016009 (2022)
39. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
40. Xu, T., Wang, H., Lu, G., Wan, F., Deng, M., Qi, P., Bezerianos, A., Guan, C., Sun, Y.: E-key: An eeg-based biometric authentication and driving fatigue detection system. IEEE Trans. Affect. Comput. **14**(2), 864–877 (2021)
41. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13. pp. 818–833. Springer (2014)
42. Zheng, W.L., Lu, B.L.: A multimodal approach to estimating vigilance using eeg and forehead eog. J. Neural Eng. **14**(2), 026017 (2017)
43. Zheng, Y., Zhao, X., Yao, L.: Copula-based transformer in eeg to assess visual discomfort induced by stereoscopic 3d. Biomed. Signal Process. Control **77**, 103803 (2022)
44. Zuo, X., Zhang, C., Cong, F., Zhao, J., Hämäläinen, T.: Driver distraction detection using bidirectional long short-term network based on multiscale entropy of eeg. IEEE Trans. Intell. Transp. Syst. **23**(10), 19309–19322 (2022)

Enhanced Cross-Task EEG Classification: Domain Adaptation with EEGNet

Vishal Pandey[1(✉)], Nikhil Panwar[1], Atharva Kumbhar[2], Partha Pratim Roy[1], and Masakazu Iwamura[3]

[1] Indian Institute of Technology Roorkee, Roorkee, Uttarakhand, India
v_pandey@cs.iitr.ac.in
[2] SCTR's Pune Institute of Computer Technology, Pune, Maharashtra, India
[3] Osaka Metropolitan University, Sakai, Osaka, Japan

Abstract. In the domain of electroencephalogram (EEG) research, the generalization of workload classification across different tasks, subjects, and channel configurations remains a significant challenge, primarily due to the field's tradition of conducting within-task studies in controlled, lab-based environments. This study pioneers the application of domain adaptation techniques to EEG data, aiming to transcend these limitations by facilitating consistent cross-task classification of workload levels. Central to our methodology is the integration of source localization techniques that render EEG data channel agnostic, thus enhancing our model's capacity to generalize across diverse channel configurations. By utilizing two disparate datasets–one derived from arithmetic tasks and the other from working memory (n-back) tasks—we implement binary classification to discern between low and high workload states. The core of our approach is a modified Convolutional Neural Network (CNN) model, EEGNet, which is specifically designed to capture the temporal and spatial dynamics inherent in EEG data. Enhanced with Maximum Classifier Discrepancy (MCD) for domain adaptation, and bolstered by source localization, this strategic combination enables an impressive 81.76% accuracy in cross-task classification. The success of our strategy in cross-task, cross-channel, and cross-subject classification not only demonstrates its potential for enhancing the generalizability of EEG data analysis but also marks a significant step forward in applying EEG-based workload classification in real-world scenarios, beyond the confines of laboratory settings. The implications of this research are vast, offering a promising avenue for the generalization of EEG data classification across various domains.

Keywords: EEG · Cognitive Workload · Cross Task · Domain Adaptation

1 Introduction

Cognitive workload, defined as the mental effort required to perform a task, plays a pivotal role in various domains such as aviation, healthcare, and human-computer interaction [1]. Accurate assessment of cognitive workload not only

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15311, pp. 354–369, 2025.
https://doi.org/10.1007/978-3-031-78195-7_24

enhances task performance but also contributes to overall well-being and safety. One promising avenue for quantifying cognitive workload is through the analysis of electroencephalography (EEG) signals, which reflect brain activity associated with different cognitive states. Understanding cognitive processes, such as reasoning and problem-solving, is key to deciphering the brain's operations amid its complex network of neurons and synapses [2]. Despite EEG's acknowledged capability in measuring cognitive workload [3,4], its application faces challenges in consistency across varying tasks and individuals due to the inherent differences in brain dynamics and signal properties. In the realm of automation, the emphasis is placed on developing machines capable of recognizing human cognitive states, aiming to enhance the system's overall effectiveness [5]. EEG data is widely used for monitoring cognitive workload in fields such as neuroscience, psychology, and human-computer interaction. However, the task-specific nature of EEG signals poses a challenge for generalizing workload classifiers across different tasks.

The EEG signals may vary among individuals and tasks, presenting both within-subject and cross subject differences, as well as variations across different tasks. These differences influence the creation of models specific to individual subjects, models applicable across different subjects, and models suitable for cross tasks [6]. Workload classifiers are trained and evaluated on the same task or dataset, limiting their generalizability to new tasks or domains. This limitation hinders the deployment of workload classification systems in real-world settings where tasks may vary widely. Addressing this gap requires the development of robust workload classifiers capable of adapting to diverse task environments—a method known as cross task classification.

To address these challenges, researchers have employed advanced signal processing techniques, feature extraction methods, and machine learning algorithms [7,8]. In particular, domain adaptation techniques have been identified as promising for aligning feature distributions between varied experimental conditions, thereby enhancing classifier robustness [9,10]. Yet, many of these studies focus predominantly on single-task scenarios, leaving a gap in comprehensive analyses that span across multiple tasks, subjects, and channels.

The concept of cross-task modeling involves training a model on one task and then testing it on another task with similarities but distinctions[11–13]. Although different cognitive tasks may trigger different cognitive resources, cognitive workload primarily revolves around the overall amount of cognitive resources utilized, rather than their specific composition. Consequently, there is potential to devise a versatile cognitive workload recognition model capable of discerning workload levels across diverse tasks [14].

Several approaches have been proposed to address the challenge of cross-task and cross-domain EEG classification. Baldwin et al. [15] attempted cross-task workload classification using artificial neural networks, but achieved poor performance below chance levels. Ke et al. [16] had more success using feature selection and support vector machine regression to identify task-independent EEG features for cross-task workload estimation. Zhang et al. [17] proposed a deep

domain adaptation framework using adversarial training to align EEG feature distributions across tasks and subjects. Dimitrakopoulos et al. [18] developed a method using functional connectivity features and sequential feature selection to achieve 87% accuracy for cross-task workload classification. More recently, Zhou et al. [19] employed a deep domain adaptation approach with feature alignment and adversarial training, improving cross-subject workload classification accuracy by 2-9% over baseline methods. Lan et al. [20] applied domain adaptation techniques to EEG-based emotion recognition achieving significant accuracy improvements of up to 13.40% across datasets, showcasing the potential for enhanced subject-independent emotion recognition. These studies demonstrate the potential of domain adaptation techniques to improve the generalization of EEG-based mental state classification across different tasks and subjects.

This paper makes several significant contributions to the advancement of EEG-based classification systems, emphasizing enhanced adaptability and generalizability across various analytical dimensions:

1. **Cross-Task and Cross-Subject Generalizability:** Our framework excels in generalizing EEG classification across diverse cognitive tasks and varied subjects, demonstrating robust performance from Mental Arithmetic to N-Back tasks. This adaptability is powered by our innovative use of domain adaptation techniques and source localization, which ensures effective classification regardless of task or subject variability. Such capabilities are crucial for applications that demand reliable cognitive assessments across different populations.
2. **Effective Domain Adaptation:** The implementation of Maximum Classifier Discrepancy (MCD) for domain adaptation significantly enhances the model's performance in cross-task scenarios. Our findings, including an increase in accuracy to 81.76% under domain adaptation, validate the efficacy of MCD in adjusting the model to handle variations arising from different experimental conditions and task complexities.
3. **Standardization of EEG Data Processing:** By standardizing EEG data into a common brain space using source localization techniques, our study sets a new precedent for ensuring consistent data comparability across sessions and channels. This approach not only aids in maintaining the integrity of EEG analyses but also supports the reproducibility of research findings across varied setups and equipment.

The remaining sections of the paper are structured as follows. Section 2 provides a brief overview of datasets, EEG pre-processing, Source Localization, Model Architecture, domain adaptation and relevant methodologies. The ensuing section, Section 3 delves into the key discoveries derived from the analysis and contrasts the outcomes to assess the effectiveness of the proposed approaches. Section 4 offers concluding remarks for the entire paper.

2 Methods

Utilizing datasets from varied cognitive tasks, the methods involve thorough preprocessing, source localization techniques, and the deployment of advanced models for classification. These steps aim to assess the generalizability and efficacy of EEG signal classification across different cognitive states.

2.1 Dataset

To explore the robustness of EEG classification across various conditions, our study leverages two distinct datasets corresponding to different cognitive tasks. The first dataset comprises EEG recordings from participants engaged in a Mental Arithmetic task and is sourced from the study by Zyma et al. [21], which is publicly available on the PhysioNet platform [22]. The second dataset is derived from a Working Memory task (n-back), obtained from the research conducted by Shin et al. [23]. These datasets provide a foundation for evaluating the generalizability of EEG classification across tasks, channels, and subjects, presenting a comprehensive perspective on the applicability of our models to varied cognitive evaluations.

Dataset 1: Mental Arithmetic Task The Mental Arithmetic task dataset in this study consists of EEG recordings from 36 healthy participants, using a Neurocom monopolar 23-channel EEG system following the 10-20 electrode placement. The EEG data, sampled at 500 Hz, include a 180-second resting phase and a 60-second active mental arithmetic phase. For balanced analysis, only the central 60 seconds of the resting phase were used, equivalent to the active task duration. We categorize the resting phase as representing low workload and the active phase as high workload conditions.

Dataset 2: Working Memory (n-back) Task The N-Back task dataset was collected from 26 healthy individuals, utilizing a 30 channel Brain Products BrainAmp EEG with a sampling frequency of 200 Hz, placed according to the international 10-5 system. The dataset included three task conditions: 0-back, 2-back, and 3-back, designed to assess varying cognitive loads. For our analysis, we specifically focused on the 0-back and 3-back tasks to represent low and high workloads, respectively.

2.2 EEG Data Processing

This section details the precise EEG data processing steps implemented to standardize and enhance the signals.

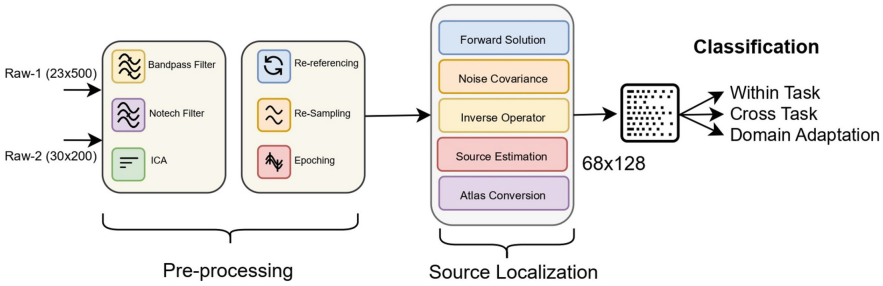


Fig. 1. Signal Processing and Classification Pipeline: Data cleaning and pre-processing with filtering and ICA, followed by source localization to make datasets channel-agnostic and classification through within-task, cross-task, and domain adaptation strategies.

Preprocessing To ensure the integrity and comparability of EEG data across different sessions and conditions, the following preprocessing steps are employed (shown in Figure 1): bandpass filtering to restrict the frequency range to 0.5–45 Hz, notch filtering at 50 Hz to remove electrical line noise, artifact correction using Independent Component Analysis (ICA) to remove eye blinks and cardiac signals, re-referencing the data to an average reference, resampling to 128 Hz to standardize sampling frequency and reduce computational load, and data epoching where fixed-length segments of 1 second with a 0.5-second overlap are created for continuous data coverage during analysis.

Source Localization Following the preprocessing of EEG data, source localization techniques are applied to convert data from electrode space to a standardized brain space using the Desikan-Killiany atlas. This step is essential for comparing and combining datasets effectively.

The process involves setting up the source space and boundary element model (BEM) using a standardized MRI subject ('fsaverage') from MNE's configuration. Specific steps include creating a source space with octahedron subdivisions (oct6), generating the BEM model and solution, and defining the transformation applicable to 'fsaverage'.

The localization to the source space is achieved through the following steps (as shown in Figure 1):

– **Forward Solution:** Compute a forward solution [24] to model the relationship between the source spaces and the EEG sensors,

$$\mathbf{F} = \mathbf{GS},$$

where $\mathbf{G}$ is the lead field matrix representing the gain of each sensor to each source and $\mathbf{S}$ denotes the source strengths.

– **Noise Covariance:** Estimate the noise covariance [25] from the data to regularize the source estimation,

$$\mathbf{C} = \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{E}_i - \overline{\mathbf{E}})(\mathbf{E}_i - \overline{\mathbf{E}})^\top,$$

where $\mathbf{E}_i$ are the EEG data epochs and $\overline{\mathbf{E}}$ is the mean over all epochs.

– **Inverse Operator:** Create an inverse operator using the forward solution and noise covariance. This operator is used to compute the neural source estimates,

$$\mathbf{M} = (\mathbf{G}^\top\mathbf{C}^{-1}\mathbf{G} + \alpha\mathbf{I})^{-1}\mathbf{G}^\top\mathbf{C}^{-1},$$

where α is the regularization parameter, and $\mathbf{I}$ is the identity matrix.

– **Source Estimation:** Apply the inverse operator to the epochs to retrieve source time courses, which are then mapped to the Desikan-Killiany atlas [26],

$$\mathbf{S}_t = \mathbf{ME},$$

and is a crucial step for determining the localization of neural activity [27].

– **Atlas Conversion:** Extract the mean time course from each brain region as defined by the atlas to facilitate direct comparisons across subjects and sessions,

$$\mathbf{T}_r = \frac{1}{N_r}\sum_{j\in R}(\mathbf{S}_{t,j}),$$

where R represents the set of sources within the region r and N_r is the number of sources in r.

This methodology allows the EEG data to be represented in a common brain space, enhancing the interpretability and comparability of the results across different datasets. Following this step, the spatial dimension of the data corresponds to 68 channels, representing the 68 brain regions defined in the Desikan-Killiany atlas. Consequently, the shape of each epoch is transformed to 68x128, indicating 68 brain regions and 128 samples per second. Figure 2 shows the output after EEG source localization which is used as the input for the classification.

2.3 Classification Methods

This section outlines the advanced classification techniques used to analyze EEG data, focusing on the models and configurations designed specifically for identifying cross-task workload levels.

Base Model (Modified EEGNet) EEGNet [9] is selected as the base model for classification due to its compact and efficient architecture specifically designed to handle the high-dimensional and noisy nature of EEG signals. This modified version of EEGNet (as shown in Table 1) is particularly tailored to align with the specific requirements of our EEG datasets, which consist of

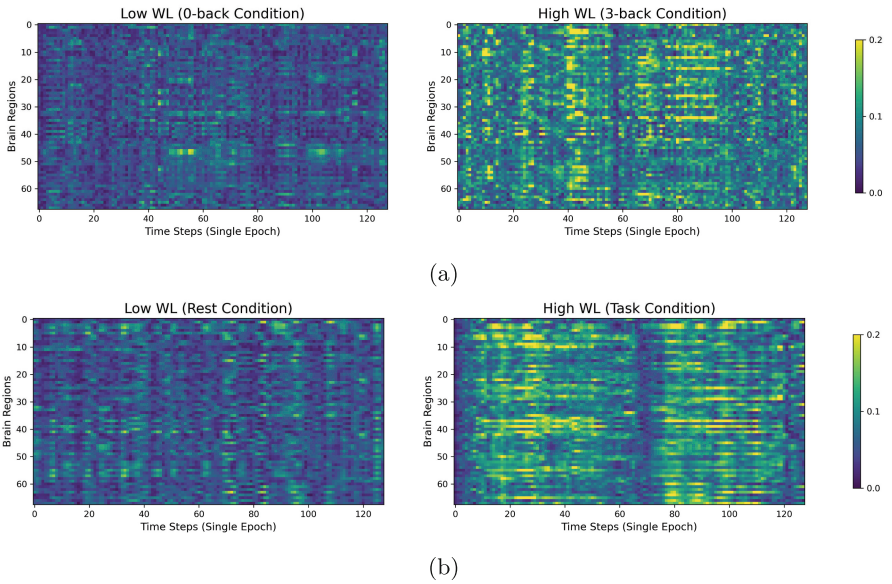


Fig. 2. Output after EEG Source Localization Visualization (a) Arithmetic task dataset (b) Working memory (n-back) task dataset

data segmented into one-second epochs, each comprising 128 time points and originating from 68 distinct brain regions or channels.

The architecture has been modified to include:

- **Temporal and Spatial Filtering:** The model uses $F1$, the number of temporal filters, to initially transform the EEG data across the time dimension. This is complemented by a depth multiplier, D, representing the number of spatial filters per channel, which helps in extracting spatial features specific to each EEG channel through depthwise convolutions.
- **Complex Feature Extraction:** The number of pointwise filters, $F2(= D \times F1)$, is optimized to further process the features, enhancing the model's ability to distinguish between different brain states or conditions effectively.
- **Classification:** The output layer is designed to reflect the number of classes, N, which categorizes the EEG data into predefined classes based on the learned features.

This configuration not only ensures that EEGNet remains lightweight and fast but also improves its accuracy and generalizability across different EEG tasks. Multiple configurations of $F1$ and D, such as (4, 2) and (8, 2), were tested. The best results were achieved with $F1 = 8$ and $D = 4$, which we have used throughout the study to optimize performance, achieving a balance between computational efficiency and the capability to perform detailed and robust classification of EEG signals.

Domain Adaptation using Maximum Classifier Discrepancy In our approach, we employ the Maximum Classifier Discrepancy (MCD) [28] technique for domain adaptation across two distinct EEG tasks: an arithmetic task and a working memory task. MCD leverages adversarial learning principles to align feature distributions from the source and target domains by maximizing the discrepancy between two classifiers, $F1$ and $F2$, both instances of a modified EEGNet model. The feature generator G transforms EEG signals, ensuring their dimensions are suitable for EEGNet classification. The alignment between domains is driven by the discrepancy loss function $d(p_1, p_2)$, which quantifies the mean absolute error between the classifiers' predictions, serving as a critical component of the adversarial training mechanism integral to MCD.

Table 1. Modified EEGNet architecture, where C represents the number of channels or brain regions, set at 68. T denotes the number of time points per epoch, which is 128 in a one-second sample. $F1$ is the number of temporal filters, D the depth multiplier indicating the number of spatial filters, $F2(= D \times F1)$ the number of pointwise filters, and N the number of classes.

Layer	Filters	Size	Output	Activation	Options
Reshape	-	-	$C \times T \times 1$	-	-
Conv2D	F1	$(1, \frac{T}{2})$	$C \times T \times F1$	Linear	Padding=same
Batch Normalization	-	-	$C \times T \times F1$	-	-
Depthwise Conv2D	D*F1	$(C, 1)$	$1 \times T \times D \times F1$	Linear	Padding=valid, Depth=D, MaxNorm(1)
Batch Normalization	-	-	$1 \times T \times D \times F1$	-	-
Activation	-	-	$1 \times T \times D \times F1$	ELU	-
Average Pooling2D	-	$(1, 4)$	$1 \times \frac{T}{4} \times D \times F1$	-	-
Dropout	-	-	$1 \times \frac{T}{4} \times D \times F1$	-	Rate=0.25
Separable Conv2D	F2	$(1, \frac{T}{8})$	$1 \times \frac{T}{8} \times F2$	Linear	Padding=same
Batch Normalization	-	-	$1 \times \frac{T}{8} \times F2$	-	-
Activation	-	-	$1 \times \frac{T}{8} \times F2$	ELU	-
Average Pooling2D	-	$(1, 8)$	$1 \times \frac{T}{32} \times F2$	-	-
Dropout	-	-	$1 \times \frac{T}{32} \times F2$	-	Rate=0.25
Flatten	-	-	$F2 \times \frac{T}{32}$	-	-
Dense	-	-	$N \times (F2 \times \frac{T}{32})$	ReLU	MaxNorm(0.25)
Dense	-	-	1	Sigmoid	MaxNorm(0.25)

Generator Architecture: The generator network is expertly constructed to transform EEG input while preserving its original shape for compatibility with EEGNet classifiers. It is a sequential convolutional model that ensures the dimensions of the input EEG data, consisting of 68 channels by 128 time points, remain intact through each layer. The use of 'same' padding and specific activation functions within the network facilitates this preservation of spatial dimensions. The detailed architecture of the generator network, including the layers, filter sizes, and activations, is delineated in Table 2, highlighting its role as an integral component in our domain adaptation framework.

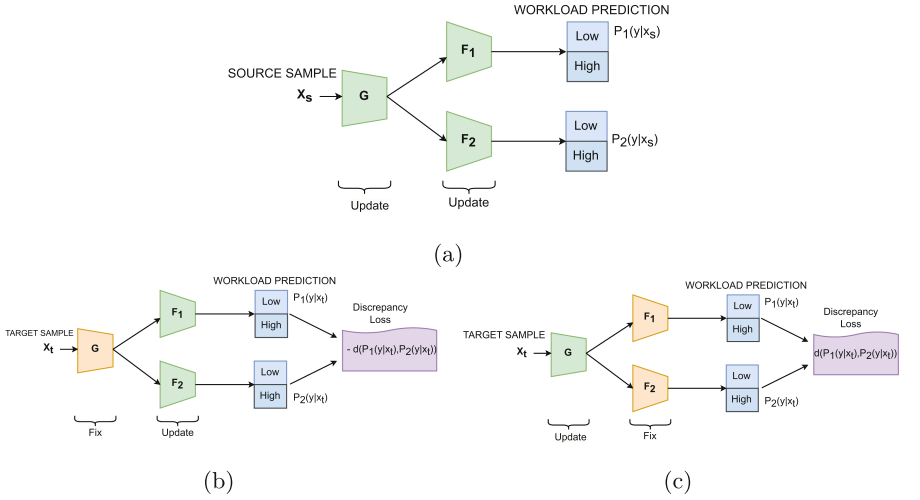


Fig. 3. Schematic of the adaptive workload prediction model: (a) Training with source samples X_s, utilizing a shared generator G and two distinct classifiers F_1 and F_2 for workload categorization. (b) showcases the maximization of the discrepancy on the target sample X_t, fixing the generator G and updating classifier F_1 and F_2. (c) illustrates the minimization of discrepancy, updating G and fixing F_1 and F_2, to refine the prediction accuracy on X_t.

Table 2. Generator Network Architecture

Layer	Filters	Size	Output	Activation	Options
InputLayer	-	-	$68 \times 128 \times 1$	-	-
Conv2D	32	$(3, 3)$	$68 \times 128 \times 32$	ReLU	Padding=same
Batch Normalization	-	-	$68 \times 128 \times 32$	-	-
Conv2D	64	$(3, 3)$	$68 \times 128 \times 64$	ReLU	Padding=same
Batch Normalization	-	-	$68 \times 128 \times 64$	-	-
Conv2D	1	$(3, 3)$	$68 \times 128 \times 1$	Sigmoid	Padding=same

Discrepancy Loss Function: A key component of our MCD approach is the discrepancy loss function, which quantifies the difference in the classifiers' probabilistic outputs. The discrepancy loss function is defined as the mean absolute error (MAE) between the output probabilities of the two classifiers:

$$d(p_1, p_2) = \frac{1}{N} \sum_{i=1}^{N} |p_1(Y|X_t^{(i)}) - p_2(Y|X_t^{(i)})|,$$

where $p_1(Y|X_t^{(i)})$ and $p_2(Y|X_t^{(i)})$ are the predicted probabilities by classifiers $F1$ and $F2$ for the target sample $X_t^{(i)}$, and N is the number of target samples.

Training Objectives: The training involves a three-step process (as shown in Figure 3), each targeting a specific objective:

1. **Step A (Source Domain Training):** Train the classifiers and generator on the source domain data to minimize the binary cross-entropy loss:

$$\min_{G,F1,F2} \mathcal{L}_{\text{CE}}(X_s, Y_s),$$

$$\mathcal{L}_{\text{CE}}(X_s, Y_s) = -\frac{1}{N} \sum_{i=1}^{N} \left[Y_s^{(i)} \log(p(Y_s^{(i)}|X_s^{(i)})) + (1 - Y_s^{(i)}) \log(1 - p(Y_s^{(i)}|X_s^{(i)})) \right].$$

2. **Step B (Discrepancy Maximization on Target Data):** Adapt the classifiers to maximize the discrepancy on the target domain data while retaining performance on the source domain:

$$\min_{F1,F2} \mathcal{L}_{\text{CE}}(X_s, Y_s) - \mathcal{L}_{\text{adv}}(X_t),$$

$$\mathcal{L}_{\text{adv}}(X_t) = \mathbb{E}_{X_t \sim X_t}[d(p1(y|X_t), p2(y|X_t))],$$

   with the discrepancy d measured as the mean absolute error between the probabilistic outputs of the classifiers.
3. **Step C (Minimization of Discrepancy on Target Data):** Optimize the generator to minimize the discrepancy between the classifiers' predictions on the target data:

$$\min_{G} \mathcal{L}_{\text{adv}}(X_t).$$

   The hyperparameter n determines the iteration count for each mini-batch within the training cycle. It serves as a tuning parameter in the original MCD framework, calibrating the equilibrium between the generator's and the classifiers' objectives during the adversarial training process.

This training regimen is iterated to refine the model's ability to classify EEG data correctly, irrespective of the task. The loss functions are tailored to ensure that the adversarial nature of MCD leads to domain-invariant feature extraction, enhancing the model's generalization capabilities.

2.4 Classification Experiments

In order to evaluate the efficacy of our domain adaptation approach using Maximum Classifier Discrepancy (MCD), we conducted a series of classification experiments. These experiments were designed to assess the performance of our model under various conditions and are divided into three main cases:

Case 1: Within-Task Classification The first case investigates the performance of our model when training and testing are conducted within the same task domain. This serves as a baseline to understand the model's capability in a controlled setting, where the feature distribution between training and testing sets is consistent.

Case 2: Cross-Task Classification without Domain Adaptation In Case 2, we assess the cross-task classification capabilities of our approach without the aid of domain adaptation techniques. We conduct two sets of experiments to evaluate the baseline generalizability of the network: first, by training on the dataset from the arithmetic task (D1) and testing on the working memory task (D2), and second, by reversing this approach–training on the working memory task (D2) and testing on the arithmetic task (D1). These assessments provide insight into the inherent ability of the network to transfer learning between different cognitive tasks without domain-specific tuning.

Case 3: Cross-Task Classification with Domain Adaptation In the third case, our focus shifts to evaluating the effectiveness of the MCD technique with the three-step training process for domain adaptation across tasks. We modulate the domain adaptation intensity by varying the hyperparameter n – denoting the number of iterations per mini-batch – with values of 2, 3, and 4. This variation allows us to observe the nuanced impacts of domain adaptation on knowledge transfer between the arithmetic and working memory tasks. For each setting, the domain adaptation procedure is conducted twice, with each dataset serving as the source and then as the target, ensuring a comprehensive evaluation of the model's adaptability to both tasks.

3 Results

In an effort to substantiate the performance of our EEG classification framework, we executed a comprehensive set of experiments. Both the Mental Arithmetic and Working Memory datasets were divided into training and testing sets with a 70-30 split. Our experiments were conducted over 50 epochs with an early stopping mechanism set at 5 epochs to prevent overfitting. We selected a batch size of 512 to optimize the computational efficiency and convergence rate. The optimization of our model was conducted using the Adam optimizer, known for its efficacy in handling sparse gradients on noisy problems. All experiments were implemented using TensorFlow, which provides a robust backend framework conducive to extensive neural network experimentation.

The results from these experiments, as detailed in Table 3, highlight the model's capability to not only adapt within specific tasks but also to generalize across different tasks, subjects, and channels–key factors for the deployment of EEG classification systems in real-world applications.

Within-Task Classification Results Within-task classification revealed a superior performance, with an accuracy of 85.23% for dataset D1. This high accuracy, coupled with an F1 score of 84.57%, suggests that the model reliably differentiates cognitive states within the same type of task. The close precision and recall scores indicate a balanced classification capability, with dataset D2 showing only a slight reduction in performance across all metrics. The success

Table 3. Classification Performance Metrics Across Cases and Conditions

Case	Condition	Accuracy	Precision	Recall	F1 Score
Within-Task	D1	**85.23**	84.76	83.68	84.57
	D2	84.67	83.59	82.45	83.54
2cmCross-Task without DA	D1 to D2	**65.12**	67.34	64.28	65.75
	D2 to D1	62.89	60.73	63.54	61.67
Cross-Task with DA	D1 to D2, $n = 2$	75.68	74.56	73.47	74.32
	D2 to D1, $n = 2$	74.21	73.89	72.58	73.45
	D1 to D2, $n = 3$	80.47	81.02	79.56	80.29
	D2 to D1, $n = 3$	79.34	78.69	77.85	78.27
	D1 to D2, $n = 4$	**81.76**	81.49	81.37	80.53
	D2 to D1, $n = 4$	80.42	79.67	80.58	80.36

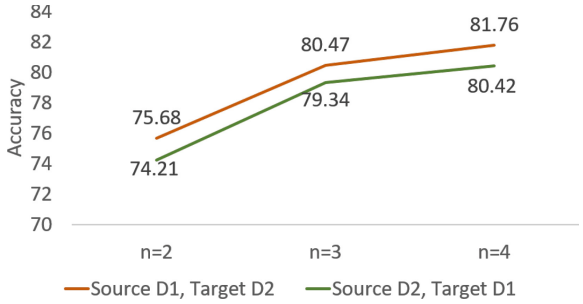


Fig. 4. Graph displays the performance trends of two domain adaptation scenarios with training iterations (n=2, n=3, n=4).

in this case underscores the model's proficiency in discerning between low and high workload conditions when the task remains constant.

Cross-Task Classification without Domain Adaptation Transitioning to cross-task classification without domain adaptation presented a notable challenge for the model, as evidenced by a significant decrease in accuracy, with the highest being 65.12% when training on D1 and testing on D2. This dip reflects the inherent difficulty in transferring knowledge between different cognitive tasks without additional adaptation mechanisms. Although precision remained relatively stable, a drop in recall for D2 when tested on D1 suggests a propensity for the model to miss true positive instances under a new domain, indicating the necessity for domain-specific tuning to improve generalizability.

Cross-Task Classification with Domain Adaptation Introducing domain adaptation markedly enhanced the model's adaptability across tasks. Notably, the accuracy improved to 81.76% for D1 tested on D2 with $n = 4$ adaptation

iterations, emphasizing the significant role of domain adaptation in cross-task classification scenarios. This represents a substantial improvement from the non-adapted models and demonstrates that the model can be effectively tuned for different cognitive tasks using the MCD approach. The increment in the number of adaptation iterations (n) from 2 to 4 yielded consistent improvements across all performance metrics as shown in Figure 4, suggesting that a more intensive domain adaptation process is beneficial for model robustness. The high F1 score for $n = 4$, 80.53%, further reinforces the model's precision and recall harmony when the domain adaptation is optimally configured.

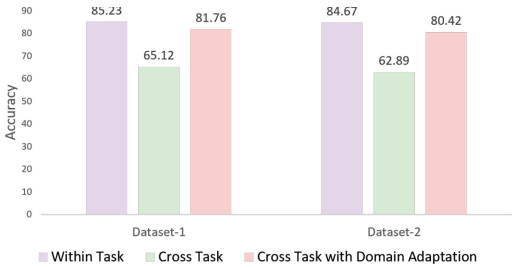


Fig. 5. Comparison of Classification Accuracies: illustrating the performance of both datasets for within task, cross-task with and without domain adaptation.

The results delineate a clear progression from within-task competency to effective cross-task classification with the integration of domain adaptation techniques (depicted in Figure 5). The gradual enhancement of performance metrics with increased domain adaptation efforts highlights the importance of this approach in extending the scope across varied cognitive evaluations.

The results of this study hold substantial implications for the advancement of EEG-based classification systems, particularly in their application across different tasks, subjects, and channels. The robust within-task performance sets a standard for the model's capabilities and provides a reference point for assessing cross-task generalizability. The differential success rates in cross-task classification underscore the challenges and reveal opportunities for enhancement through domain adaptation. Notably, the improvement in accuracy and F1 scores in the presence of domain adaptation underscores its critical role in surmounting variabilities inherent to cross-task, cross-subject, and cross-channel scenarios. This adaptability is crucial for the practical deployment of EEG classifiers in dynamic real-world settings, where the ability to interpret neural signals consistently, despite variations in task context, subject responses, and channel configurations, is essential. The positive outcomes observed with domain adaptation not only highlight its necessity but also affirm its efficacy in promoting the widespread applicability of EEG classification models.

4 Conclusion

This study has made significant strides in the field of EEG-based classification systems, demonstrating robust capabilities in handling variability across different tasks, subjects, and channels. Through the application of advanced preprocessing techniques, domain adaptation methods, and innovative neural network architectures, we have shown that it is possible to significantly enhance the adaptability and generalizability of EEG classifiers. The key achievements of this study include the successful generalization of the EEG classification model across varied cognitive tasks and among diverse participant groups. By employing the Maximum Classifier Discrepancy (MCD) technique, we improved the model's accuracy in cross-task scenarios, which is a substantial step forward from traditional EEG classifiers that are often limited to specific task conditions. Our approach has proven particularly effective, with notable improvements in classification accuracy reaching up to 81.76% in complex domain adaptation settings. Additionally, the standardization of EEG data through source localization to a common brain space has enhanced the comparability and consistency of data analysis across studies and setups.The implications of our findings are profound, extending beyond the immediate scope of EEG classification. The ability to accurately classify EEG data across varying conditions suggests potential applications in personalized medicine, where EEG-based diagnostics and treatments could be tailored to individual physiological and cognitive profiles. Furthermore, our methodological innovations provide a framework that can be adapted to other modalities of biomedical data, potentially transforming approaches to neural monitoring and intervention.While our results are promising, they also highlight areas for further research. One limitation of the current study is the focus on controlled laboratory tasks. Future work could extend these methods to more naturalistic settings to better capture the complexities of real-world cognitive states. Another limitation is the focus on a binary classification task; future work will extend our approach to multiclass problems to evaluate the performance and robustness of MCD in more complex classification scenarios. Further research should also explore the scalability of the proposed models in larger, more heterogeneous populations to validate the generalizability of the findings. Moreover, exploring the impact of varying the hyperparameters, such as the number of adaptation iterations (n), could provide deeper insights into the dynamics of domain adaptation in neural networks.

In conclusion, this study not only advances our understanding of EEG signal classification but also sets a benchmark for future research in this rapidly evolving field. By addressing both methodological challenges and practical applications, the contributions of this research pave the way for the next generation of EEG classification systems that are more adaptable, reliable, and widely applicable in clinical and research settings. The potential of these systems to revolutionize how we understand and interact with the human brain is immense, promising significant advancements in neuroscience, medicine, and beyond.

References

1. Gupta, S.S., Taori, T.J., Ladekar, M.Y., Manthalkar, R.R., et al.: Classification of cross task cognitive workload using deep recurrent network with modelling of temporal dynamics. Biomed. Signal Process. Control **70**, 103070 (2021)
2. Sporns, O., Tononi, G., Edelman, G.M.: Connectivity and complexity: the relationship between neuroanatomy and brain dynamics. Neural Netw. **13**(8–9), 909–922 (2000)
3. Chakladar, D.D., Dey, S., Roy, P.P., Dogra, D.P.: Eeg-based mental workload estimation using deep blstm-lstm network and evolutionary algorithm. Biomed. Signal Process. Control **60**, 101989 (2020)
4. Siddhad, G., Gupta, A., Dogra, D.P., Roy, P.P.: Efficacy of transformer networks for classification of eeg data. Biomed. Signal Process. Control **87**, 105488 (2024)
5. P. Ramakrishnan, B. Balasingam, and F. Biondi, "Cognitive load estimation for adaptive human–machine system automation," in *Learning control*, pp. 35–58, Elsevier, 2021
6. Zhou, Y., Huang, S., Xu, Z., Wang, P., et al.: Cognitive workload recognition using eeg signals and machine learning: A review. IEEE Transactions on Cognitive and Developmental Systems **14**(3), 799–818 (2021)
7. P. Kaushik, I. Tripathi, and P. P. Roy, "Motor activity recognition using eeg data and ensemble of stacked blstm-lstm network and transformer model," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, 2023
8. Panwar, N., Pandey, V., Roy, P.P.: Eeg-cognet: A deep learning framework for cognitive state assessment using eeg brain connectivity. Biomed. Signal Process. Control **98**, 106770 (2024)
9. Lawhern, V.J., Solon, A.J., Waytowich, N.R., Gordon, S.M., et al.: Eegnet: a compact convolutional neural network for eeg-based brain-computer interfaces. J. Neural Eng. **15**(5), 056013 (2018)
10. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. Knowl. Data Eng. **22**(10), 1345–1359 (2009)
11. Ke, Y., Jiang, T., Liu, S., Cao, Y., et al.: Cross-task consistency of electroencephalography-based mental workload indicators: comparisons between power spectral density and task-irrelevant auditory event-related potentials. Front. Neurosci. **15**, 703139 (2021)
12. Zhou, Y., Xu, Z., Niu, Y., Wang, P., et al.: Cross-task cognitive workload recognition based on eeg and domain adaptation. IEEE Trans. Neural Syst. Rehabil. Eng. **30**, 50–60 (2022)
13. Jiménez-Guarneros, M., Gómez-Gil, P.: Custom domain adaptation: A new method for cross-subject, eeg-based cognitive load recognition. IEEE Signal Process. Lett. **27**, 750–754 (2020)
14. Ke, Y., Qi, H., He, F., Liu, S., et al.: An eeg-based mental workload estimator trained on working memory task can work well under simulated multi-attribute task. Front. Hum. Neurosci. **8**, 703 (2014)
15. Baldwin, C.L., Penaranda, B.: Adaptive training using an artificial neural network and eeg metrics for within-and cross-task workload classification. Neuroimage **59**(1), 48–56 (2012)
16. Ke, Y., Qi, H., Zhang, L., Chen, S., et al.: Towards an effective cross-task mental workload recognition model using electroencephalography based on feature selection and support vector machine regression. Int. J. Psychophysiol. **98**(2), 157–166 (2015)

17. Zhang, P., Wang, X., Zhang, W., Chen, J.: Learning spatial-spectral-temporal eeg features with recurrent 3d convolutional neural networks for cross-task mental workload assessment. IEEE Trans. Neural Syst. Rehabil. Eng. **27**(1), 31–42 (2018)
18. Dimitrakopoulos, G.N., Kakkos, I., Dai, Z., Lim, J., et al.: Task-independent mental workload classification based upon common multiband eeg cortical connectivity. IEEE Trans. Neural Syst. Rehabil. Eng. **25**(11), 1940–1949 (2017)
19. Zhou, Y., Wang, P., Gong, P., Wei, F., et al.: Cross-subject cognitive workload recognition based on eeg and deep domain adaptation. IEEE Trans. Instrum. Meas. **72**, 1–12 (2023)
20. Lan, Z., Sourina, O., Wang, L., Scherer, R., et al.: Domain adaptation techniques for eeg-based emotion recognition: a comparative study on two public datasets. IEEE Transactions on Cognitive and Developmental Systems **11**(1), 85–94 (2018)
21. Zyma, I., Tukaev, S., Seleznov, I., Kiyono, K., et al.: Electroencephalograms during mental arithmetic task performance. Data **4**(1), 14 (2019)
22. A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, *et al.*, "Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals," *circulation*, vol. 101, no. 23, pp. e215–e220, 2000
23. Shin, J., Von Lühmann, A., Kim, D.-W., Mehnert, J., et al.: Simultaneous acquisition of eeg and nirs during cognitive tasks for an open access dataset. Scientific data **5**(1), 1–16 (2018)
24. Hämäläinen, M.S., Ilmoniemi, R.J.: Interpreting magnetic fields of the brain: minimum norm estimates. Medical & biological engineering & computing **32**, 35–42 (1994)
25. Engemann, D.A., Gramfort, A.: Automated model selection in covariance estimation and spatial whitening of meg and eeg signals. Neuroimage **108**, 328–342 (2015)
26. Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., et al.: An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. Neuroimage **31**(3), 968–980 (2006)
27. Dale, A.M., Fischl, B., Sereno, M.I.: Cortical surface-based analysis: I. segmentation and surface reconstruction. Neuroimage **9**(2), 179–194 (1999)
28. K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3723–3732, 2018

Advanced Framework for Early Congestive Heart Failure Detection Using Electrocardiogram Data and Ensemble Learning Models

Aditya Oza(✉), Sanskriti Patel, Abhishek Shrivastava, and Santosh Kumar

IIIT, Naya Raipur, Naya Raipur, India
aditya21102@iiitnr.edu.in

Abstract. Our research addresses the pressing issue of congestive heart failure (CHF), a critical cardiovascular condition characterized by the heart's diminished ability to pump blood effectively, resulting in fluid accumulation. Current diagnostic methods often face challenges in signal processing and manual Electrocardiogram (ECG) analysis, leading to reduced accuracy and diagnostic complexities. To tackle these challenges, we introduce an innovative framework that integrates global QRS average (gQRS) detection of RR peaks and intervals from ECG data. We then apply advanced machine learning models such as K-Nearest Neighbours (KNN), Support Vector Machine (SVM), Random Forest (RF), and XGBoost, specifically tailored for CHF diagnosis. What distinguishes our approach is the strategic use of ensemble learning, combining the predictive strengths of XGBoost and RF algorithms. This fusion optimizes diagnostic outcomes and demonstrates significant improvements in CHF detection accuracy, marking a notable advancement in clinical diagnostics. Our research underscores the potential of ensemble learning methodologies in enhancing diagnostic accuracy and clinical decision-making for CHF. By leveraging cutting-edge technologies and methods, we aim to revolutionize cardiovascular health monitoring and contribute to more effective patient care strategies. This innovative approach not only achieves a high accuracy rate of 99.56% but also significantly reduces processing time, making our research highly impactful and promising for practical healthcare applications.

Keywords: Congestive Heart Failure · Electrocardiogram · Ensemble learning.

1 Introduction

Congestive heart failure (CHF) [1,2] is a significant global health issue, affecting approximately 26 million people worldwide [3]. Despite medical advancements,

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15311, pp. 370–385, 2025.
https://doi.org/10.1007/978-3-031-78195-7_25

early detection of CHF remains challenging due to the subtlety of its early symptoms and the complexity of its underlying causes. Traditional methods for analyzing electrocardiogram (ECG) signals often miss early-stage CHF markers, resulting in late diagnoses when treatment options are limited and less effective.

Early prediction of CHF is critically needed to improve patient outcomes and alleviate the burden on healthcare systems. Detecting CHF at an early stage allows for timely interventions, lifestyle modifications, and medical treatments that can significantly slow disease progression, enhance quality of life, and reduce mortality rates. Traditional methods, which include statistical and machine learning (ML) approaches, are valuable for identifying subtle patterns in ECG data indicative of early-stage CHF.

Our research highlights the superiority of traditional ML techniques over deep learning (DL) methods for early CHF detection, leveraging ECG data's critical role in cardiovascular diagnostics. Through advanced ML methods, we extract intricate features from raw ECG signals that traditional approaches often miss. A rigorous preprocessing protocol ensures precision by addressing missing values, redundancies, and errors. This includes feature engineering and data extraction to derive essential information from ECG signals [4–6]. We introduce a novel framework incorporating ensemble learning, specifically using XGBoost and Random Forest (RF), to enhance model performance [7,8]. This approach significantly improves early CHF detection, offering more accurate predictions and revolutionizing cardiovascular healthcare by enabling early intervention and personalized care.

1.1 Rationale for Choosing Traditional Methods

While recent advancements in DL have shown promising results in various domains, we have opted to focus on traditional ML methods for this study due to several key reasons. Firstly, DL models often require large datasets to avoid overfitting and to generalize well. Our dataset, though comprehensive, may not be sufficiently large to train deep models effectively without overfitting. Secondly, traditional ML methods are computationally more efficient, which is crucial for applications requiring quick deployment and lower computational resources. Thirdly, traditional models offer better interpretability, allowing for easier understanding and communication of the model's decision-making process to clinicians and stakeholders. By emphasizing computational efficiency and interpretability, our approach ensures that the models are not only effective but also practical for real-world deployment in varied clinical settings.

1.2 Contributions

The major contributions of this research are given below:

1. We propose a novel framework for the early prediction of CHF using ensemble-based fusion techniques, integrating the strengths of XGBoost, SVM, RF, and KNN models for improved predictive accuracy.

2. The effectiveness of the proposed framework has been rigorously evaluated on the BIDMC CHF and MIT-BIH datasets, demonstrating superior accuracy and reduced computational time compared to existing methods.

The remaining of this work is framed as follows: Section II contains the literature review, exploring research gaps and challenges. Section III discusses methods and material, including dataset description and proposed methods. Section IV presents results and discussion, while Section V concludes the work and outlines future scope.

2 Literature Survey

Savarese et al. [1] reviewed significant advancements in diagnosing heart conditions, with a focus on CHF. Zou et al. [2] introduced an architecture combining LSTM networks with DCNNs, achieving 97.62% accuracy in real-time CHF detection. Shrivastava et al. [4] developed a system for detecting myocardial infarction (MI) using three feature selection methods and eight ML algorithms, enhancing performance and reducing complexity. Other studies [5–7] transitioned to federated learning with an RF approach, achieving a 95% accuracy in CHF diagnosis, while the HBA-FRCNN technique addressed ECG signal noise, attaining 97.65% accuracy in predicting chronic heart failure.

Rai et al. [8] explored the application of various ML and DL methods for cardiac disease detection, offering insights into advanced diagnostic techniques. Khan et al. [9] studied prevailing ML models for predicting cardiac arrests, emphasizing the need for rigorous evaluation to improve healthcare predictions. Their findings highlight the importance of better threat evaluation to enhance outcomes and resource allocation in CVD diagnosis. Bhaskarpandit et al. [10] demonstrated significant advancements in cardiac diagnosis using eigendomain deep representation learning for 12-lead ECG trace images, highlighting the potential of ML and DL techniques to enhance diagnostic precision and patient care.

De Marco et al. [29] highlighted the significant role of DL in ECG analysis, specifically for the identification of premature ventricular contractions (PVC). Using the MIT-BIH Arrhythmia Database, they classified QRS complexes through five deep neural networks: LSTM, AlexNet, GoogleNet, Inception V3, and ResNet-50. The experiments demonstrated high efficiency and reliability, with ResNet-50 achieving 99.8% accuracy and 99.2% F1-score, and Inception V3 achieving 98.8% accuracy and 98.8% F1-score.

Observations from the literature review reveal several gaps:

1. Model complexity often reduces interpretability, hindering collaboration between medical professionals and data scientists.
2. Many models lack time-efficient methods suitable for seamless clinical integration, limiting rapid decision-making.
3. Existing ECG preprocessing methods are insufficiently robust, leading to inaccuracies in analysis.

4. Lack of model explainability hinders clinician trust and adoption in practice.
5. Justification for the practical use of advanced ML models in improving patient outcomes is often inadequate.

3 Methods and Materials

To address the objective of early detection of CHF[10] based on ECG data[11], we present a comprehensive methodology encompassing several key steps. Figure 1 demonstrates how the ECG dataset is preprocessed and the features with maximum importance for the classification of healthy and unhealthy ECGs are taken, various ML models are used including KNN, SVM, RF, XGBoost, and also an ensemble learning approach was employed using RF and XGBoost. The performance is then compared to find the best-performing model in classifying the healthy and unhealthy ECGs.

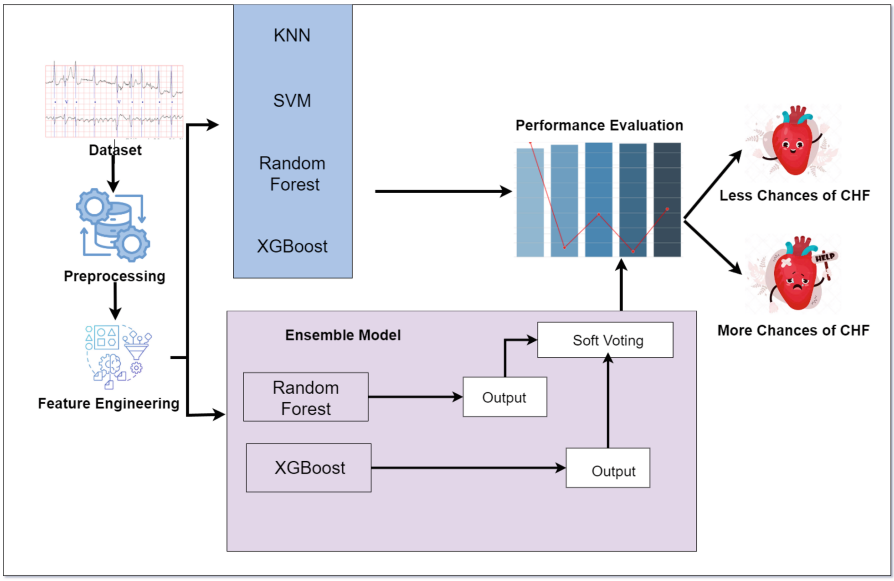


Fig. 1. Overview of the proposed model: starting with the ECG dataset on the left, we progress through preprocessing, feature engineering, application of various ML models, and finally compare their performance.

3.1 Dataset Description

In this research, we utilize the BIDMC CHF and MIT-BIH datasets [12,13] to ensure the quality of our training data, which is crucial for our model's effectiveness. We systematically collect and preprocess data, transforming raw ECG

recordings to align with best practices. Drawing from leading cardiovascular research, our approach aims to enhance the model's generalizability and reliability.

Table 1. Databases used in our research

Database	NHYA Class	# Subjects	# Males (age)	# Females (age)	# EB
BIDMC	CHF	15	11(22 – 71 years)	4(54 – 63 years)	20,000
MIT-BIH Arrhythmia	NSR	18	5(26 – 45 years)	13(20 – 50 years)	36,000

BIDMC Congestive Heart Failure Database [11] The BIDMC database includes ECG recordings from 15 individuals (11 men aged 22-71 and 4 women aged 54-63) with severe CHF (NYHA class 3–4). These recordings provide valuable insights into cardiac dynamics in severe heart failure.
MIT-BIH NSRDB [12] This database contains 18 recordings from patients without notable arrhythmias, including 13 women (ages 20-50) and 5 men (ages 26-45). Each recording, made with ambulatory ECG recorders, lasts around 20 hours and features two ECG signals captured at 250 samples per second.

3.2 Data Preprocessing

Effective preprocessing of ECG data is essential to ensure accurate analysis. This study uses a standardization process to address noise and discrepancies in ECG recordings, ensuring data consistency [1,14,15]. The first step involves noise reduction, focusing on non-QRS variations, using two Moving Average Cascades (MACs) with different impulse response lengths (140 ms and 25 ms). This approach preserves QRS peak amplitudes while reducing slow waves like T waves and baseline drifts [2].

$$X[i] = \frac{1}{N} \sum_{j=0}^{N} X[i+N] \tag{1}$$

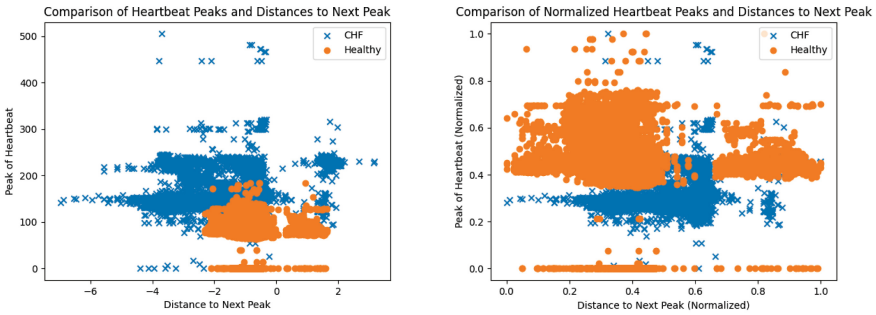
Next, QRS complexes are enhanced using derivative filters, with the optimal filter selected based on a Signal Quality Index (SQI):

$$\text{SQI} = \frac{k_s + mD_s}{k_n + mD_n} \tag{2}$$

Here, mD_s represents QRS-specific derivatives, and mD_n captures noise, with the highest SQI filter applied [6]. QRS detection is then performed by comparing the absolute derivative value to an adaptive threshold:

$$T = \min(\overline{D}, 2.5 \cdot T^{(0)}) \quad \max(\overline{D}, 0.5 \cdot T^{(0)}) \tag{3}$$

The threshold adjusts dynamically based on the distance from the last QRS detection to reduce false positives. Heartbeat occurrence is estimated by locating



(a) Before Data Normalization (b) After Data Normalization

Fig. 2. Diagram represents the heartbeat comparison (a) Before Data Normalization and (b) After Data Normalization.

the peak of the signed derivative signal. Finally, normalization using MinMaxScaler standardizes feature scales, ensuring consistency in model training, with scatter plots visualizing the normalized data for CHF and normal conditions [4]. Figure 2 shows how normalization helped to standardize the data.

3.3 Feature extraction

Feature engineering is a crucial step in extracting meaningful information from the ECG signals [16–18]. The following key features are engineered to capture essential aspects of cardiac dynamics:

QRS Wave The gQRS detection technique plays a crucial role in our investigation, accurately identifying QRS complexes in ECG waveforms. This precision is essential for meaningful feature extraction and reliable predictive models for early CHF detection. Rigorous validation confirms the technique's suitability and effectiveness. Figure 3 shows the ECG after the gQRS detection.

$$Sp = \frac{Correct\ QRS\ Predicted}{Total\ number\ of\ true\ QRS\ peaks} \tag{4}$$

$$Pp = \frac{Correct\ QRS\ Predicted}{Total\ number\ of\ QRS\ Predicted} \tag{5}$$

$$F1 = 2 \times \frac{Sp \times Pp}{Sp + Pp} \tag{6}$$

The QRS complex represents the depolarization of the ventricles and is a crucial feature in ECG analysis. Its duration ($QRS\_Duration$) can be calculated as the time taken from the onset to the offset of the QRS complex:

$$QRS\_Duration = QRS\_Offset - QRS\_Onset \tag{7}$$

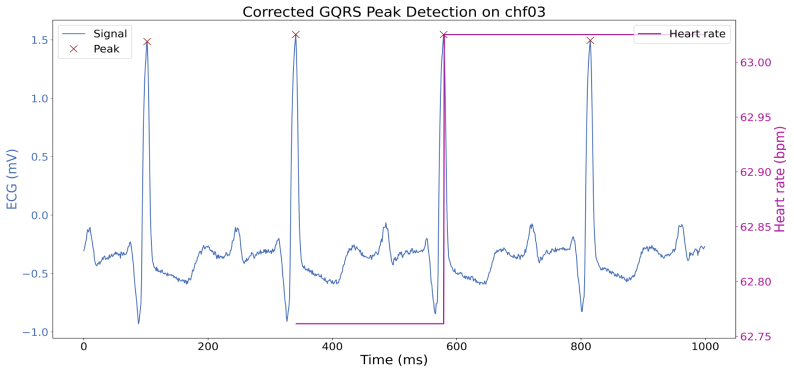


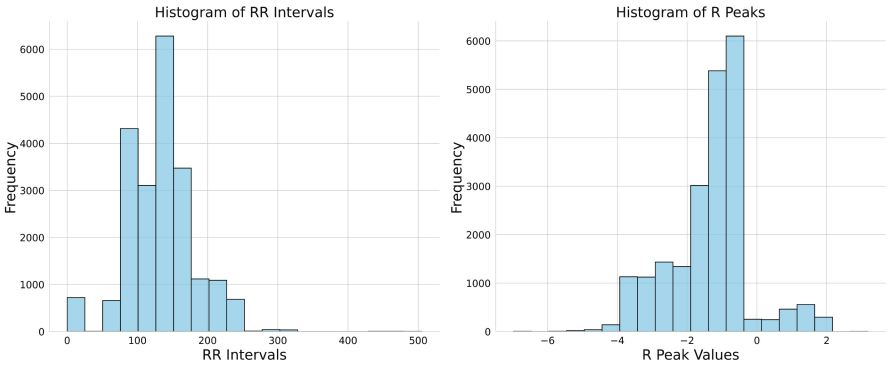
Fig. 3. Corrected gQRS R-Peak Detection used to extract R Peaks and RR Intervals

The Inter-Beat Interval, also known as the RR interval, reflects the period between two R-peaks. It is a fundamental measure of heart rate variability (HRV) and is computed as:

$$IBI = R_n - R_{n-1} \tag{8}$$

where R_n and R_{n-1} are the locations of consecutive R-peaks.

To ensure consistency in feature scaling, normalization is applied to the extracted features. Figure 4: This figure (a) and (b) provide a comparison of the distribution of RR-Intervals and R Peaks, respectively. Understanding these distributions is crucial as they represent key features in ECG data that can help in analyzing heart rhythm patterns and detecting abnormalities.



(a) RR-Intervals Distribution (b) R Peaks Distribution

Fig. 4. Diagram represents the distribution comparison (a) RR-Intervals distribution and (b) R Peaks distribution.

3.4 Machine Learning Models

In this research, we employed a range of ML models to address the binary classification problem associated with QRS peaks and intervals. The selected models include K-Nearest Neighbors (KNN), SVM [18], RF[18,19], and XGBoost[19]. KNN is a non-parametric technique that classifies data points based on the majority class of their nearest neighbors. The classification decision rule can be formulated as follows:

$$\hat{y} = \text{argmax} \left(\sum_{i=1}^{K} I(y_i = j) \right) \tag{10}$$

SVM, a robust classifier, maximizes the margin between classes by identifying the optimal hyperplane. It can be formulated as an optimization problem:

$$\min_{w,b,\xi} \frac{1}{2}||w||^2 + C \sum_{i=1}^{n} \xi_i \tag{11}$$

subject to:

$$y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \tag{12}$$

RF is an ensemble learning method that creates multiple decision trees during training and combines their predictions:

$$\hat{y} = \text{mode}(\text{predictions}_{\text{trees}}) \tag{13}$$

XGBoost, a boosting algorithm, enhances model performance by iteratively combining weak learners:

$$\hat{y} = \sum_{i=1}^{N} \text{learner}_i(x) \tag{14}$$

3.5 Ensemble Learning Approach

Ensemble learning combines multiple models to improve prediction accuracy. In this research, we used an ensemble of XGBoost and RF models to predict CHF using RR peaks and intervals. XGBoost excels in capturing complex relationships, while RF reduces overfitting by constructing an ensemble of decision trees. Combining these models into an ensemble leverages their complementary strengths, leading to a robust and accurate prediction mechanism.

Soft Voting The Voting Classifier employs soft voting, where each classifier contributes class probabilities. The final prediction ($\hat{y}$) is determined by the weighted sum of the predicted probabilities:

$$\hat{y} = \arg\max_i \sum_j w_j P_{ij} \tag{15}$$

Weight Assignment Weights (w_j) are assigned based on the classifiers' performance, often using the inverse of the classification error (E_j):

$$w_j = \frac{1}{1 - E_j} \tag{16}$$

3.6 Overfitting Concerns and Mitigation

High accuracy levels in ML models often raise concerns about potential overfitting. To ensure that our model generalizes well and does not merely memorize the training data, we implemented several strategies to address and mitigate overfitting.

Cross-Validation We employed k-fold cross-validation where the dataset is divided into k equally-sized folds. The model is trained on k-1 folds and tested on the remaining folds. This process is repeated k times, with each fold serving as the test set once. The average performance across all folds provides a reliable estimate of the model's performance on unseen data. The results of our 10-fold cross-validation are presented in Table 2.

Table 2. Cross-Validation Results

Fold	Accuracy	Precision	Recall	F1-Score
1	99.50%	99.48%	99.52%	99.50%
2	99.55%	99.53%	99.56%	99.54%
3	99.60%	99.58%	99.61%	99.59%
4	99.52%	99.50%	99.53%	99.51%
5	99.58%	99.56%	99.59%	99.57%
6	99.54%	99.52%	99.55%	99.53%
7	99.57%	99.55%	99.58%	99.56%
8	99.61%	99.59%	99.62%	99.60%
9	99.53%	99.51%	99.54%	99.52%
10	99.56%	99.54%	99.57%	99.55%
Average	**99.56%**	**99.54%**	**99.57%**	**99.55%**

Performance on Separate Test Set We evaluated our model on a separate test set that was not used during training or validation. The test set performance metrics are consistent with the cross-validation results, confirming that the model maintains high performance on unseen data. The test set results are presented in Table 3.

Table 3. Test Set Performance

Metric	Accuracy	Precision	Recall	F1-Score
Value	99.56%	99.63%	99.18%	99.41%

4 Result and Discussion

This section evaluates our model's performance compared to existing approaches. The model achieves high accuracy with minimal computational time, validated through common benchmark settings. We assess the model using four key metrics: accuracy, precision, recall, and F1-score, providing a comprehensive evaluation. **Table** 4 compares the performance of classifiers, including XGBoost, KNN, RF, SVM, and the Ensemble Model, across these metrics. The Ensemble Model combines the strengths of RF and XGBoost, achieving the highest accuracy at 99.56%, with a precision of 0.9963, recall of 0.9918, and F1 score of 0.9941. This model significantly reduces misclassifications and enhances overall performance and generalization, making it the most robust among the evaluated classifiers.

Table 4. Performance Metrics of Different Classifiers

Classifier	Accuracy (%)	Precision	Recall	F1 Score	Explanation/Inference
XGBoost	98.59	0.9896	0.9731	0.9813	The XGBoost classifier shows high accuracy and balanced precision and recall, indicating robust performance in classifying both CHF and NSR segments.
KNN	98.17	0.9804	0.9551	0.9676	The KNN classifier achieves strong accuracy with slightly lower precision and recall compared to XGBoost, indicating effective classification but with some margin for improvement in precision.
RF	99.36	0.9909	0.9941	0.9925	The RF classifier demonstrates near-perfect accuracy with excellent precision and recall, highlighting its capability to accurately classify both CHF and NSR segments with minimal misclassifications.
SVM	95.48	0.9689	0.8682	0.9151	The SVM classifier shows good accuracy but lower recall compared to other models, indicating a higher rate of misclassifications for CHF segments despite effective NSR segment classification.
Ensemble Model	99.56	0.9963	0.9918	0.9941	The ensemble model combines the strengths of RF and XGBoost, achieving high accuracy and balanced precision and recall. This model significantly reduces misclassifications and enhances overall performance and generalization.

The confusion matrix in Figure 5 illustrates the performance of the ensemble model with True Positives (TP) of 2173, True Negatives (TN) of 4128, False Positives (FP) of 8, and False Negatives (FN) of 18. This indicates the model's high

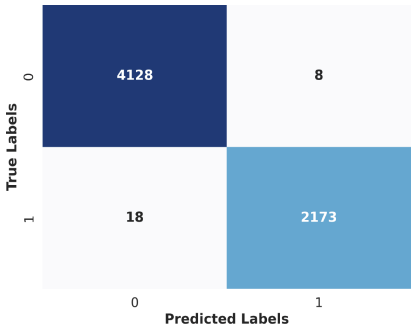


Fig. 5. Confusion Matrix from classification using the ensemble model

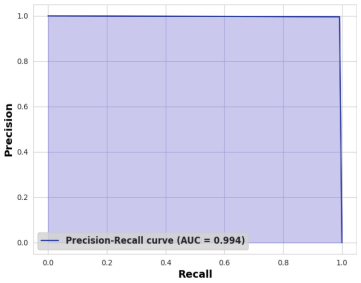


Fig. 6. Precision-Recall Curve for the ensemble model

accuracy in correctly identifying both CHF and NSR segments. The Precision-Recall Curve in Figure 6 demonstrates the ensemble model's high precision and recall performance, achieving an Area Under the Curve (AUC) of 0.994. This highlights the model's effectiveness in distinguishing between CHF and NSR segments across different thresholds.

4.1 Comparison of Model Accuracy and Time Complexity

In real-time detection of critical diseases, accuracy and time play a vital role in early diagnosis and accurate prediction in rural areas. We perform the comparative analysis of accuracy and time taken by models based on benchmark settings. The models are executed over i5 intel processor with 40 core, 16 GB RAM, 512 GB hard drive under Windows operating system. Table 5 summarizes the accuracy and time complexity comparison of different ML models.

Table 5. Comparison of Accuracies and Time Complexity of All Models

Classifier Model	**Accuracy (%)**	**Time Taken (seconds)**	**Time Complexity (O())**
SVM	95.48	3.85	$O(n_{\text{samples}}^2 \times n_{\text{features}})$
KNN	98.17	0.34	$O(n_{\text{features}} \times n_{\text{samples}} \log(n_{\text{samples}}))$
RF	99.36	1.45	$O(n_{\text{features}} \log(n_{\text{features}}) \times n_{\text{features}} \times Trees)$
XGBoost	98.59	0.21	O(n × m × log(m))
Ensemble	99.56	1.63	O(n × m × log(m))

From the time complexity comparison, it's evident that the KNN and XGBoost classifiers have the lowest time requirements, making them efficient choices for real-time applications. The SVM classifier, although accurate, has

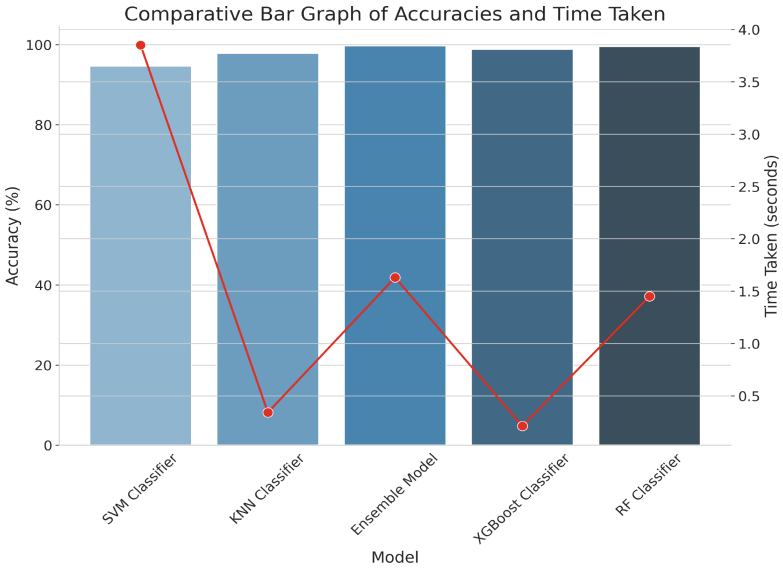


Fig. 7. The diagram represents the performance of the models: first SVM, second KNN, third RF, fourth XGBoost, and fifth Ensemble model.

a higher time complexity, which could impact its suitability for time-sensitive tasks. The ensemble model shows slightly higher time requirements due to the combined training and prediction processes of XGBoost and RF. Figure 7 illustrates the accuracy and time taken for each model, providing a visual comparison of their performance metrics.

4.2 Comparison with Existing Research

Numerous investigations have examined the application of ML methodologies for CHF prediction using ECG data. We contrast the performance of our proposed ensemble learning approach with that of existing research studies as summarized in Table 6. Our proposed ensemble learning model achieves an accuracy of 99.56%, outperforming existing methods in CHF prediction. Our approach demonstrates effectiveness and robustness in accurately classifying ECG segments, as evidenced by high recall, precision, and F1-score values for both CHF and non-CHF classes.

4.3 Comparative Analysis of Traditional and Deep Learning Methods for CHF Detection

This study compares traditional ML methods with DL approaches for early CHF detection using ECG data, focusing on performance, interpretability, and practical applicability in clinical settings. The comparison is summarized as follows:

Table 6. Result Analysis

Ref.	Dataset used	Methodology used	Accuracy
[2]	NSR-RR, CHF-RR	Unet++	89.83%
[6]	MIT-BIH, BIDMC CHF	Faster RCNN classifier	98%
[8]	BIDMC-CHF, PTBDB	DA-DRRNet	98.57%
[9]	Mendeley data source	Eigendomain DRL approach	98.68%
[12]	from Catholic University of Leuven	Artificial Neural Network	90.00%
[20]	MIT-BIH, BIDMC CHF	ECG-Convolution-Vision Transformer Network	98.88%
[27]	MIT-BIH, BIDMC CHF	Stockwell Transform and Hybrid Classification Scheme	98.78%
Proposed	**MIT-BIH, BIDMC CHF**	**Ensemble Learning**	**99.56%**

1. **Performance Metrics:** Both traditional and DL models were evaluated using metrics like accuracy, precision, recall, F1-score, and AUC-ROC. While DL models often show slightly higher accuracy and AUC-ROC, traditional methods remain competitive, particularly with smaller datasets.
2. **Interpretability:** Traditional methods, such as decision trees and random forests, offer clear interpretability, crucial in clinical settings for gaining trust and acceptance among healthcare professionals. Conversely, DL models, though accurate, operate as "black boxes," presenting a challenge for clinical adoption due to their lack of transparency.
3. **Data Requirements:** Traditional methods perform well even with limited data, making them more practical in situations where large ECG datasets are hard to obtain, unlike DL models which require vast amounts of data.
4. **Computational Efficiency:** Traditional methods are computationally efficient and can be implemented on standard medical equipment, making them more accessible and cost-effective, especially in resource-constrained settings.

In conclusion, while DL methods offer higher accuracy, traditional methods excel in interpretability, data efficiency, and computational requirements. The choice of method should balance accuracy with clinical needs, data availability, and computational resources.

5 Conclusion and Future Directions

This research focuses on the urgent need for precise CHF detection, vital for prompt intervention. Our proposed ensemble learning and fusion model, utilizing

ECG data, demonstrates promising early CHF detection. Integration of diverse ML algorithms and ECG-derived features enhances prediction accuracy and efficiency. Fusion techniques further boost performance by integrating information from multiple sources, validated with real-world datasets. Our comprehensive approach establishes a robust groundwork, achieving remarkable predictive accuracy (99.56%) in distinguishing NSR from CHF cases. We also evaluated each model's time complexity, which is a crucial parameter for real-time clinical applications. Our all-encompassing method not only solves the pressing requirement for precise CHF identification but also establishes the foundation for further improvements. Future research aims to refine and extend this model for timely MI detection, enhancing cardiovascular health monitoring in clinical practice.

References

1. Savarese, G., et al.: Global burden of heart failure: a comprehensive and updated review of epidemiology. In: Cardiovasc. Res., pp. 3272–3287. (2023). https://doi.org/10.1093/cvr/cvac013. Erratum. In: Cardiovasc Res. 2023 Jun 13;119(6):1453 PMID: 35150240
2. L. Zou,et al. "Automatic Detection of Congestive Heart Failure Based on Multiscale Residual UNet++: From Centralized Learning to Federated Learning," in IEEE Transactions on Instrumentation and Measurement, vol. 72, pp. 1-13, 2023, Art no. 4001013, https://doi.org/10.1109/TIM.2022.3227955
3. Malik A, et al. Congestive Heart Failure. [Updated 2023 Nov 5]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK430873/
4. A. Shrivastava et al., "A Novel Hybrid Model for Predictive Analysis of Myocardial Infarction using Advanced Machine Learning Techniques," 2023 OITS International Conference on Information Technology (OCIT), Raipur, India, 2023, pp. 381-386, https://doi.org/10.1109/OCIT59427.2023.10430780
5. Ning, et al.: Automatic detection of congestive heart failure based on a hybrid deep learning algorithm in the internet of medical things. IEEE Internet Things J. **8**(16), 12550–12558 (2020)
6. S. Irin Sherly et al., An efficient honey badger based Faster region CNN for chronic heart Failure prediction, Biomedical Signal Processing and Control, Volume 79, Part 2,2023
7. Baral, et. al. "A Literature Review for Detection and Projection of Cardiovascular Disease Using Machine Learning." EAI Endorsed Transactions on Internet of Things 10 (2024)
8. Prabhakararao, E., et al.: Congestive Heart Failure Detection From ECG Signals Using Deep Residual Neural Network. IEEE Transactions on Systems, Man, and Cybernetics: Systems **53**(5), 3008–3018 (2023). https://doi.org/10.1109/TSMC.2022.3221843
9. Bhaskarpandit, et al.: Detection of Myocardial Infarction From 12-Lead ECG Trace Images Using Eigendomain Deep Representation Learning. IEEE Trans. Instrum. Meas. **72**, 1–12 (2023)
10. Rani et al. (2024). An Extensive Review of Machine Learning and Deep Learning Techniques on Heart Disease Classification and Prediction. Archives of Computational Methods in Engineering, 1-19

11. A. S et al. "ECG Classification and Arrhythmia Detection Using Wavelet Transform and Convolutional Neural Network," 2021 International Conference on Communication, Control and Information Sciences (ICCISc), Idukki, India, 2021, pp. 1-5, https://doi.org/10.1109/ICCISc52257.2021.9485012
12. Bibicu, D., et al.: Cardiac Cycle Phase Estimation in 2-D Echocardiographic Images Using an Artificial Neural Network. IEEE Trans. Biomed. Eng. **60**(5), 1273–1279 (2013). https://doi.org/10.1109/TBME.2012.2231864
13. Goldberger, et al.: PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. **101**(23), e215–e220 (2000)
14. Ju, R., et al.: 3D-CNN-SPP: A Patient Risk Prediction System From Electronic Health Records via 3D CNN and Spatial Pyramid Pooling. IEEE Transactions on Emerging Topics in Computational Intelligence **5**(2), 247–261 (2021). https://doi.org/10.1109/TETCI.2019.2960474
15. Karami, E., et al.: Adaptive Polar Active Contour for Segmentation and Tracking in Ultrasound Videos. IEEE Trans. Circuits Syst. Video Technol. **29**(4), 1209–1222 (2019). https://doi.org/10.1109/TCSVT.2018.2818072
16. Sharma, 2023, December). Optimizing Knowledge Transfer in Sequential Models: Leveraging Residual Connections in Flow Transfer Learning for Lung Cancer Classification. In *Proceedings of the Fourteenth Indian Conference on Computer Vision, Graphics and Image Processing* (pp. 1-8)
17. Ortiz-Gonzalez, A., et al.: Optical Flow-Guided Cine MRI Segmentation With Learned Corrections. IEEE Trans. Med. Imaging **43**(3), 940–953 (2024). https://doi.org/10.1109/TMI.2023.3325766
18. D. Morillo-Velepucha, et al. "Congestive heart failure prediction based on feature selection and machine learning algorithms," 2022 17th Iberian Conference on Information Systems and Technologies (CISTI), Madrid, Spain, 2022, pp. 1-6, https://doi.org/10.23919/CISTI54924.2022.9820312
19. Melillo, P., et al.: Classification Tree for Risk Assessment in Patients Suffering From Congestive Heart Failure via Long-Term Heart Rate Variability. IEEE J. Biomed. Health Inform. **17**(3), 727–733 (2013). https://doi.org/10.1109/JBHI.2013.2244902
20. Kaiser, A., "Towards a method for early detection of congestive heart failure with an electrocardiogram and acoustic transducers,âĂİ, et al.: IEEE Signal Processing in Medicine and Biology Symposium (SPMB). New York, NY, USA **2012**, 1–5 (2012). https://doi.org/10.1109/SPMB.2012.6469460
21. Zhang, Y., "Congestive Heart Failure Detection Via Short-Time Electrocardiographic Monitoring For Fast Reference Advice In Urgent Medical Conditions,âĂİ, et al.: 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). Honolulu, HI, USA **2018**, 2256–2259 (2018). https://doi.org/10.1109/EMBC.2018.8512888
22. Valenza, G., et al.: Mortality Prediction in Severe Congestive Heart Failure Patients With Multifractal Point-Process Modeling of Heartbeat Dynamics. IEEE Trans. Biomed. Eng. **65**(10), 2345–2354 (2018). https://doi.org/10.1109/TBME.2018.2797158
23. Mei-Yi Wu, et al.,Radio-contrast medium exposure and dialysis risk in patients with chronic kidney disease and congestive heart failure: A case-only study,International Journal of Cardiology,Volume324,2021,Pages 199-204,ISSN 0167-5273, https://doi.org/10.1016/j.ijcard.2020.09.014.
24. J. Zhang et al., "MLBF-Net: A Multi-Lead-Branch Fusion Network for Multi-Class Arrhythmia Classification Using 12-Lead ECG," in IEEE Journal of Translational

Engineering in Health and Medicine, vol. 9, pp. 1-11, 2021, Art no. 1900211, https://doi.org/10.1109/JTEHM.2021.3064675
25. V. Jahmunah,et al.Computer-aided diagnosis of congestive heart failure using ECG signals – A review, Physica Medica,Volume 62,2019,Pages 95-104,ISSN 1120-1797, https://doi.org/10.1016/j.ejmp.2019.05.004
26. Hou, B., et al.: LSTM-Based Auto-Encoder Model for ECG Arrhythmias Classification. IEEE Trans. Instrum. Meas. **69**(4), 1232–1240 (2020). https://doi.org/10.1109/TIM.2019.2910342
27. R.K. Tripathy, et al., Automated detection of congestive heart failure from electrocardiogram signal using Stockwell transform and hybrid classification scheme, Computer Methods and Programs in Biomedicine, Volume 173,2019, Pages 53-65,ISSN 0169-2607, https://doi.org/10.1016/j.cmpb.2019.03.008
28. Guan,et al., "CobNet: Cross Attention on Object and Background for Few-Shot Segmentation." In 2022 26th International Conference on Pattern Recognition (ICPR), pp. 39-45. IEEE, 2022
29. F. De Marco, D. Finlay and R. R. Bond, "Classification of Premature Ventricular Contraction Using Deep Learning," 2020 Computing in Cardiology, Rimini, Italy, 2020, pp. 1-4, https://doi.org/10.22489/CinC.2020.311
30. Porumb, M., Iadanza, E., Massaro, S., Leandro Pecchia, A., convolutional neural network approach to detect congestive heart failure, Biomedical Signal Processing and Control, Volume 55,: 101597. ISSN **1746–8094**,(2020). https://doi.org/10.1016/j.bspc.2019.101597
31. De Marco, F., Ferrucci, F., Risi, M., Tortora, G.: Classification of QRS complexes to detect Premature Ventricular Contraction using machine learning techniques. PLoS ONE **17**(8), e0268555 (2022). https://doi.org/10.1371/journal.pone.0268555

Hierarchical Classification of Frontotemporal Dementia Subtypes Utilizing Tabular-to-Image Data Conversion with Deep Learning Methods

Km Poonam(✉), Venkata Sathwik Kotra, Rajlakshmi Guha, and Partha P. Chakrabarti

Indian Institute of Technology Kharagpur, Kharagpur 721302, West Bengal, India
poonamk@iitkgp.ac.in, sathwik1631@kgpian.iitkgp.ac.in, rajg@cet.iitkgp.ac.in, ppchak@cse.iitkgp.ac.in

Abstract. Frontotemporal dementia (FTD) is a group of neurodegenerative disorders characterized by progressive damage to the frontal and temporal lobes of the brain. FTD includes notable changes in social behavior, personality, and language function difficulties. The primary FTD spectrum disorders include behavioral variant FTD, non-fluent variant primary progressive aphasia, and semantic variant. With an aging population, FTD prevalence increases, emphasizing the critical need for research, care, and awareness. This study introduces novel hierarchical classification methods that enhance model prediction capabilities. The dataset used in this work consists of neuropsychiatric tabular data from the Frontotemporal Lobar Degeneration Neuroimaging Initiative database. Innovative methods are needed to effectively leverage deep learning for FTD classification, especially when limited to tabular data. To address this challenge, we proposed an innovative technique, the Tensorised Image Generator (TIG), which transforms tabular data into grid-based image representations or tensors. The TIG algorithm optimizes tabular data visualization by ranking pairwise feature distances and then creating a distance matrix and strategically positioning features in the image to preserve spatial correlation among features. The algorithm further enhances the grid structure with line drawing and intensity adjustments. The results on these images demonstrate high accuracy in detecting FTD subtypes using a Convolutional Neural Networks ensemble with multi-layer perceptron, achieving a test accuracy of 88.89%, which is a 2% increment of hierarchical approach over flat machine learning methods and around 5% increment for combined hierarchical classification with image-based technique compared to best flat method across most metrics (accuracy, precision, recall, and F1-score).

Keywords: Frontotemporal Dementia · Hierarchical Classification · Tabular-to-Images Conversion · Deep-Learning

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15311, pp. 386–401, 2025.
https://doi.org/10.1007/978-3-031-78195-7_26

1 Introduction

Frontotemporal dementia (FTD) presents a significant challenge in the field of neurodegenerative diseases, characterized by a progressive deterioration in behavior, language, and executive functions [1]. Early and accurate classification of FTD subtypes is crucial for effective management and personalized treatment strategies. Frontotemporal dementia has primarily three clinical subtypes, namely, Behavioral Variant Frontotemporal Dementia (BV), Non-fluent Variant Primary Progressive Aphasia (PNFA), and Semantic Variant Primary Progressive Aphasia (SV) [2][3]. The diagnosis of FTD does not rely on a single test. Clinicians must carefully examine the patient's symptoms and signs while also ruling out other potential causes. Early FTD diagnosis can be challenging, as the symptoms often overlap with those of other disorders. To assess the patient's condition, clinicians typically conduct extensive testing of memory, reasoning, and other cognitive functions using neuropsychiatric inventories. In the given dataset, we have neuropsychological test scores.

Recent advancements in deep learning have revolutionized the field of medical image analysis. Convolutional Neural Networks (CNNs), in particular, have demonstrated remarkable success in extracting meaningful features from image data, leading to significant improvements in disease diagnosis and prognosis. However, traditional CNN approaches rely on image data, limiting their applicability to tasks where image-based information is readily available. In this context, there exists a need for innovative methodologies that can effectively leverage Deep Learning (DL) techniques for FTD subtype classification, even in scenarios where only tabular data is available. One approach to applying deep learning to tabular data involves transforming tabular data into images. Currently, several converters are accessible for this purpose [8][19]. Some of these converters assign individual pixels in an image to represent each sample of the dataset, though the arrangement of these pixels can differ between converters. For instance, the method suggested in study [19] maps each feature to a grayscale pixel based on its corresponding coordinates in a reduced 2D space, with pixel intensity reflecting the features. In contrast, the approach by Zhu et al. [8] organizes pixels based on distance metric, placing features with shorter distances closer together in the generated grayscale image. Bazgir et al. [20] employ a blend of distance metrics and dimensionality reduction techniques to assign grayscale pixel values to individual features. We propose an innovative method that transforms tabular data into images to utilize image-based DL models for FTD classification.

In medical practice, a single-subject classification model would be more informative than a group analysis because of the FTD subtype patients' subtle patterns and overlying patterns. This argument motivated us to design an automated hierarchy-based classification model to analyze each subject at an individual level. In a previous study [4], the authors used a support vector machines classifier on cortical thickness data to compute an individual network for each subject. Based on this analysis, each subject was then classified into one of the diagnostic categories: mild cognitive impairment, Alzheimer's disease, or cognitively normal. Additionally, some other studies [5][6] have investigated the use of AI methods to discriminate between the FTD and CN groups using MRI data.

In this work, we have presented a novel method of converting tabular data into grid-based images, followed by hierarchical classification of FTD subtypes in which we perform the classification on different levels that will classify each subject into one of the class labels (CN, BV, PNFA, and SV). This approach aligns with a range of ML/DL methods and imitates the natural decision-making process of clinicians. If a clinician detects any abnormal findings beyond those explained by typical aging in the preliminary assessment, they must rule out dementia. Similarly, clinicians typically deduce the disorder through logical inference when a subject exhibits behavioral or linguistic changes suggestive of FTD. In the suggested method, first, cognitively normal subjects are separated from FTD patients. Further, the group of FTD patients is classified into behavioral variant and primary progressive aphasia. Eventually, the primary progressive aphasia group classifies subjects into non-fluent aphasia and semantic variants. By transforming tabular data into image representations, we aim to harness the power of CNNs for the hierarchical classification of FTD subtypes, thereby enhancing diagnostic accuracy and improving patient outcomes. This study contributes in the following ways:

- **A New Method For Tabular Data to Images Conversion:** We propose an approach to enhance the representation of tabular data by converting it into grid-based images, which enables better visualization and representation of the spatial correlation among features within the data. Leveraging the power of CNNs and pre-trained models (ResNet-50 and VGG-16) for the classification task.
- **Hierarchical Classification:** Our research conducts hierarchical classification on the generated images and compares the results with direct 4-class classification to evaluate performance improvements.

The results demonstrate high accuracy in detecting FTD subtypes using a CNN ensemble with MLP, achieving a test accuracy of 88.89%, representing almost a 5% increase over the non-hierarchical (flat) best-known method. The proposed hierarchical approach generally outperforms the flat approach across most metrics (accuracy, precision, recall, and F1-score) and models.
The rest of the paper is organized as follows: the “Materials” section presents the dataset and data preprocessing methods, and the "Proposed Methodology" section demonstrates the proposed framework, algorithm, and model training and prediction. The "Results and Discussion" section presents the findings from the experiments. Subsequently, the study’s conclusions are deliberated in the "Conclusion" section.

2 Material

2.1 Dataset

The dataset used in this study was retrieved from the Frontotemporal Lobar Degeneration Neuroimaging Initiative (FTLDNI). The dataset contains 288 (CN (134), BV (75), SV (39), and PNFA (40)) subjects for baseline; a detailed feature description study is presented in [15].

2.2 Data Pre-processing

We selected 36 different features, each representing a unique neuropsychological test. These features have been carefully selected from a larger dataset, likely due to having the least missing values. This selection ensures data completeness and reliability of the analysis, offering valuable insights into the cognitive and behavioral functions of the patients. We employed a mean imputer technique to handle missing data. This technique involves replacing missing values in the dataset with the mean value of the corresponding feature.

3 Proposed Methodology

This work aims to transform tabular neuropsychiatric data into image representations and then use these images to build a predictive model for classifying the clinical status of FTD subtypes. Our proposed system consists of two main tasks: first, converting tabular data to images by utilizing the Image Generator for Tabular Data (IGTD) method [8] and introducing the new Tensorised Image Generator (TIG) algorithm to transform tabular data into images, and second, employing a hierarchical classification approach to categorize individuals into diagnostic labels, CN, BV, PNFA, and SV. We applied DL methods to the image data, taking advantage of DL's powers in image recognition. We employ ML techniques to the original tabular data, harnessing the power of the ML. Additionally, we performed a flat classification, directly assigning individuals to one of the four diagnostic labels (CN, BV, PNFA, and SV) using both image and tabular data. This flat classification serves as a benchmark for comparing the performance of our hierarchical approach. The proposed framework depicted in Fig. 1 offers a novel way to handle neuropsychiatric data, combining data representation techniques with DL models to improve diagnostic accuracy and provide valuable insights for clinical decision-making.

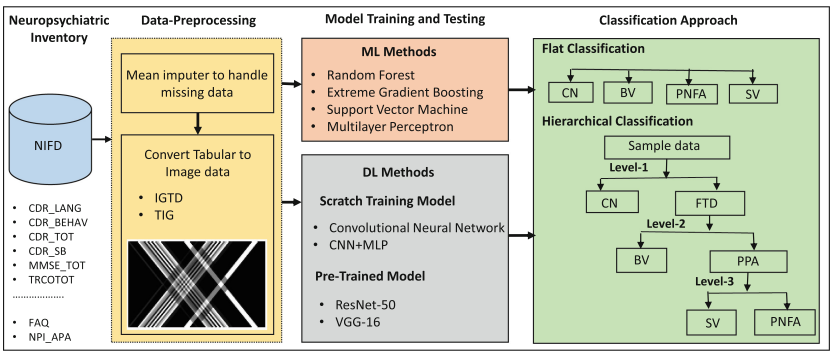


Fig. 1. Proposed framework for the development process. IGTD: Image Generator for Tabular Data, TIG: Tensorised Image Generator.

3.1 Conversion of Tabular Data to Images

The main aim of our study is to use the power of CNNs over the tabular data for classification tasks. However, CNNs work well only on images. So, each data sample is converted into a single image. We used the algorithm Image Generator for Tabular Data (IGTD) [8] to generate images for tabular data, and from these generated images, we have created our own image representations by using the Tensorised Image Generator (TIG) algorithm; it involves transforming the tabular data into a grid-based image representation, which can be viewed as a tensor that better represents the data.

Image Generator for Tabular Data (IGTD): This algorithm [8] converts tabular data into images by strategically assigning features to specific pixel positions. This assignment ensures that similar features are positioned near one another within the resulting image. The algorithm aims to find an optimal assignment by reducing the gap between the ranking of distances among features and their corresponding pixel placements within the image, and the intermediate sub-figures are shown in Fig. 3. The algorithm is to transform each sample x_i into an $N_r \times N_c$ image where $N_r \times N_c$=N where N is the number of features in the dataset. Pairwise distances between features are computed using a distance metric like Euclidean distance. Subsequently, these distances are ranked so small distances receive lower ranks while longer distances are assigned higher ones. Let R be a Rank Matrix of size $N \times N$; each cell (i, j) represents the rank value of the distance between i^{th} and j^{th} feature. When this is transformed into an image, the grey level in the image indicates the rank, i.e., the larger the distance, the higher the rank. On the other hand, Q is a $N \times N$ Rank matrix of pixel distances in the image initially. To transform the tabular data into images each feature needs to be assigned to a pixel position of the image. In order to find the optimal positioning of features, we have designed an error function (eq. 1) and minimized the error function so that most related features will be together in the image generated.

$$error(R, D) = \Sigma_{i=2}^{n} \Sigma_{j=1}^{n} diff(r_{i,j}, q_{i,j}) \tag{1}$$

Tensorised Image Generator (TIG): The IGTD algorithm creates an image from a data sample with n features, arranging related features close together and unrelated features farther apart. The goal is to show how each feature i relates to all other features by drawing lines in a grid. These lines intersect at points that indicate how closely related the features are. To do this, we draw lines diagonally in the grid, using pixel intensities to represent the features. Each line is two pixels wide, and we mark the i^{th} row and j^{th} column with the feature's pixel intensity. The result is a set of intersecting lines that form triangular patterns, as shown in Fig. 2, illustrating the relationships between features.

In our proposed TIG images, we use a sorted order for the features and draw lines with intensities matching those from the IGTD algorithm. The line width is adjustable; here, we've chosen 2 pixels. The size of the grid (and thus the final image) depends on the number of features and the line width, calculated as $(N_f \times N_p \times 2) - N_p$, where N_f is the number of features and N_p is the line width in pixels. The lines are drawn to intersect within a specific triangular

area, ensuring a clear visualization of the relationships. The Sample image of the IGTD and TIG over the dataset is shown in Fig. 4.

Algorithm 1: Image Generator for Tabular Data (IGTD) [8]

Input: Tabular data features
Output: Generated image representation

Compute pairwise distances between features using a distance measure Euclidean distance, and a larger distance represents a higher rank;
Rank the pairwise distances to create a Rank Matrix R of size $N \times N$;
Initialize a rank matrix Q of pixel distances in the image to be transformed;
Assign each feature to a pixel position of the image;
Define an error function $error(R, D)$ to minimize the discrepancy between the ranking of distances among features and the ranking of distances among their corresponding pixels in the image;
for $i = 2$ *to* n **do**
    for $j = 1$ *to* n **do**
        Compute the difference between rank values $r_{i,j}$ and $q_{i,j}$;
        Increase the error with the difference $diff(r_{i,j}, q_{i,j})$;
    end
end
Minimize the error function to optimize the positioning of features in the image;
Generate the image representation by transforming the tabular data into an $N_r \times N_c$ image, where $N_r \times N_c = N$ and N is the number of features in the dataset.

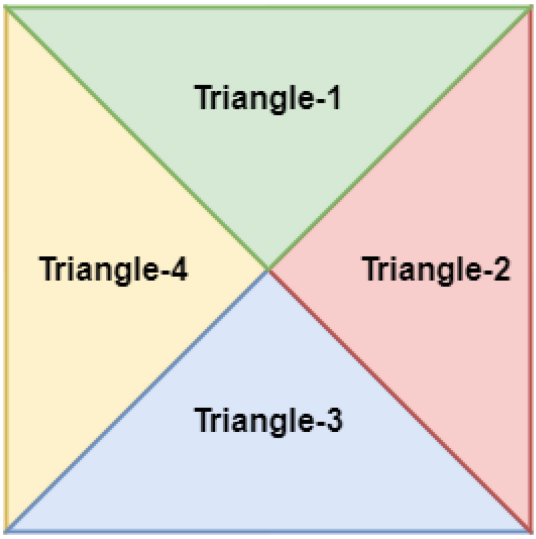


Fig. 2. Representation of the Image Grid

Algorithm 2: Proposed Tensorised Image Generator (TIG) Algorithm

Input: Index: list of features to include in the grid
Intensity: intensity value to assign to each pixel
Array: 2D array representing the grid
No_features: number of features
No_of_pixels: number of pixels
Break_point: breakpoint value
Output: grid: final generated grid image

```
for each IGTD image do
    for each feature in the image do
        Determine the index of the feature and get its intensity value;
        Generate the grid by calling the create_grid function for both
        "right" and "down" directions;
    end
end
for each row in the grid do
    for each column in the grid do
        if pixel value is -1 then
            Set it to 0;
        end
    end
end
Function create_grid(arr, feature, direction, pixels, breakpoint,
intensity):
if direction is "right" then
    for i from 1 to breakpoint + 1 do
        for j from 0 to pixels do
            if arr[i][j + pixels*feature] is not -1 then
                arr[i][j + pixels*feature] = average(arr[i][j + pixels*feature],
                intensity);
            else
                arr[i][j + pixels*feature] = intensity;
if direction is "down" then
    for i from breakpoint to 1 do
        for j from 0 to pixels do
            if arr[i][j + pixels*feature] is not -1 then
                arr[i][j + pixels*feature] = average(arr[i][j + pixels*feature],
                intensity);
            else
                arr[i][j + pixels*feature] = intensity;
Return: modified grid image;
```

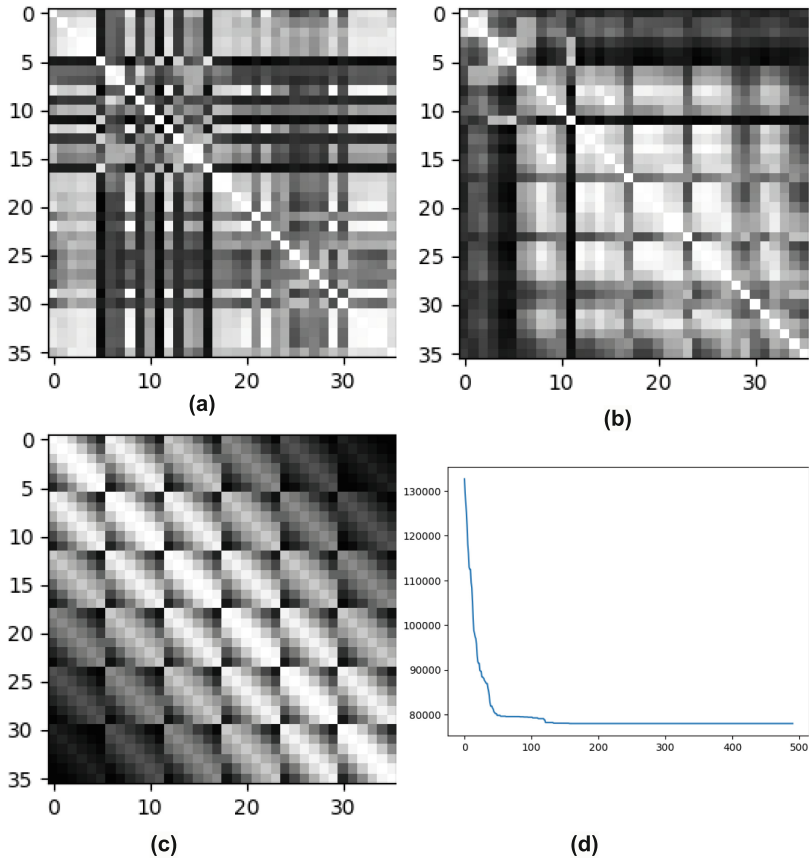
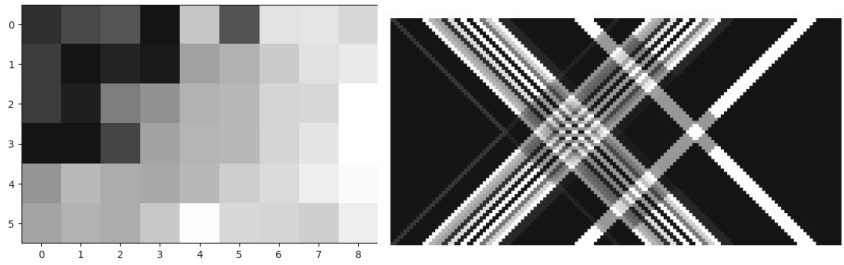


Fig. 3. Distance matrices generated by the IGTD. (a) a ranking matrix of feature distances calculated using Euclidean distance; (b) a ranking matrix of feature distances post-optimization; (c) a ranking matrix of pixel distances calculated using Euclidean distance; (d) Error minimization graph, the difference between matrices (a) and (c).



(a) Image generated using IGTD. (b) Image generated using TIG

Fig. 4. Image representation generated from tabular data.

3.2 Convolutional Neural Networks

We adapted the Convolutional Neural Networks (CNNs) architecture with two convolutional layers, two max-pooling layers, a flattening layer, and a Fully Connected (FC) layer represents a common and effective deep-learning model for image classification tasks [9]. In the architecture presented in Fig. 5, the two convolutional layers perform feature extraction by applying a set of learnable filters to the input image. Each convolutional layer is followed by a max-pooling layer, which reduces the spatial dimensions of the feature maps and retains the most relevant information. The flattening layer transforms the 2D feature maps into a 1D vector, preparing the data for the subsequent FC layer. The fully connected layer is the final layer in the network, takes the flattened features, and performs classification, making it capable of distinguishing between various class labels. This architecture is well-suited for tasks like image recognition and is known for its ability to automatically learn hierarchical features from the input data, making it a crucial component in computer vision applications. Since we are fine-tuning the model as per the dataset, there is no fixed number of kernels and kernel sizes, so the number of kernels and kernel sizes is represented with variables in the image.

3.3 Ensemble CNN with MLP

We integrated CNN-extracted features with raw tabular data, leveraging the strengths of both image-based and tabular representations for classification tasks. Initially, the CNN processed each image to extract high-level features, capturing intricate patterns and characteristics. Subsequently, these extracted features were combined with the neuropsychological test score data, forming concatenated datasets enriched with image-derived information. This concatenated data was then fed into an MLP [14], which effectively utilized the fused features to perform classification for flat and hierarchical approaches.

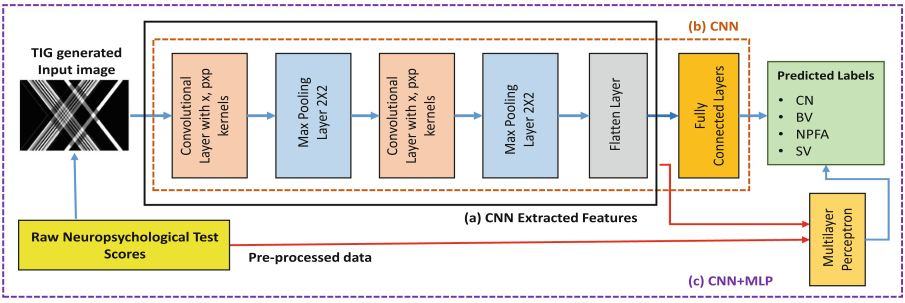


Fig. 5. Covolutional Neural Networks (CNNs) ensemble with Multilayer Perceptron (MLP) architecture.

Model Training: We used the grid-search cross-validation approach to find the optimal hyperparameters. Grid search was conducted using a 5-fold cross-validation strategy, where the dataset was divided into 5 subsets. The model was trained and evaluated 5 times, with each iteration utilizing a different subset as the validation set while the rest were employed for training. Model hyperparameter tuning was conducted by evaluating the error between the predicted and actual outputs, employing binary cross-entropy, and sparse categorical cross-entropy was chosen as the loss function for hierarchical and flat classification, respectively. Updating a neural network's weights and biases during training involves propagating the error backward from the output layer to the input layer. This backpropagation process calculates the gradient of the loss function with respect to each parameter in the network using the chain rule of calculus. By iteratively adjusting these parameters in the direction that minimizes the loss, the network learns to make better predictions over time. Table 1 shows the optimal hyper-parameters used in the model's training and validation.

Model Complexity: The model's selected hyperparameters include the number of convolution layers, dense layers, pool size, filter dimensions, number of filters, and kernel/filter size. Among these operations, convolution, pooling, and dense layers hold significance. Each layer's parameter count represents the number of learnable elements. While the input layer determines the shape without learnable parameters, pooling layers reduce feature map dimensions and parameter count, which reduces computational complexity. Nevertheless, pooling layers lack trainable parameters. Studies have highlighted the sensitivity of batch size and learning rate in influencing computational complexity [13]. Larger batch sizes elevate model computational complexity. Therefore, multiplying these parameters is essential for determining the model's precise complexity. The convolution layer's time complexity can be computed according to eq. 2.

$$(\Sigma_{n=1}^{d} k_{n-1} \cdot s_n^2 \cdot f_n \cdot l_n^2) \cdot r_l \cdot b_l \tag{2}$$

In this context, d denotes the convolutional layer's depth, l_n represents the output feature map's length, and f_n indicates the number of filters in the n^{th} layer. Additionally, s_n signifies the filter's length, k_{n-1} defines the number of input channels in the $(l)^{th}$ layer, r_l represents the learning rate, and b_l denotes the batch size.

In a fully connected layer, every parameter is interconnected with each other. This layer serves to link preceding layers with the output layer and comprises varying numbers of neurons, impacting the output size. Therefore, it is necessary to calculate the parameters for each FC layer, including factors like the input image's height and width, the number of neurons, and the input dimension. Finally, the total complexity of all FC layers in the model is obtained by summing up the complexities of individual layers. The time complexity of the fully connected layer can then be determined using eq. 3.

$$\Sigma_{l=1}^{f} D \cdot W \cdot H \cdot N \tag{3}$$

Here, l represents FC layer depth; D, W, H, and N define the dimension of the input/output channel, the width of the input, the height of the input, and the number of outputs, respectively. The time complexity for the MLP is defined as the time taken to extract features from CNN and the fully connected layer used for classification:

$$((\Sigma_{n=1}^{d} k_{n-1} \cdot S_n^2 \cdot f_n \cdot l_n^2) \cdot r_1 \cdot b_1) + (\Sigma_{l=1}^{f} D \cdot W \cdot H \cdot N) \tag{4}$$

Table 1. Approximated hyperparameters for 2D CNN ensemble with MLP model training.

Hyperparameter	Search Space	Optimal Value
Learning Rate	[0.01, 0.001, 0.0001]	0.001, 0.001
Batch Size	[8, 10, 16, 20, 32]	20, 20
Conv. Layers	[2, 3, 4]	2
Kernel Size	[(2,2),(2,2)], [(3,3),(3,3)], [(4,4),(4,4)]	(2,2), (2,2)
No. of Kernels	[(32,64),(64,128),(128,256), (32,32),(64,64),(128,128)]	(32,32), (64,64)
Dropout	[0.0, 0.1, 0.2, 0.3, 0.4, 0.5]	0.2, 0.2
Bayesian Optimizer	SGD, RMSprop, Adagrad, Adadelta, Adam	Adam, Adam

3.4 Hierarchical Classification

In traditional classification, also referred to as flat classification, the model predicts the correct label for the input data without considering any hierarchical structure among the classes. On the other hand, hierarchical classification involves organizing the classes into a hierarchical structure where each class at a certain level represents a subset of classes at a higher level, and the classification process occurs in a hierarchical manner, typically through a series of binary classification tasks at each level of the hierarchy [7]. At each level, the model distinguishes between one class and all other classes grouped into a single class. If there are n classes in total, there will be n-1 levels of hierarchy, as each level involves a binary classification task.

The dataset consists of 4 classes: CN, BV, PNFA, and SV, variants of FTD. Here, we can divide the hierarchy into three levels, as shown in Fig. 6. The first level of the hierarchy is where we group the BV, SV, and PNFA into a common class called FTD. Since all three are variants of FTD, we group them into this FTD class in Level-1. So now the 4-class classification has been boiled down to a 2-class classification, i.e., CN vs FTD. We will build a classifier for Level-1, which correctly classifies between CN and FTD. In Level-2, we can drop all

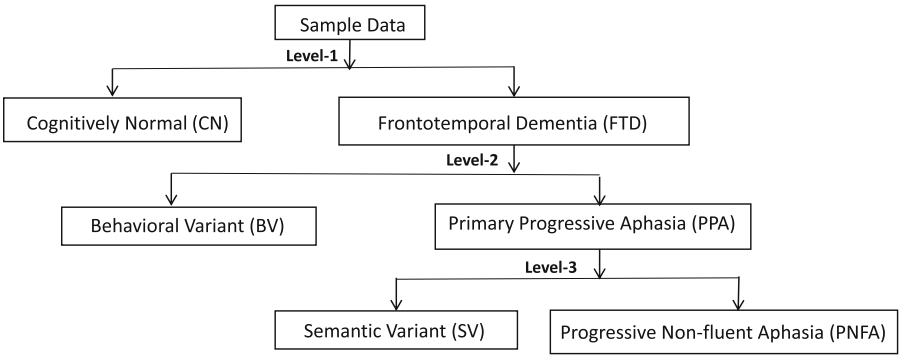


Fig. 6. Hierarchical Classification Approach.

the data of the class CN because the Level-2 classifier only classifies among the FTD Variants. Now, we have 3 classes: BV, SV, and PNFA. Now, we can divide these into two classes, BV and Primary Progressive Aphasia (PPA), i.e., SV and PNFA, which are grouped into PPA since they are the types of PPA. The Level-2 Classifier will be trained over the data with BV and PPA class labels. So, Level-2 will classify over BV and PPA. Now, we drop all the data samples with class BV. Level-3 Classifier trained over the SV and PNFA data samples. It will do the classification SV and PNFA. We have broken down the 4-class flat classification into 3 levels of hierarchical classification. The overall accuracy of the hierarchical classification is defined as (total accuracy = $(\Sigma_{i=1}^{N}\ Accuracy_i)$/N), N is the total number of levels in the hierarchy and $Accuracy_i$ is the accuracy at level i.

4 Results and Discussion

Comparison with existing methods: Table 2 shows the classification performance of both approaches (flat and hierarchical) by employing ML/DL models and comparison with existing methods. All ML models (Random Forest (RF), Extreme Gradient Boosting (XGBoost), Support Vector Machine (SVM), and MLP) were trained on tabular data, while the DL models (CNNs [9], ResNet-50 [10], and VGG-16 [11]) were trained using images generated through algorithms (IGTD and TIG). We also conducted the ensemble CNN+MLP experiment in which we directly concatenated the raw data to the features extracted by CNN and performed classification by training CNN only. A wrapper function was created to efficiently obtain the metric accuracy to evaluate the trained models. The robustness of the models was estimated by the k-fold cross-validation. It can be observed that XGBoost performs better in flat and hierarchical classification approaches, and the DL method's performance in terms of test accuracy over the image data generated from the IGTD representation is improved for hierarchical classification. We can state that the hierarchical approach indicates improvement compared to flat classification across all models except the ensemble of CNN and MLP for IGTD image representation. Table 2 highlights the DL

method's performance over TIG and IGTD images, and the results show high accuracy in detecting FTD subtypes using a CNN ensemble with MLP, achieving a test accuracy of 88.89%, representing an around 5% increase over the flat best-known methods [18] [17] [16].

Table 2. Comparison with existing methods and performance of machine learning models over the tabular data and deep learning method on the images generated from IGTD and TIG algorithms.

	Methods	Flat Accuracy (%)	Hierarchical Accuracy (%)
Existing Methods (MRI Image Data)	Torso et.al (2020) KNN [17]	75.75	-
	Ma et. al (2024) DNN[18]	84.00	-
	Rogeau et. al (2024) 3D-CNN [16]	84.10	-
Tabular Data	Random Forest	82.75	85.23
	Extreme Gradient Boosting	83.90	86.51
	Support Vector Machines	81.89	86.08
	Multi-layer Perceptron	81.03	85.21
IGTD Image	CNN	82.41	82.96
	ResNet-50	66.18	78.61
	VGG-16	68.16	76.74
	CNN+MLP	82.06	78.50
TIG Image (Proposed Method)	CNN	84.48	84.19
	ResNet-50	72.64	83.06
	VGG-16	71.04	76.54
	CNN+MLP	**85.76**	**88.04**
IGTD & TIG Images	**CNN+MLP**	**86.00**	**88.89**

Classification Performance: Table 3 presents performance evaluation metrics for classification models, comparing with the baseline and proposed approaches. It provides precision, recall, F1-score, and accuracy values for each class (CV, BV, PNFA, and SV) under both flat and hierarchical approaches. Insights from the table include: the proposed hierarchical approach generally outperforms the flat approach across most metrics and models. RF and XGBoost models show improved performance in the hierarchical approach compared to the flat approach. After careful evaluation of different models, the CNN+MLP ensemble model achieves competitive results, with slightly lower precision but higher recall and F1-score compared to other models. Overall, the proposed hierarchical approach, particularly employing the CNN+MLP ensemble model, outperforms in accurately classifying FTD subtypes. Despite its more complex architecture, the optimized CNN+MLP ensemble model consistently produced results without showing any noticeable bias towards any particular class.

Key Findings: Transforming tabular data into images using converters encodes feature distance information into a 2D spatial configuration. This visual representation enhances the CNN's ability to capture underlying patterns in the data enriched with spatial information effectively. In cases where the dataset exhibits complexities that hinder effective classification, employing CNNs can be a viable solution to enhance performance. We presented the new algorithm TIG, which can better capture the spatial 2D correlation of the features than the existing techniques IGTD.

Combining CNN-extracted features from TIG-generated images with raw tabular data creates a more comprehensive feature set that captures both spatial correlations and direct numerical relationships, leading to improved classification performance. The ensemble method, which combines the strengths of CNNs and MLPs, demonstrates better performance than other methods. This suggests that hybrid models can effectively harness the complementary strengths of different algorithms to achieve better results.

The integration of features from two different modalities proved effective in capturing diverse features and enhancing FTD subtype classification accuracy, particularly valuable in addressing the complex and heterogeneous nature of FTD. Leveraging the natural hierarchy of FTD subtypes, our proposed model offers clinicians a more comprehensive understanding of classification, potentially improving diagnostic accuracy and patient care.

Table 3. Performance evaluation metrics of classification models for baseline and proposed approach. CN: patients with normal cognitive function; BV: patients primarily affecting behavior and personality; SV: patients characterized by language and semantic memory impairments; PNFA: patients primarily impacting language fluency and grammatical processing.

	Model	Approach	Precision				Recall				F1-Score				Accuracy
			CN	BV	PNFA	SV	CN	BV	PNFA	SV	CN	BV	PNFA	SV	
Baseline	Random Forest	Flat	0.91	0.73	0.95	0.82	0.94	0.76	0.46	0.75	0.92	0.74	0.62	0.78	0.82
		Hierarchical	0.93	0.77	0.86	0.87	0.96	0.76	0.78	0.84	0.94	0.77	0.81	0.85	0.85
	XGB	Flat	0.92	0.80	0.77	0.73	0.95	0.71	0.58	0.87	0.93	0.75	0.66	0.79	0.83
		Hierarchical	0.94	**0.81**	0.84	0.91	0.96	0.76	0.83	**0.87**	0.95	**0.78**	0.83	**0.89**	0.86
Proposed Method	CNN+MLP	Flat	**0.95**	0.78	0.73	0.78	0.95	0.76	0.62	0.62	0.95	**0.78**	0.67	0.69	0.85
		Hierarchical	**0.95**	**0.78**	**0.95**	**0.92**	**0.97**	**0.77**	**0.87**	0.83	**0.96**	**0.78**	**0.91**	0.87	**0.88**

5 Conclusion

The results suggest that implementing a hierarchical classification strategy has resulted in enhanced accuracy. Additionally, incorporating TIG and combining CNN with MLP have contributed to further improvements in accuracy compared to conventional machine learning models on the FTLDNI dataset. The proposed ensemble approach results in better classification performance and robustness. This study provides a detailed understanding of the heterogeneity within the

FTD spectrum. By categorizing FTD subtypes hierarchically, clinicians can better tailor diagnostic and treatment strategies to match each subtype's specific characteristics and progression patterns. The tabular-to-image conversion approach can be adapted to various domains and applications, providing new insights and opportunities for data analysis and modeling.
The limitation of this work is that we primarily focused on exploring the effectiveness of 2D CNNs for tabular data by converting it into images. While 1D CNNs can be directly applied to raw tabular data, we chose to emphasize 2D CNNs due to the potential advantages of leveraging spatial information. We did not extensively investigate the potential of 1D CNNs in this context.
In the future, we aim to delve deeper into optimizing CNN architectures specifically for image representations derived from tabular data. This includes investigating alternative methods for converting tabular data into images and exploring the potential of 1D CNNs. Extending this methodology to other health datasets and analyzing model performance will also be a key focus. Additionally, we plan to conduct a computational cost analysis to compare the efficiency of CNNs against traditional ML models in terms of resource utilization.

Acknowledgement. The dataset utilized in this research was sourced from the Frontotemporal Lobar Degeneration Neuroimaging Initiative (FTLDNI). We express our gratitude to the data access committee for granting approval to access the FTLDNI dataset.

References

1. Sawyer, Russell & Shinagawa, Shunichiro & Miller, Bruce. (2020). Frontotemporal dementia. https://doi.org/10.1016/B978-0-12-813866-3.00003-5
2. Bang, J., Spina, S., Miller, B.L.: Non-Alzheimer's dementia 1: frontotemporal dementia. Lancet (London Engl.) **386**(10004), 1672 (2015)
3. Antonioni, A., Raho, E.M., Lopriore, P., Pace, A.P., Latino, R.R., Assogna, M., Mancuso, M., Gragnaniello, D., Granieri, E., Pugliatti, M., Di Lorenzo, F., Koch, G.: Frontotemporal Dementia, Where Do We Stand? A Narrative Review. Int. J. Mol. Sci. **24**(14), 11732 (2023). https://doi.org/10.3390/ijms241411732
4. D. Dai, H. He, J. Vogelstein and Z. Hou, Network-based classification using the cortical thickness of AD patients, in: International Workshop on Machine Learning in Medical Imaging, Springer, 2011, pp. 193–200
5. M. Bruun, J. Koikkalainen, H.F. Rhodius-Meester, M. Baroni, L. Gjerum, M. van Gils, H. Soininen, A.M. Remes, P. Hartikainen, G. Walde- mar et al., Detecting frontotemporal dementia syndromes using MRI biomarkers, NeuroImage: Clinical 22 (2019), 101711
6. Torso, M., Bozzali, M., Cercignani, M., Jenkinson, M., Chance, S.: Using diffusion tensor imaging to detect cortical changes in frontotemporal dementia subtypes. Sci. Rep. **10**(1), 1–11 (2020)
7. Jun Pyo Kim, Jeonghun Kim, Yu Hyun Park, Seong Beom Park, Jin San Lee, Sole Yoo, Eun-Joo Kim, Hee Jin Kim, Duk L Na, Jesse A Brown, et al. Machine learning based hierarchical classification of frontotemporal dementia and Alzheimer's disease. NeuroImage: Clinical, 23:101811, 2019

8. Yitan Zhu, Thomas Brettin, Fangfang Xia, Alexander Partin, Maulik Shukla, Hyunseung Yoo, Yvonne A Evrard, James H Doroshow, and Rick L Stevens. Converting tabular data into images for deep learning with convolutional neural networks. Scientific reports, 11(1):11325, 2021
9. Nash, R. (2015). An Introduction to Convolutional Neural Networks. ArXiv. /abs/1511.08458
10. He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. ArXiv. /abs/1512.03385
11. Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. ArXiv. /abs/1409.1556
12. Hu, J., Qing, Z., Liu, R., Zhang, X., Lv, P., Wang, M., Wang, Y., He, K., Gao, Y., Zhang, B.: Deep learning-based classification and voxel-based visualization of frontotemporal dementia and Alzheimer's disease. Front. Neurosci. **14**, 1468 (2021)
13. Kandel, I., Castelli, M.: The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset. ICT express **6**(4), 312–315 (2020)
14. Wan, S., Liang, Y., Zhang, Y., Guizani, M.: Deep multi-layer perceptron classifier for behavior analysis to estimate parkinson's disease severity using smartphones. IEEE Access **6**, 36825–36833 (2018). https://doi.org/10.1109/ACCESS.2018.2851382
15. Poonam, K., Guha, R., Chakrabarti, P.P. (2023). Frontotemporal Dementia Detection Model Based on Explainable Machine Learning Approach. In: Chandran K R, S., N, S., A, B., Hamead H, S. (eds) Computational Intelligence in Data Science. ICCIDS 2023. IFIP Advances in Information and Communication Technology, vol 673. Springer, Cham. https://doi.org/10.1007/978-3-031-38296-3_18
16. Rogeau, A., Hives, F., Bordier, C., Lahousse, H., Roca, V., Lebouvier, T., Pasquier, F., Huglo, D., Semah, F., Lopes, R.: A 3D convolutional neural network to classify subjects as Alzheimer's disease, frontotemporal dementia or healthy controls using brain 18F-FDG PET. Neuroimage **288**, 120530 (2024). https://doi.org/10.1016/j.neuroimage.2024.120530
17. Torso, M., Bozzali, M., Cercignani, M., Jenkinson, M., Chance, S.: Using diffusion tensor imaging to detect cortical changes in frontotemporal dementia subtypes. Sci. Rep. **10**(1), 1–11 (2020)
18. Ma, D., Stocks, J., Rosen, H., Kantarci, K., Lockhart, S.N., Bateman, J.R., Craft, S., Gurcan, M.N., Popuri, K., Beg, M.F., Wang, L.: Differential diagnosis of frontotemporal dementia subtypes with explainable deep learning on structural MRI. Front. Neurosci. **18**, 1331677 (2024). https://doi.org/10.3389/fnins.2024.1331677
19. Sharma, A., Vans, E., Shigemizu, D., Boroevich, K.A., Tsunoda, T.: DeepInsight: A methodology to transform a non-image data to an image for convolution neural network architecture. Sci. Rep. **9**(1), 1–7 (2019)
20. Bazgir O, Zhang R, Dhruba SR, Rahman R, Ghosh S, Pal R. Representation of features as images with neighborhood dependencies for compatibility with convolutional neural networks. Nature communications.2020; 11(1):4391. https://doi.org/10.1038/s41467-020-18197-y PMID: 32873806

Identifying Alzheimer's Disease Across Cognitive Impairment Spectrum Using Acoustic Features Only

Hyo Jin Jon, Hyuntaek Jung, Longbin Jin, and Eun Yi Kim(✉)

AI&Computer Vision Lab, Konkuk University, Seoul, Korea
{hyojin2011,busan199,jinlongbin,eykim}@konkuk.ac.kr

Abstract. Early diagnosis of Alzheimer's Disease (AD) holds key in delaying cognitive decline and managing the progression of dementia. The Mini-Mental State Examination (MMSE) serves as a valuable tool for evaluating cognitive impairment, aiding in the identification of AD severity and early-stage diagnosis. Traditionally, prediction models aiming to predict MMSE scores from speech have utilized the mean squared error (MSE) as the loss function. However, MSE tends to bias predictions towards the average, which limits its ability to discriminate AD severity effectively. To address this limitation, we propose a novel Weighted MSE-CE loss function aimed at improving AD discrimination by accurately predicting MMSE scores. It reaches the goal through the integration of cross entropy (CE) into mean squared error (MSE), leveraging the Bernoulli penalty and distance-based weights. Furthermore, our approach relies solely on language-agnostic acoustic features, eliminating the requirement for transcription. Consequently, it gains the potential to be applied to individuals speaking various languages. Experimental results on the ADReSSo dataset demonstrate that our method reduces the RMSE to 4.55, outperforming other acoustic-based approaches. Additionally, our predictions achieved an impressive 77.46% accuracy in AD detection, highlighting its effectiveness in assessing AD severity.

Keywords: Alzheimer's Disease · MMSE score prediction · AD Detection · weighted MSE-CE Loss · Acoustic Features.

1 Introduction

Dementia is a widespread cognitive disorder affecting millions of people worldwide, including Alzheimer's Disease (AD). Early detection of AD is crucial for timely intervention and treatment aimed at delaying cognitive decline and mitigating the risk of progression to dementia. However, diagnosing AD in its early stages remains challenging, often due to symptoms being mistaken for normal aging by both patients and medical professionals [4].

The Mini-Mental State Examination (MMSE) [8] is a survey for screening cognitive impairment and dementia, including AD, which is commonly used

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15311, pp. 402–413, 2025.
https://doi.org/10.1007/978-3-031-78195-7_27

worldwide. It contains a brief questionnaire for assessing various cognitive functions. With scores ranging from 0 to 30, higher scores indicate better cognitive ability. The identification between Healthy Control (HC) and probable AD is commonly made based on a criterion score around 27 [15,30]. Notably, the assessment of the MMSE allows for not only the detection of severity but also the diagnosis of early-stage AD.

Recent research has been dedicated to identifying AD patients based on their speech patterns, as individuals with AD commonly experience language impairments like aphasia [10,26]. Thus, speech emerges as a crucial clue for facilitating easier AD diagnosis and cognitive status monitoring. The ADReSSo [19] is one of the most representative AD datasets, which targets AD recognition in spontaneous speech. Within the dataset, existing studies have primarily categorized subjects as either HC or probable AD, with less attention on assessing the severity of the disease through MMSE scores. However, identifying severity is crucial for early detection and understanding AD progression, highlighting the critical need for accurately predicting MMSE scores. In predicting MMSE scores, the predictions typically rely on acoustic and linguistic features of speech. These features are derived from speech recordings and corresponding transcripts. However, due to the absence of manual transcripts in ADReSSo, unlike in the prior ADReSS dataset [18], efforts were made to generate transcripts before linguistic feature extraction. This adjustment was necessary because linguistic features have shown better performance compared to acoustic features in the ADReSS dataset [5,7,14,24,31].

Despite the superior performance of linguistic features, they are limited by their dependency on Automatic Speech Recognition (ASR) and lack of scalability for diverse languages. Within the ADReSSo dataset, studies [19,23] have demonstrated better results in MMSE score prediction by utilizing ASR-driven linguistic features. Nevertheless, the usage of linguistic features caused the reliance on ASR performance [22], and changing the language may significantly impair performance. Consequently, our focus is directed toward utilizing exclusively acoustic features.

In this paper, we introduce a novel weighted MSE-CE loss to facilitate AD severity discrimination by accurately predicting MMSE scores. The traditional loss function, mean squared error (MSE), lacks discrimination power for AD due to its tendency for predicted MMSE scores to converge towards the average score of the dataset. To prevent this convergence, we integrate cross entropy (CE) into MSE. By segmenting the inference process into two phases, we consider both classification for subdivided score intervals and its subsequent regression for expected scores. Within the process, we penalize the ambiguous probabilities during score inference to prevent discrepancies between the predicted probability and expected score. Also, we consider the distance property between intervals in the classification phase since traditional CE disregards the relationship between labels. Furthermore, unlike conventional approaches relying on linguistic features extracted from ASR-driven texts, our approach exclusively leverages acoustic features from speech recordings. This distinction provides our method

with the potential to be applicable to users across diverse linguistic backgrounds. Experimental results on the ADReSSo dataset demonstrate the lowest RMSE of 4.55 in MMSE score prediction among acoustic-based approaches. Subsequently, the predictions achieve an impressive 77.46% accuracy in AD detection through simple thresholding. These results underscore the effectiveness of our loss function for AD discrimination as well as MMSE score prediction. In summary, the contributions of this paper are as follows:

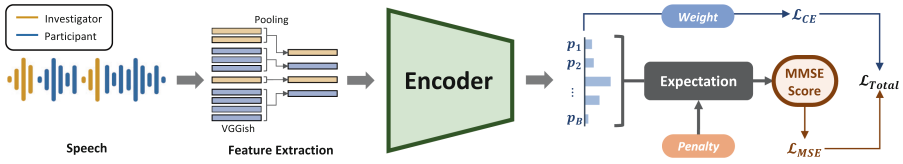


Fig. 1. Overview of proposed method for MMSE score prediction.

- We propose the weighted MSE-CE loss, designed to facilitate AD severity discrimination by accurately predicting MMSE scores.
- By utilizing exclusively acoustic features, we eliminate the need for transcription, thereby reducing costs and promising language-agnostic models.
- Experimental results demonstrate significant AD discrimination power, achieved by superior MMSE score prediction performance among acoustic-based approaches.

2 Method

In this paper, our goal is to identify the spectrum of AD severity through accurate MMSE score prediction. To achieve this, we propose weighted MSE-CE loss, which addresses the limitation of traditional MSE loss. Before delving into details, we offer a description of the overall model architecture.

2.1 Model Architecture

A convolutional recurrent neural network (CRNN)-based model, which was developed by Koo et al. [14], is employed to leverage sequential information from speech. The overview of the proposed MMSE score prediction model is depicted in Fig. 1.

Initially, VGGish [11] features are extracted from the Mel-Spectrogram of the speech recording. These features, each comprising 128 dimensions, are extracted every 960ms to capture the acoustic context. Subsequently, the features are segmented along each speaker turn, and global max pooling is applied to capture the most salient features.

The data then flows through the CRNN encoder, which is designed to capture information within and across sequences. This encoder includes an attention layer, squeeze and excitation blocks [12], a convolutional neural network, bidirectional long short-term memory layers, and fully connected (FC) layers. This architecture enables the model to learn complex patterns within the input features, capturing temporal dependencies and focusing on relevant features while enhancing representation. Additionally, the speaker's ID information is incorporated by channel-wise concatenation with VGGish features post-attention layer, utilizing an FC layer. This integration ensures the model can leverage speaker-specific characteristics to improve its predictions.

Finally, the score inference process is segmented into two sequential phases to exploit the novel weighted MSE-CE loss. The encoder generates the probabilities for subdivided score intervals. Then, the MMSE score is predicted using an expected formula. Further details will be presented in the next section.

2.2 Weighted MSE-CE Loss

To achieve our goal, we introduce a novel loss function designed to overcome the main drawback of the traditional MSE loss. While MSE is commonly used in MMSE score prediction due to its simplicity and intuitiveness [5,14,22,24,25], it tends to converge predictions towards the dataset's average value, limiting their discrimination power for AD. To prevent this convergence, we draw inspiration from the computer vision domain [1] and integrate the CE loss into the MSE loss. In this paper, we initially partition the entire MMSE score distribution of 0 to 30 into several regularly sized segments. Subsequently, the score prediction process unfolds into two phases: classification for the score segment and regression for the score within the class. The classification phase aims to discern the spectrum of AD severity, ranging from early to severe cognitive impairments, while the regression phase focuses on predicting exact MMSE scores. The score is inferred using the expectation formula depicted in Equation 2. Here, $p \in \{p_i\}_{i=1}^{B}$ and s represent the forecasted probabilities and score, respectively, with i indicating the index for the intervals subdivided into B bins. Additionally, y denotes the true MMSE score, while its corresponding one-hot encoded vector is denoted by $\hat{y} \in \{\hat{y}_i\}_{i=1}^{B}$.

$$\mathcal{L}_{MSE-CE} = MSE(y, s) + CE(\hat{y}, p) \tag{1}$$

$$s = E(p) = \frac{30}{B} \sum_{i=1}^{B} p_i \tag{2}$$

$$MSE(y, s) = (y - s)^2 \tag{3}$$

$$CE(\hat{y}, p) = -\sum_{i=1}^{B} \hat{y}_i \cdot \log p_i \tag{4}$$

Bernoulli Penalty. Despite the incorporation of an additional classification phase, the model encounters challenges in acquiring explicit probabilities for each interval, which hinders the consistency between the two phases. In other words, a discrepancy occurs between the predicted probability and expected score. To address this discrepancy caused by the uncertainty, we introduce a penalty for ambiguous probabilities during the score inference. The penalty leverages the variance of the Bernoulli distribution, which exhibits the highest value when the probability reaches 0.5. By applying this penalty, the model is encouraged to clarify its probability predictions to obtain the expected MMSE score. The penalized expectation formula is outlined in Equation 5, where λ denotes a penalty parameter.

$$s^* = E(p) - \lambda \cdot \frac{30}{B} \sum_{i=1}^{B} (1 - p_i) p_i \tag{5}$$

Weighted Cross Entropy. Unlike common classification labels, the subdivided score intervals possess a distance property between them. However, traditional CE overlooks this property since it only considers the probability of the correct interval. Recognizing this distinction, we propose weighted CE in the classification phase. By employing absolute distance as a weight, our proposed loss function reflects the interval misalignments. Additionally, we incorporate an offset of 1 to prevent the loss from reaching 0 when the correct interval has the highest probability. Equation 6 and 7 provide detailed formulations of the absolute distance and weighted CE, with d representing the distance between the correct and predicted intervals.

$$d = \left| \underset{i}{argmax}(\hat{y}_i) - \underset{i}{argmax}(p_i) \right| \tag{6}$$

$$CE_{Weighted}(\hat{y}, p) = -(1 + d) \cdot \sum_{i=1}^{B} \hat{y}_i \cdot \log p_i \tag{7}$$

Total Loss Function. The proposed loss function combines a penalized MSE with weighted CE to accurately predict MMSE scores and discern the spectrum of AD severity. Each loss strengthens the synergy between them complementarily, contributing to the overall effectiveness of the model. The total loss function is depicted in Equation 8.

$$\mathcal{L}_{Total} = MSE(y, s^*) + CE_{Weighted}(\hat{y}, p) \tag{8}$$

3 Experimental Results

In this paper, we propose a novel loss function to facilitate AD severity discrimination by accurate MMSE score prediction. To validate our proposal, we

present experimental results conducted on the ADReSSo dataset [19]. We report RMSE value for MMSE score prediction, while also considering the corresponding accuracy for AD detection. Furthermore, we compare our method with existing approaches that utilize various features, including linguistic properties.

3.1 Experimental Settings

ADReSSo Dataset. The ADReSSo dataset [19] was collected for dementia recognition from spontaneous speech. It provides speech recordings with labels for AD and MMSE scores. Participants were tasked with describing a Cookie Theft picture from the Boston Diagnostic Aphasia Examination [9]. The recording includes participant speech and investigator commentary, segmented by speaker turns. The dataset, designed to be unbiased in gender and age, varies in length from 23s to 279s. Notably, Table 1 illustrates differences between probable AD and HC in MMSE scores, suggesting the potential for AD identification through accurate MMSE score prediction. The training set comprises 166 samples, with 71 samples in the test set. The evaluation is conducted on the test set.

Table 1. Statisitic of MMSE scores on the ADReSSo dataset.

AD Label	Train (Mean / Std.)	Test (Mean / Std.)
Probable AD	17.44 / 5.33	18.86 / 5.80
HC	28.99 / 1.15	28.91 / 1.25
Total	22.93 / 6.99	23.89 / 6.56

Implementation Details. The experiments utilize an identical model architecture. In terms of the baseline model using MSE, the single value obtained through Sigmoid is adjusted to fit within a range of 0 to 30. Conversely, our proposal segments the scores into 10 interval bins and sets the penalty parameter λ to 1. Training exclusively employs MMSE score labels. In order to evaluate AD detection as well, we establish a threshold for the predicted MMSE scores at 27, as suggested by Kukull et al. [15]. This threshold-based classification using AD labels facilitates the validation of our model's ability to discriminate AD based on the predicted MMSE scores. Each model contains about 33 million parameters, requiring approximately 3 hours of training time on a single NVIDIA RTX 3090 GPU.

3.2 Main Results

Our aim is to demonstrate the effectiveness of proposed weighted MSE-CE loss in facilitating AD discrimination and predicting MMSE scores. We begin by

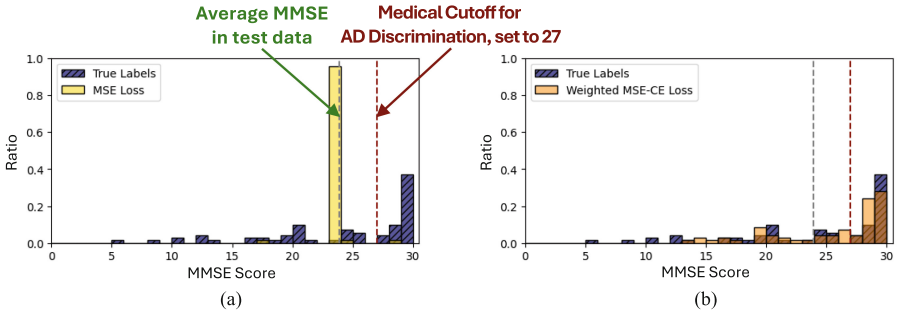


Fig. 2. Distribution of true and predicted MMSE scores along the loss function: (a) when using only MSE, predictions converge to the average MMSE score in the test data, failing to discriminate between AD and normal. (b) Weighted MSE-CE loss enables correct MMSE score prediction and AD discrimination.

Table 2. Main results of MMSE score prediction and AD detection on the ADReSSo datasets.

Task	Method	Metrics	
		RMSE	ACC (%)
AD detection	$A1$. CE Loss	-	77.46
MMSE score prediction	$M1$. MSE Loss	6.60	49.30
	$M2$. MSE-CE Loss	5.49	66.20
	$M3$. + Bernoulli Penalty	4.79	70.42
	$M4$. + Absolute Distance	**4.55**	**77.46**

establishing baseline models for AD detection $A1$ and MMSE score prediction $M1$ using traditional loss functions. Subsequently, we progressively integrate components into our approach: incorporating CE into MSE $M2$, a Bernoulli penalty within the scoring process $M3$, and weighted CE in the classification phase $M4$.

Quantitative Results. Table 2 illustrates the quantitative improvements achieved by each component. Initially, $M1$ shows poor performance, particularly in AD detection, indicating the limitation of traditional MSE loss in distinguishing AD. However, significant enhancements are observed in our proposals. $M2$ demonstrates the most substantial improvements in both tasks, indicating that integrating the classification phase significantly refines predictions to be more accurate and discriminative. Subsequently, $M3$ and $M4$ primarily enhance each task, respectively. $M3$ mainly contributes to MMSE score prediction, suggesting that mitigating discrepancies between predicted probabilities and expected scores leads to more accurate predictions. Conversely, $M4$ primarily enhances AD detection, indicating that considering the relationship between interval labels

enhances discrimination power. Additionally, consistent enhancements on both tasks indicate that as the accuracy of MMSE score prediction increases, the discrimination power for AD improves. Consequently, achieving an accuracy equivalent to A1 and the lowest RMSE of 4.55 underscores the effectiveness of accurate MMSE score prediction in AD discrimination.

Qualitative Results. Visualizing the prediction distributions for $M1$ and $M4$ confirms the effectiveness of our proposed loss function. In Fig. 2(a), the traditional MSE loss leads to predictions converging towards the average score of the dataset. Since the scores of probable AD significantly reduce the average score, most predictions fall below the discrimination threshold, making them unable to discern AD. Conversely, in Fig. 2(b), our proposed loss function aligns predictions closely with the true MMSE scores, enabling effective identification of AD severity. This visual representation highlights the superior precision of our approach in MMSE score prediction.

Table 3. Performance comparison of MMSE score prediction among various methods.

Feature Type	Method	RMSE
Linguistic	Agbavor and Liang [2]	5.46
	Luz et al. [19]	5.28
	Pérez-Toro et al. [23]	4.56
	Rohanian et al. [25]	4.49
	Mirheidari et al. [21]	4.45
	Zhu et al. [33]	4.44
	Pappagari et al. [22]	3.85
Linguistic + Acoustic	Shah and Aryal [27]	5.56
	Pérez-Toro et al. [23]	4.79
	Pappagari et al. [22]	4.62
	Rohanian et al. [25]	4.26
Acoustic	Agbavor and Liang [2]	6.25
	Luz et al. [19]	6.09
	Pérez-Toro et al. [23]	5.35
	Proposed Method	**4.55**

3.3 Comparison

We provide a comprehensive comparison with existing methods for MMSE score prediction using the ADReSSo dataset, as shown in Table 3. So far, most studies

have favored linguistic features over acoustic ones due to their superior performance [2,19,23]. Consequently, Mirheidari et al. [21] and Zhu et al. [33] exclusively focused on using ASR to predict MMSE scores. Notably, Pappagari et al. [22] achieved outstanding results with an RMSE of 3.85 using the BERT model with ASR, capturing speech disfluency from language-specific ASR-driven transcripts. Additionally, Rohanian et al. [25] and Shah and Aryal [27] combined acoustic features with linguistic ones to enhance performance. However, the usage of linguistic features requires the transcription via ASR, confining their applicability within a single language. Recently, there has been a growing interest in acoustic features due to their immediate extraction from recordings and language-agnostic nature. Nevertheless, the potential of acoustic features in MMSE score prediction remains underexplored. In acoustic-based approaches, Pérez-Toro et al. [23] achieved the best RMSE of 5.35, employing prosody features, x-vector, and VAD valence and dominance for linear regression.

In this paper, we emphasize the exclusive utilization of acoustic features for MMSE score prediction, aiming to maximize their potential. Our proposed method achieves a state-of-the-art RMSE of 4.55 among models using only acoustic features. It represents a significant improvement of approximately 15% (RMSE of 0.80) compared to the existing best performance. Furthermore, our competitive performance against models involving linguistic features underscores the comparable effectiveness of acoustic features.

4 Discussion

The ADReSS challenge [18] has played a crucial role in advancing machine learning techniques for identifying AD and predicting MMSE scores through spontaneous speech. Participants [5,7,14,24,31] in the challenge predominantly directed their attention towards linguistic features due to the availability of both audio recordings and manual transcripts in the dataset. However, the reliance on manual transcripts presents significant challenges in terms of cost and time. Subsequently, the ADReSSo challenge [19] shifted towards the dataset containing only audio recordings, leading to increased reliance on ASR for transcription [22,23,25]. However, this approach presents limitations, as the diagnostic model's performance is contingent upon the accuracy of the ASR system, thereby hindering scalability across various languages and speech patterns. Hence, the utilization of acoustic features became mainstream in the latest ADReSS-M [17] challenge due to its inclusion of diverse languages across both the training and test sets [13,20,28,32].

In response to this evolving trend, our approach emphasizes the utilization of acoustic features exclusively, eliminating the need for transcription entirely. Thanks to this approach, we gain the potential to be applicable to users across diverse languages. Moreover, this simplifies the feature extraction process and eliminates the reliance on ASR performance, as transcription is not required. Our proposed weighted MSE-CE loss achieves state-of-the-art performance in MMSE score prediction among acoustic-based approaches, and the notably high

accuracy in AD detection underscores the competitiveness of acoustic features. These results represent a significant advance towards practical diagnostics with global applicability. Moving forward, our approach requires validation in a multilingual environment with various acoustic features to further demonstrate its potential. Therefore, we plan to enhance the robustness of our findings by incorporating a broader array of acoustic features, such as Wav2Vec 2.0 [3], X-vector [29], and I-vector [6]. Additionally, we aim to extend our approach to multilingual datasets, including ADReSS-M and TAUKADIAL [16], in future work.

5 Conclusion

In this paper, we introduce a novel Weighted MSE-CE loss function designed to enhance both MMSE score prediction and AD discrimination. We address the limitation of traditional methods by preventing predictions from converging solely to the average MMSE score in the dataset. This is achieved through the integration of cross entropy (CE) into mean squared error (MSE), leveraging the Bernoulli penalty and distance-based weights. Experimental results on the ADReSSo dataset demonstrate the lowest RMSE of 4.55 in MMSE score prediction among acoustic-based approaches. Moreover, our predictions achieve an impressive 77.46% accuracy in AD detection. Consequently, our novel Weighted MSE-CE loss function, in combination with acoustic feature-based modeling, represents a significant advancement in MMSE score prediction and AD severity discrimination. This holds promise for enhancing early diagnosis and management strategies in AD.

Acknowledgements. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) under the metaverse support program to nurture the best talents (IITP-2024-RS-2023-00256615) grant funded by the Korea government(MSIT).

References

1. Abdelrahman, A.A., Hempel, T., Khalifa, A., Al-Hamadi, A., Dinges, L.: L2CS-Net : Fine-Grained Gaze Estimation in Unconstrained Environments. In: 2023 8th International Conference on Frontiers of Signal Processing (ICFSP). pp. 98–102 (2023)
2. Agbavor, F., Liang, H.: Predicting dementia from spontaneous speech using large language models. PLOS Digital Health **1**(12), 1–14 (12 2022). https://doi.org/10.1371/journal.pdig.0000168, https://doi.org/10.1371/journal.pdig.0000168
3. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 12449–12460. Curran Associates, Inc. (2020)
4. Belmin, J., Min, L., Roth, C., Reuben, D., Wenger, N.: Assessment and management of patients with cognitive impairment and dementia in primary care. J. Nutr. Health Aging **16**(5), 462–467 (2012)

5. Cummins, N., Pan, Y., Ren, Z., Fritsch, J., Nallanthighal, V.S., Christensen, H., Blackburn, D., Schuller, B.W., Magimai-Doss, M., Strik, H., Härmä, A.: A Comparison of Acoustic and Linguistics Methodologies for Alzheimer's Dementia Recognition. In: Proc. Interspeech 2020. pp. 2182–2186 (2020)
6. Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P., Ouellet, P.: Front-end factor analysis for speaker verification. IEEE Trans. Audio Speech Lang. Process. **19**(4), 788–798 (2011). https://doi.org/10.1109/TASL.2010.2064307
7. Farzana, S., Parde, N.: Exploring MMSE Score Prediction Using Verbal and Non-Verbal Cues. In: Proc. Interspeech 2020. pp. 2207–2211 (2020)
8. Folstein, M.F., Folstein, S.E., McHugh, P.R.: "mini-mental state": A practical method for grading the cognitive state of patients for the clinician. J. Psychiatr. Res. **12**(3), 189–198 (1975)
9. Goodglass, H., Kaplan, E.: Boston diagnostic aphasia examination booklet. Lea & Febiger (1983)
10. Harnish, S.M.: Anomia and Anomic Aphasia: Implications for Lexical Processing. In: The Oxford Handbook of Aphasia and Language Disorders. Oxford University Press (2018)
11. Hershey, S., Chaudhuri, S., Ellis, D.P.W., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., Slaney, M., Weiss, R.J., Wilson, K.: Cnn architectures for large-scale audio classification. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 131–135 (2017)
12. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7132–7141 (2018)
13. Jin, L., Oh, Y., Kim, H., Jung, H., Jon, H.J., Shin, J.E., Kim, E.Y.: Consen: Complementary and simultaneous ensemble for alzheimer's disease detection and mmse score prediction. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–2 (2023)
14. Koo, J., Lee, J.H., Pyo, J., Jo, Y., Lee, K.: Exploiting Multi-Modal Features from Pre-Trained Networks for Alzheimer's Dementia Recognition. In: Proc. Interspeech 2020. pp. 2217–2221 (2020)
15. Kukull, W., Larson, E., Teri, L., Bowen, J., McCormick, W., Pfanschmidt, M.: The mini-mental state examination score and the clinical diagnosis of dementia. J. Clin. Epidemiol. **47**(9), 1061–1067 (1994)
16. Luz, S., Garcia, S.d.l.F., Haider, F., Fromm, D., MacWhinney, B., Lanzi, A., Chang, Y.N., Chou, C.J., Liu, Y.C.: Connected speech-based cognitive assessment in chinese and english (2024). https://doi.org/10.48550/ARXIV.2404.nnnnn, final DOI to be assigned
17. Luz, S., Haider, F., Fromm, D., Lazarou, I., Kompatsiaris, I., MacWhinney, B.: Multilingual alzheimer's dementia recognition through spontaneous speech: A signal processing grand challenge. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–2 (2023)
18. Luz, S., Haider, F., de la Fuente, S., Fromm, D., MacWhinney, B.: Alzheimer's Dementia Recognition Through Spontaneous Speech: The ADReSS Challenge. In: Proc. Interspeech 2020. pp. 2172–2176 (2020)
19. Luz, S., Haider, F., de la Fuente, S., Fromm, D., MacWhinney, B.: Detecting Cognitive Decline Using Speech Only: The ADReSSo Challenge. In: Proc. Interspeech 2021. pp. 3780–3784 (2021)
20. Mei, K., Ding, X., Liu, Y., Guo, Z., Xu, F., Li, X., Naren, T., Yuan, J., Ling, Z.: The ustc system for adress-m challenge. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–2 (2023)

21. Mirheidari, B., Blackburn, D., Christensen, H.: Automatic cognitive assessment: Combining sparse datasets with disparate cognitive scores. In: Proc. Interspeech 2022. pp. 2463–2467 (2022). https://doi.org/10.21437/Interspeech.2022-10205
22. Pappagari, R., Cho, J., Joshi, S., Moro-Velázquez, L., Żelasko, P., Villalba, J., Dehak, N.: Automatic Detection and Assessment of Alzheimer Disease Using Speech and Language Technologies in Low-Resource Scenarios. In: Proc. Interspeech 2021. pp. 3825–3829 (2021)
23. Pérez-Toro, P., Bayerl, S., Arias-Vergara, T., Vásquez-Correa, J., Klumpp, P., Schuster, M., Nöth, E., Orozco-Arroyave, J., Riedhammer, K.: Influence of the Interviewer on the Automatic Assessment of Alzheimer's Disease in the Context of the ADReSSo Challenge. In: Proc. Interspeech 2021. pp. 3785–3789 (2021)
24. Rohanian, M., Hough, J., Purver, M.: Multi-Modal Fusion with Gating Using Audio, Lexical and Disfluency Features for Alzheimer's Dementia Recognition from Spontaneous Speech. In: Proc. Interspeech 2020. pp. 2187–2191 (2020)
25. Rohanian, M., Hough, J., Purver, M.: Alzheimer's Dementia Recognition Using Acoustic, Lexical, Disfluency and Speech Pause Features Robust to Noisy Inputs. In: Proc. Interspeech 2021. pp. 3820–3824 (2021)
26. Rohrer, J.D., Rossor, M.N., Warren, J.D.: Alzheimer's pathology in primary progressive aphasia. Neurobiol. Aging **33**(4), 744–752 (2012)
27. Shah, U., Aryal, S.K.: Experimenting with multimodal autoML: Detection and evaluation of alzheimer's disease (2023), https://openreview.net/forum?id=nSqrgBKBGkv
28. Shah, Z., Qi, S.A., Wang, F., Farrokh, M., Tasnim, M., Stroulia, E., Greiner, R., Plitsis, M., Katsamanis, A.: Exploring language-agnostic speech representations using domain knowledge for detecting alzheimer's dementia. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–2 (2023)
29. Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S.: X-vectors: Robust dnn embeddings for speaker recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5329–5333 (2018). https://doi.org/10.1109/ICASSP.2018.8461375
30. Spering, C.C., Hobson, V., Lucas, J.A., Menon, C.V., Hall, J.R., O'Bryant, S.E.: Diagnostic Accuracy of the MMSE in Detecting Probable and Possible Alzheimer's Disease in Ethnically Diverse Highly Educated Individuals: An Analysis of the NACC Database. The Journals of Gerontology: Series A **67**(8), 890–896 (2012)
31. Syed, M.S.S., Syed, Z.S., Lech, M., Pirogova, E.: Automated Screening for Alzheimer's Dementia Through Spontaneous Speech. In: Proc. Interspeech 2020. pp. 2222–2226 (2020)
32. Tamm, B., Vandenberghe, R., Van Hamme, H.: Cross-lingual transfer learning for alzheimer's detection from spontaneous speech. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–2 (2023)
33. Zhu, Y., Obyat, A., Liang, X., Batsis, J.A., Roth, R.M.: Wavbert: Exploiting semantic and non-semantic speech using wav2vec and bert for dementia detection. In: Proc. Interspeech 2021. pp. 3790–3794 (2021). https://doi.org/10.21437/Interspeech.2021-332

Policy Gradient-Driven Noise Mask

Mehmet Can Yavuz(✉) and Yang Yang

Department of Radiology and Biomedical Imaging, University of California, San Francisco, USA
{mehmetcan.yavuz,yang.yang4}@ucsf.edu

Abstract. Deep learning classifiers face significant challenges when dealing with heterogeneous multi-modal and multi-organ biomedical datasets. The low-level feature distinguishability limited to imaging-modality hinders the classifiers' ability to learn high-level semantic relationships, resulting in sub-optimal performance. To address this issue, image augmentation strategies are employed as regularization techniques. While additive noise input during network training is a well-established augmentation as regularization method, modern pipelines often favor more robust techniques such as dropout and weight decay. This preference stems from the observation that combining these established techniques with noise input can adversely affect model performance.

In this study, we propose a novel *pretraining* pipeline that learns to generate conditional noise mask specifically tailored to improve performance on multi-modal and multi-organ datasets. As a reinforcement learning algorithm, our approach employs a dual-component system comprising a very light-weight policy network that learns to sample conditional noise using a differentiable beta distribution as well as a classifier network. The policy network is trained using the reinforce algorithm to generate image-specific noise masks that regularize the classifier during pretraining. A key aspect is that the policy network's role is limited to obtaining an *intermediate (or heated) model* before fine-tuning. During inference, the policy network is omitted, allowing direct comparison between the baseline and noise-regularized models.

We conducted experiments and related analyses on RadImageNet datasets. Results demonstrate that fine-tuning the intermediate models consistently outperforms conventional training algorithms on both classification and generalization to unseen concept tasks.

Keywords: Pretraining · Medical Imaging · Classification Accuracy · RadImageNet · Policy Gradient Method · Reinforcement Learning

1 Introduction

Image classification is a fundamental task in computer vision that involves assigning labels or categories to images based on their visual content. Traditional approaches to image classification have relied on conventional supervised learning techniques, where the model is trained on a labeled dataset. However, reinforcement learning (RL) has emerged as a promising enhancement to the training

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15311, pp. 414–431, 2025.
https://doi.org/10.1007/978-3-031-78195-7_28

process, enabling models to learn optimal classification policies through interaction with an environment[1][2]. In particular, policy gradient methods in RL offer a powerful framework for directly optimizing classification performance[3].

RL-based approaches offer a promising solution to address the challenges posed by image classification datasets[1,4]. By learning optimal policies for feature extraction and classification through interaction with the environment, RL algorithms can adapt to the variations in low-level features and capture the relevant high-level semantic relationships [2]. This enables the development of more robust and accurate image classification models.

Medical imaging datasets, such as MRI scans, ultrasound images, and CT scans, often exhibit significant heterogeneity and variations in low-level features due to differences in acquisition protocols, imaging modalities, and anatomical regions an example illustration is in Figure 1. These variations manifest as differences in brightness, contrast, and noise levels, which can be readily discerned even through visual inspection of image histograms. Such low-level distinguishability hinders the learning of high-level semantic relationships. The challenge is to capture and homogenize these low-level image features to improve classifier performance.

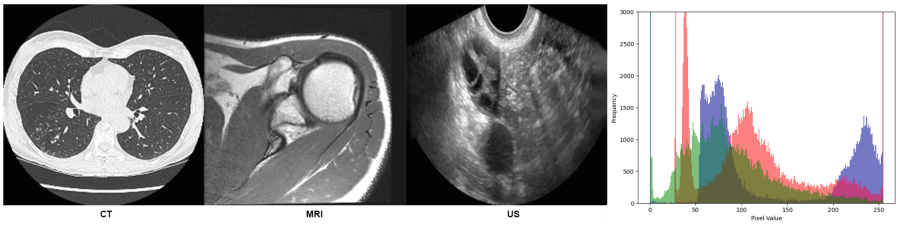


Fig. 1. Heterogeneity in medical imaging datasets (RadImageNet [5]) across modalities and anatomical regions. (Left) CT scan of the lungs. (Second left) MRI scan of the shoulder. (Second Right) Ultrasound image of the ovary. (Right) Pixel intensity histograms as indicator for low-level image features of CT (blue), MRI (red), and ultrasound (green) images, illustrating variations in brightness, contrast, and noise levels.

Previously, injecting random noise into the input data during training has been proposed as a regularization approach to improve generalization [6]. However, the prevalence of other powerful regularizers like weight decay and dropout raises the question of whether additional noise-based regularization is beneficial [7]. Moreover, prior work has highlighted the potential negative impact of noise-based regularization, an important consideration given that modern training pipelines typically omit additive or multiplicative noise [8].

In this paper, we present several key contributions to tackle the challenges encountered by deep learning classifiers when working with heterogeneous multimodal and multi-organ biomedical datasets. Our primary contribution is a novel *pretraining* pipeline that learns to generate conditional noise masks specifically

designed to enhance performance on these datasets. We propose a reinforcement learning system that consists of a lightweight policy network and a classifier network as shown in Figure 2. During pretraining, the policy network is optimized to generate image-specific noise masks that regularize the classifier, effectively improving its performance on complex biomedical datasets. This approach enables the classifier to better handle the heterogeneity and multi-modality of the data, leading to more accurate and robust predictions in various biomedical applications.

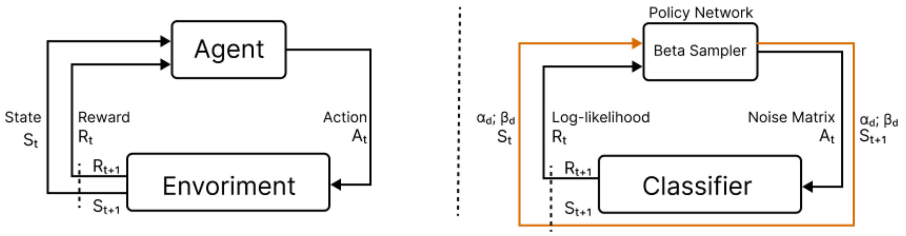


Fig. 2. Schematic diagram of Reinforcement learning. At left agent takes action and change the state in environment and gain reward. At right, beta sampler (policy network) generates noise matrix and classifier as a differentiable environment computes the log-likelihood and updates the state variables α and β.

In addition to our novel pretraining pipeline, we perform comprehensive experiments and analyses on RadImageNet datasets to validate the effectiveness of our proposed approach. The results consistently show that fine-tuning the intermediate (or heated) models obtained through our *pretraining* pipeline outperforms conventional training algorithms on both classification tasks and generalization to unseen concepts. This superior performance underscores the potential of our reinforcement learning-based noise regularization technique in enhancing the robustness and adaptability of deep learning classifiers when faced with challenging biomedical imaging scenarios.

In summary, our contributions significantly advance the state-of-the-art in deep learning for heterogeneous multi-modal and multi-organ biomedical datasets. By introducing a novel reinforcement learning-based approach, we effectively address the limitations of existing regularization techniques and provide a powerful tool for improving the performance of deep learning classifiers in complex biomedical imaging scenarios.

The remainder of this paper is organized as follows. The next sub-section provides an overview of related work in the field of medical image classification and reinforcement learning methodologies. Section 2 describes our proposed pretraining algorithm and background in detail, including training procedure. Section 3 presents the dataset, experimental setup and Section 4 includes results along with a discussion of the findings. Finally, Section 5 outlines and concludes the paper.

1.1 Literature Review

Holmstrom et al. were among the pioneers in addressing this issue by using additive noise in back-propagation training, which can be seen as an early form of regularization to prevent overfitting [9]. It is extended to the concept of the domain of speech recognition, demonstrating the effectiveness of noisy training for deep neural networks [10].

Bishop provided a theoretical foundation for training with noise, showing that it is equivalent to Tikhonov regularization, which adds a penalty term to the loss function to control the complexity of the model [6]. This concept is explored image recognition with deep neural networks in the presence of noise, demonstrating that distortions can be both a challenge and an opportunity for model training [7]. Enhancing the generalisation abilities of neural networks (NNs) through integrating noise such as MixUp or Dropout during training has emerged as a powerful and adaptable technique. Despite the proven efficacy of noise in NN training, there is no consensus regarding which noise sources, types and placements yield maximal benefits in generalisation and confidence calibration. [11]

It is proposed deep neural network architectures that are robust to adversarial examples, which are inputs crafted to deceive the model into making incorrect predictions [12]. This work is part of a broader effort to develop models that maintain high performance in the presence of input perturbations.

Dropout is introduced, a simple yet effective technique to prevent neural networks from overfitting [13]. Dropout works by randomly omitting a subset of features during training, which encourages the model to learn more robust features. Further contributed to this field by introducing Cutout, a regularization method that randomly masks out sections of input images during training, forcing the network to focus on less prominent features [14].

The use of semantic segmentation for masking and cropping input images has proven to be a significant aid in medical imaging classification tasks. The proposal of a novel joint-training deep reinforcement learning framework for image augmentation called Adversarial Policy Gradient Augmentation (APGA) that shows promising results on medical imaging tasks [15].

RadImageNet, represents a significant step forward in the domain of medical imaging [5]. It is an open radiologic dataset designed to facilitate effective transfer learning in deep learning research. The MedMNIST Classification Decathlon, as presented by Yang et al., is a lightweight benchmark for medical image analysis designed to assess the capabilities of automated machine learning (AutoML) solutions [16].

2 Policy Gradient-Driven Noise Mask

Our novel pretraining pipeline streamlines heterogeneous biomedical data through a three-pronged approach:

– Adapting the noise masks to each specific image, accounting for variations in modality and organ type.

– Using reinforcement learning to optimize the noise masks for improved classifier performance.
– Separating the noise mask generation (pretraining) from the final model (fine-tuning), allowing for more flexible and effective regularization.

We describe the mathematical formulation of the action taken by the policy network and the subsequent interaction with the environment, which leads to the computation of the loss function used for training. The order of steps follows the Figure 3 from left to right.

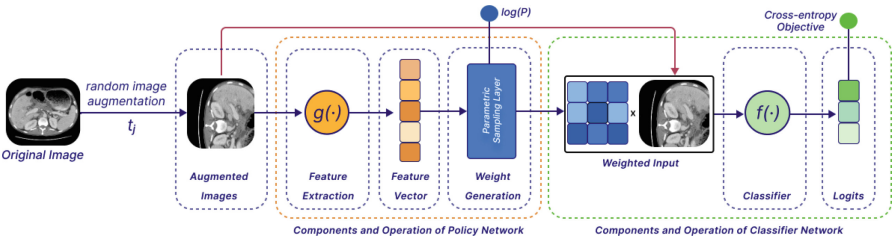


Fig. 3. Diagram of our proposed pipeline using deep learning, illustrating the process from original image through stochastic masking, feature extraction, beta sampling and classification to produce a prediction with a cross-entropy objective. The blue and green color parameters to compute objective function.

Image preprocessing is the initial stage, where raw images are prepared for further processing. This step may include normalization, resizing, and other image augmentation techniques. Mathematically, if I represents the raw image, the preprocessing step can be represented as:

$$I_{pre} = t_j(I). \tag{1}$$

where I_{pre} is the preprocessed image and the stochastic function $t_j(\cdot)$ is used to obtain random augmentations of the input image j.

Action Given an pre-processed input image I_{pre}, the policy network computes image specific parameters α_{image} and β_{image} based on the dataset specific parameters represented by $\alpha_{dataset}$ and $\beta_{dataset}$. The updated parameters are obtained as follows:

$$\alpha_{image}, \beta_{image}, \alpha_{dataset}, \beta_{dataset} = \text{PolicyNet}(I, \alpha_{dataset}, \beta_{dataset}). \tag{2}$$

PolicyNet consists of feature extrator network $g(\cdot)$ and beta sampling operation based on features as shown in Figure 3. The extracted feature vector let the beta sampler network to generate image specific noise mask as shown in Figure 4.

These parameters are then exponentiated to ensure they are positive, as required by the Beta distribution:

$$\alpha' = e^{\alpha_{image}}, \quad \beta' = e^{\beta_{image}}. \tag{3}$$

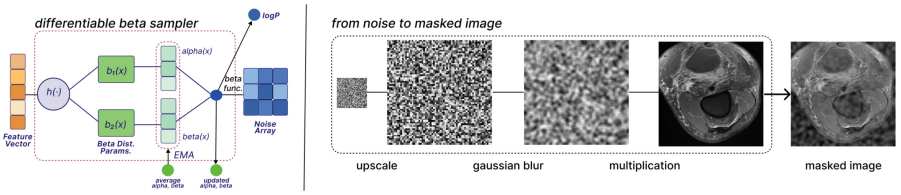


Fig. 4. Diagram of a policy network architecture (at left) showing the flow from input feature tensor to weighted tensor output. The network processes the input through a function $h(\cdot)$, projects the feature vector into Beta distribution parameters $b_1(x)$ and $b_2(x)$, derives alpha and beta values for the Beta function, calculates log probability (logP). The outputs are visualized over the dashed region as colorful circles. As an example (at right), the process of applying a stochastic masking to a medical image, showcasing the transformation from a noise matrix to the final masked image, which is part of the image processing pipeline involving steps such as noise matrix acquisition, upsampling, blurring, and applying.

A Beta distribution $\mathcal{B}$ is then defined using these parameters and a mask M is sampled from this distribution:

$$M \sim \mathcal{B}(\alpha', \beta'). \tag{4}$$

The mask is reshaped to match the dimensions of the input image and then interpolated if necessary to match the image dimensions.

We also introduce post-processing steps that involve upsampling and blurring as shown in Figure 4. Firstly, we upsample the low-resolution noise matrix obtained from the policy network to obtain image regions with similar coefficient values, improving the correlation between neighboring pixels. Secondly, we apply a blurring operation to the upsampled noise matrix to avoid sharp transitions between different regions during the convolution operation, smoothing out the boundaries between areas with different noise values.

Environment The model takes the element-wise product of the input image I and the mask M_{in} to get the output:

$$O = \text{Model}(I_{pre} \odot M_{\text{in}}). \tag{5}$$

The environment in this context is implicit. It consists of: (1) The distribution of input images I that the model encounters. (2) The task-specific criteria that determine the reward, which is based on the model's output O and the target labels T. (3) The Beta distribution from which the mask M_{in} is sampled, which forms part of the action space of the policy.

Policy Gradient Objective The policy gradient method is a fundamental approach in reinforcement learning for optimizing a policy function $\pi_\theta(a|s)$, where θ represents the policy parameters, a denotes the action taken, and s is the current state. The primary objective is to maximize the expected return $J(\theta)$. This is achieved by computing the gradient of the objective function with respect to the policy parameters θ, given by:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi\theta}\left[\sum_{\eta=0}^{T} \nabla_\theta \log \pi_\theta(a_\eta|s_\eta) R(\tau)\right]. \tag{6}$$

Here, η represents the time-step, and τ denotes a trajectory sampled from the policy π_θ. This formulation allows for the update of policy parameters in the direction that increases the expected return.

The policy gradient objective comprises two key components: (1) The log-probability output from the stochastic policy network. (2) The value of the reward function.

For the specific application described, the log-probability of the occurrence of the mask M_{in} under the Beta distribution is computed as:

$$\log(P) = \mathcal{B}.logP(M_{\text{in}}). \tag{7}$$

This log-probability represents the output of the policy network.

The reward function, in this case, is defined as the cross-entropy between the output logits and the corresponding labels for the input image (which represents the state).

The loss function, combining these elements, is formulated as the mean of the product of the exponentiated log probabilities and the criterion applied to the output and the target:

$$\mathcal{L} = \text{mean}\left(\log\left(\sum e^{\log(P)}\right) \times \text{Criterion}(O, T)\right). \tag{8}$$

where O represents the output logits and T denotes the target labels. This formulation of the policy gradient method provides a framework for optimizing the policy in a reinforcement learning context, specifically tailored to the task of mask generation under a Beta distribution.

Image and Dataset Specific Shape Parameters The beta distribution is a continuous probability distribution defined on the interval $(0, 1)$ and is parameterized by two positive shape parameters, typically denoted as α and β. The probability density function (PDF) of the beta distribution with parameters α and β is given by:

$$f(x; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 < x < 1. \tag{9}$$

where $\Gamma(\cdot)$ is the gamma function, which is a generalization of the factorial function to complex numbers.

The beta distribution is a versatile distribution that can take on various shapes depending on the values of the shape parameters α and β. The extracted feature vectors are transformed by a projection layer $h(\cdot)$ (as shown in Figure 4), which generates two separate tensors - alpha and beta - the beta distribution parameters. The beta distribution parameters - alpha and beta - are used to sample a random tensor via the beta function. This beta matrix contains values between 0 and 1. The logP output provides the log probability of this tensor under the beta distribution, which could be useful for training or interpreting the model.

The alpha and beta parameters from the last iteration step are the input for the policy network and the update function for alpha and beta are determined by recursive exponential moving averages formula:

$$\alpha_{new,image} = \tau_i \cdot \alpha_{dataset} + (1 - \tau_i) \cdot \alpha_{image}. \tag{10}$$

$$\beta_{new,image} = \tau_i \cdot \beta_{dataset} + (1 - \tau_i) \cdot \beta_{image}. \tag{11}$$

$$\alpha_{dataset} = \tau_d \cdot \alpha_{dataset} + (1 - \tau_d) \cdot \alpha_{mean,image}. \tag{12}$$

$$\beta_{dataset} = \tau_d \cdot \beta_{dataset} + (1 - \tau_d) \cdot \beta_{mean,image}. \tag{13}$$

The alpha and beta parameters in the given formulas, in a recursive manner, are used to update the Beta distribution parameters for an image-level as well as dataset-level.

The formulas use exponential moving averages to update the α and β parameters for both the current image and the overall dataset:

- $\alpha_{dataset}$ and $\beta_{dataset}$ are updated by taking a weighted average of their previous values and the mean of the image-level parameters across the dataset ($\alpha_{mean,image}$, $\beta_{mean,image}$).
- $\alpha_{new,image}$ and $\beta_{new,image}$ are updated by taking a weighted average of the dataset-level parameters ($\alpha_{dataset}$, $\beta_{dataset}$) and the current image-level parameters (α_{image}, β_{image}). The weight is controlled by τ_i, typically set to 0.9.

2.1 The Algorithm

The pseudo-code for the proposed algorithm is given in Algorithm 1: **Initialization** The algorithm begins with the random initialization of model parameters. Additionally, dataset level parameters α_d and β_d are also initialized. These parameters are crucial as they will be updated throughout the training process to optimize the regularization mechanism. **Optimization** The core of the algorithm is an iterative process that continues until a specified termination condition is met, which in this case is convergence. **Compute Alpha and Beta**

Result: Optimized model parameters
Randomly initialization model parameters and dataset level α_d, β_d;
while *not converged* **do**
  compute the α, β*;*
  $\alpha_i, \beta_i, \alpha_d, \beta_d \leftarrow$ update mask params$(image, \alpha_i, \beta_i, \alpha_d, \beta_d)$;
  dist $\leftarrow$ Beta(α_i, β_i);
  sample the noise matrix;
  noise matrix $\leftarrow$ dist.sample();
  take action;
  cross entropy obj. $\leftarrow$ classifier.step$(image,$ noise matrix$)$;
  loss $\leftarrow$ dist.log_prob(noise matrix) $\times$ cross entropy obj.;
  loss.backward();
end

Algorithm 1: The policy gradient method algorithm for training policy network together with classifier mechanism.

At each iteration, the algorithm computes the values of α and β. These are the parameters of the Beta distribution, which is used to model the stochastic nature of the regularization mechanism. **Update Mask Parameters** The function 'update mask params' is called with the current image and the alpha and beta parameters for both the individual and dataset levels. This function adjusts the parameters to better fit the data as the algorithm learns. **Sample Noise Mask** A noise mask is sampled from the Beta distribution parameterized by the updated α_i and β_i. This noise mask represents the probabilistic decisions made by the regularization mechanism at this stage of training. **Take Action** The classifier takes an action based on the current image and the sampled weight matrix. This step involves the classifier making a prediction which is then used to calculate the objective function, in this case, the cross-entropy loss. **Calculate Loss and Backpropagate** The loss is calculated by taking the log probability of the sampled weight matrix from the Beta distribution and scaling it by the cross-entropy objective. This loss is then backpropagated through the network to update the model parameters in a direction that minimizes the loss. The loop continues until convergence.

3 Experiments

Datasets For pre-training our models, we are utilizing stratified (train/val/test) split of RadImageNet [5], a large-scale multi-modal and multi-organ medical imaging dataset (see Appendix A). The split we prepare let us to justify model performance using different training techniques. This diverse dataset should help our model learn general features and representations as well as dataset specific comparison. To evaluate the performance of our pre-trained model, we are using the enhanced MedMNIST Classification Decathlon [16], which includes the original 10 MedMNIST datasets, as well as 2 additional MRI datasets and 1 ultrasound (US) dataset (seee Appendix B). This comprehensive benchmark covers

a wide range of medical imaging tasks, modalities (e.g., X-ray, CT, MRI, US, Microscope, OCT), and anatomical regions. By assessing our model's performance on the enhanced MedMNIST Decathlon, we can determine how well it generalizes across various medical imaging applications.

Implementation Details [1][2] The training pipeline is configured through a set of hyper-parameters. The main model to train is Resnet-50 and the policy network is always a leight-weight network such as Resnet-10t. The batch size is 32 for each 8xV100 GPU with effective batch size is 256. Each training phase takes 90 epochs with SGD optimizer with learning rate 0.1, momentum 0.9 and weight decay 1e-4. The step learning rate scheduler reduce by 1/10 in 30 epochs cycle. The Resnet-10t policy network starts with 0.01 learning rate and same momentum and weight decay and using cosine annealing learning rate scheduler. The initial values for $\alpha_{dataset}$ and $\beta_{dataset}$ are random and sampled from normal distribution.

For unseen concepts, we extract the features from the freezed backbone network and using MLP for unseen concept generalization or using Logistic Regression for low-shot adaptability. Both pre-training phases use AdamW with default hyper-parameters until convergences (with no accuracy increment for 5 epochs).

Baselines, Ablation Study and Optimal Model For the ablation study, we determine the optimal hyperparameters for the upscale coefficient, kernel size, and stride through a systematic search. In ablation study, we employ Resnet-10t as the backbone and the policy network due to its compact size and ease of optimization. These experiments provide a comprehensive understanding of the model's behavior under different settings and help identify the most suitable configuration for the given task. In the evaluation task, we use Resnet50 as backbone and Resnet-10t as policy network models. Two model are compared: a baseline model and one improved training with a Gradient Policy technique. The performance (macro) metrics considered included *Precision*, *Recall*, *F1 Score*, *AUROC*, and *Balanced Accuracy*.

Case Analysis We investigate the scenario where no upscaling is applied, and instead, pixel-level noise is directly introduced to the input. *Baseline Performance:* Without any noise model applied, the performance metrics serve as a baseline. *Different Noise Models:* The application of Gaussian and Uniform noise models, following fine-tuning, at 32x32 and 64x64 noise matrix. *Pure Noise Conditions:* Under conditions simulating pure noise (noise matrix equal to image size, 224x224).

Generalization to unseen concepts We evaluate the generalization performance of our model on unseen concepts using the protocol proposed [17]. For

[1] https://github.com/ogrenenmakine/Policy-Gradient-Driven-Noise-Mask
[2] https://huggingface.co/ogrenenmakine/Policy-Gradient-Noise-Mask

RadImageNet, except for the modalities, the samples and classes from MedMNIST are unseen concepts. The model is pretrained on three datasets: ImageNet-1K (IN1K), RadImageNet (RadIN), and RadImageNet using Gradient Policy (Grad. P. RadIN). We then extract features for each downstream dataset and evaluate the performance using a randomly initialized multi-layer perceptron.

How fast can models adapt to unseen concepts? We evaluate the model performance for unseen concepts using low-shots proposed in [17]. We use CT, MRI, US and XR samples from MedMNIST dataset and the sample numbers are 8, 16, 32, 64, 128 and 256, respectively. The model is pretrained on three datasets: ImageNet-1K (IN1K), RadImageNet (RadIN), and RadImageNet using Gradient Policy (Grad. P. RadIN).

Table 1. Entropy and performance comparison of normal and heated models with different input types during policy gradient-driven training on RadImageNet. Bolds indicate the best balanced accuracy (higher is better). The underlined score is the best entropy (lower is better.)

Model	Input	Entropy	Precision	Recall	F1	ROC	B.Acc.
(Normal)	Normalized	0.3802	0.5929	0.5014	0.5226	0.9884	0.5014
(CutMix)	Normalized	0.3259	0.6022	0.5225	0.5444	0.9875	**0.5225**
(Heated)	Normalized	0.7156	0.2906	0.2825	0.2417	0.9497	0.2825
(Heated)	Noisy	0.3294	0.5967	0.5136	0.5402	0.9898	0.5136
(Finetuned Heated)	Normalized	<u>0.3177</u>	0.6034	0.5211	0.5468	0.9900	**0.5211**

4 Results & Discussion

We start by our experiment by explaining the intermediate (or heated) model concept experimentally in Table 1. The results demonstrate the evolution of model entropy through different stages of training and input types. Initially, the heated model shows high entropy (0.7156) with normalized input, indicating a state of uncertainty. When trained with noisy input, the entropy decreases significantly (0.3294), suggesting improved robustness. After finetuning on the target domain, the heated model achieves the lowest entropy (0.3177) among all models, including normal and CutMix variants. This low entropy, combined with competitive performance metrics, indicates that the finetuned heated model has learned more effectively from the data compared to other approaches. Notably, while the finetuned heated model and CutMix model have similar accuracy (balanced accuracy of 0.5211 and 0.5225, respectively), the finetuned model's lower entropy suggests more confident and potentially more reliable predictions.

Table 2 compares the performance of lightweight Resnet-10 and baseline Resnet-50 models with and without the gradient policy technique in different

augmentations settings (hard augmentation: CutMix [18]: soft augmentation: AugMix [19], AutoAug [20], RandAug [21]). The intermediate model obtained by policy gradient technique is fine-tuned RadImageNet. For both model sizes, applying the gradient policy improves all metrics, compare to normal training and provides competitive results with CutMix.

Table 2. Performance Metrics for ResNet Models with Best Parameter Settings in RadImageNet

Technique	Precision	Recall	F1	ROC	B.Acc.
Baseline Model (ResNet-50)					
Heavy Augmentations					
Grad.P.⊙AutoAug → *FT*	0.6034	0.5211	0.5468	0.9900	**0.5211**
CutMix⊙AutoAug	0.6022	0.5225	0.5444	0.9875	**0.5225**
Augmentations					
RandAug	0.5877	0.5102	0.5313	0.9898	0.5102
AugMix	0.6037	0.5078	0.5326	0.9892	0.5078
AutoAug	0.5929	0.5014	0.5226	0.9884	0.5014
Lightweight Model (ResNet-10t)					
Heavy Augmentations					
Grad.P.⊙AutoAug → *FT*	0.5672	0.4573	0.4838	0.9890	**0.4573**
CutMix⊙AutoAug	0.5405	0.4292	0.4523	0.9871	0.4292
Augmentations					
RandAug	0.5110	0.4060	0.4275	0.9864	0.4060
AugMix	0.5111	0.4096	0.4297	0.9867	0.4096
AutoAug	0.5386	0.4262	0.4479	0.9871	0.4262

Note: Grad.P.: Gradient-Policy Noise Mask (this work) using best parameter setting: Noise Matrix 64x64, K:13 and S:6 Gaussian Blurring; FT: Finetuning; CutMix [18], AugMix [19], AutoAug [20], RandAug [21]. 2

Table 3 shows the results on enhanced medical imaging datasets. It reports the F1 scores of feature extractor Resnet-50 backbone (pretrained with ImageNet-1K, RadImageNet, and RadImageNet intermediate model with gradient policy, respectively) and evaluated using MLP over extracted feature vectors. The intermediate model with gradient policy achieves the highest F1 scores on 11 out of 13 datasets, demonstrating strong generalization to unseen medical imaging concepts. The improvements are especially significant for the small-scale datasets. It is categorized into three: the small datasets is less than 10,000 samples, mid-scale dataset range is between 10,000 and 30,000 samples, and the large-scale datasets are over 100,000 samples.

Figures 5 reports the few-shot adaptability of each CT, MRI and US modalities in 8 different dataset and 3 pre-trained networks (ImageNet-1K, RadIm-

Table 3. The F1 scores (higher is better) indicate the generalization to unseen concepts by MedMNIST dataset [16] using Resnet-50 backbone. The results columns are for ImageNet, RadImageNet pretrained weights as well as Gradient Policy pretrained intermediate (heated) models. The t-distribution %95 confidence scores are also provided for small (7 trials) and mid-scale (3 trials) datasets.

Dataset	IN1K	RadIN	Grad. P. RadIN
Resnet-50			
Small Sets (7 Trials Averaged)			
Breast US	0.6662±0.0394	0.6866±0.0312	**0.7649±0.0154**
Breast Cancer US	0.5932±0.0137	0.5659±0.0170	**0.6563±0.0111**
BrainTumor MR	0.8431±0.0093	0.8275±0.0079	**0.8880±0.0041**
Brain MR	0.2258±0.0187	0.2257±0.0154	**0.4351±0.0345**
Pneumonia XR	0.8344±0.0093	0.8343±0.0085	**0.8554±0.0101**
Mid-scale Sets (3 Trials Averaged)			
Blood Cell Mic.	0.9099±0.0024	0.8686±0.0066	0.**9261±0.0024**
Dermatoscope	**0.4998±0.0179**	0.3596±0.0602	0.4643±0.0266
OrganA CT	0.7272±0.0043	0.7339±0.0134	**0.8108±0.0033**
OrganC CT	0.6579±0.0119	0.6772±0.0094	**0.7438±0.0037**
OrganS CT	0.5622±0.0068	0.5649±0.0121	**0.6170±0.0009**
Large-Scale Sets (1 Trial)			
Retinal OCT	0.5160	0.5220	**0.5736**
Colon Pathology	**0.8138**	0.7880	0.8042
Tissue Mic.	0.3082	0.3447	**0.4033**

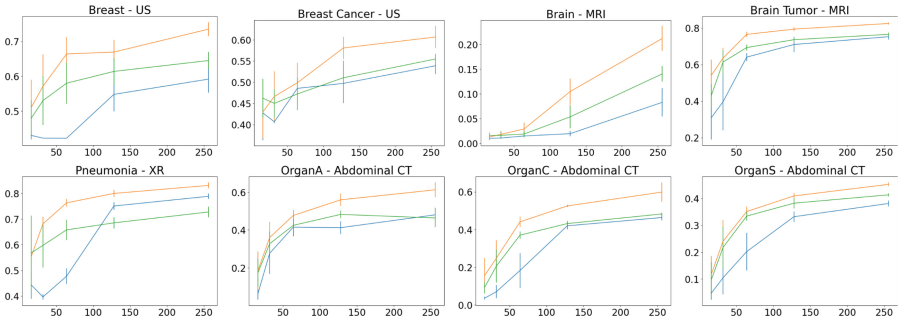


Fig. 5. Each figure represents a different unseen but related dataset: Breast - US, Breast Cancer - US, Brain - MRI, Brain Tumor - MRI, Pneumonia - XR, OrganA, OrganC, OrganS - Abdominal CT. Different color curves corresponds to different pretrained model on few-shot adaptability. The orange curves represent the performance of Gradient Policy RadImageNet, the green curves show the results for ImageNet, and the blue curves indicate the performance of ImageNet pretrained models. The evaluation is carried out for 8, 16, 32, 64, 128, 256 samples and 10 trials in each sample size. The vertical lines are the t-statistics %95 confidence interval.

Table 4. Performance Metrics for Different Noise Models and Noise Matrix Sizes for RadImageNet [5] by ResNet-10t model.

Features	Matrix	Precision	Recall	F1	AUROC	B. Accuracy
Different Noise Models						
Baseline Model	-	0.5386	0.4262	0.4479	0.9871	0.4262
Gaussian + Blurring	32x32	0.5354	0.4273	0.4494	0.9871	0.4273
Uniform + Blurring	32x32	0.5206	0.4278	0.4480	0.9872	0.4278
Gaussian + Blurring	64x64	0.5507	0.4279	0.4497	0.9874	0.4279
Uniform + Blurring	64x64	0.5504	0.4299	0.4499	0.9861	0.4274
Pure Noise						
Baseline Model	-	0.5386	0.4262	0.4479	0.9871	0.4262
Gaussian	224x224	0.5312	0.4236	0.4441	0.9864	0.4253
Gaussian + Blurring	224x224	0.5396	0.4312	0.4534	0.9876	0.4312
Pure Noise	224x224	0.5340	0.4314	0.4555	0.9876	0.4314
Pure Noise + Blurring	224x224	NaN	NaN	NaN	NaN	NaN

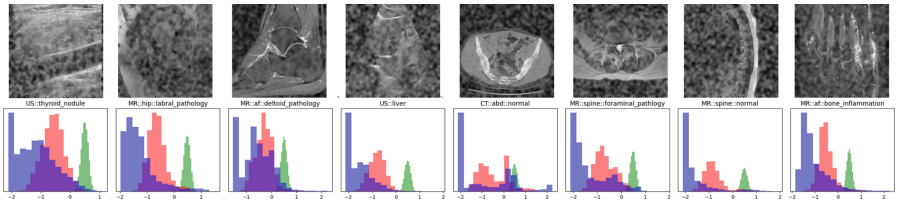


Fig. 6. Composite image displaying a series of medical scans from RadImageNet [5] (1st row) with corresponding histograms (2nd row), where blue bars represent the original image pixel intensity distribution, red bars indicate the masked image pixel intensity distribution, and green bars show the noise pixel intensity distribution.

ageNet, and RadImageNet with gradient policy) for (8, 16, 32, 64, 128, 256 samples and 10 trials in each sample size). The orange curves are pre-trained model with gradient policy which is consistently better in few-show adaptability.

Table 4 shows the impact of different noise models and noise matrix sizes on the performance of the lightweight Resnet-10t model. Gaussian noise with a noise matrix of 64x64 yields the best precision, while uniform noise with a noise matrix of 64x64 gives the highest recall and accuracy. However, the differences between noise models are relatively small. Using pure noise matrix of size 224x224 leads to slightly lower but comparable performance to the baseline.

Figure 6 presents medical scans with their respective histograms, indicating low-level features and pixel intensity distributions. The stochastic masking operation performed by the policy network modifies the skewness and center of distribution using pixel-wise multiplication, enhancing the image representation for the classifier and achieving a form of homogenization.

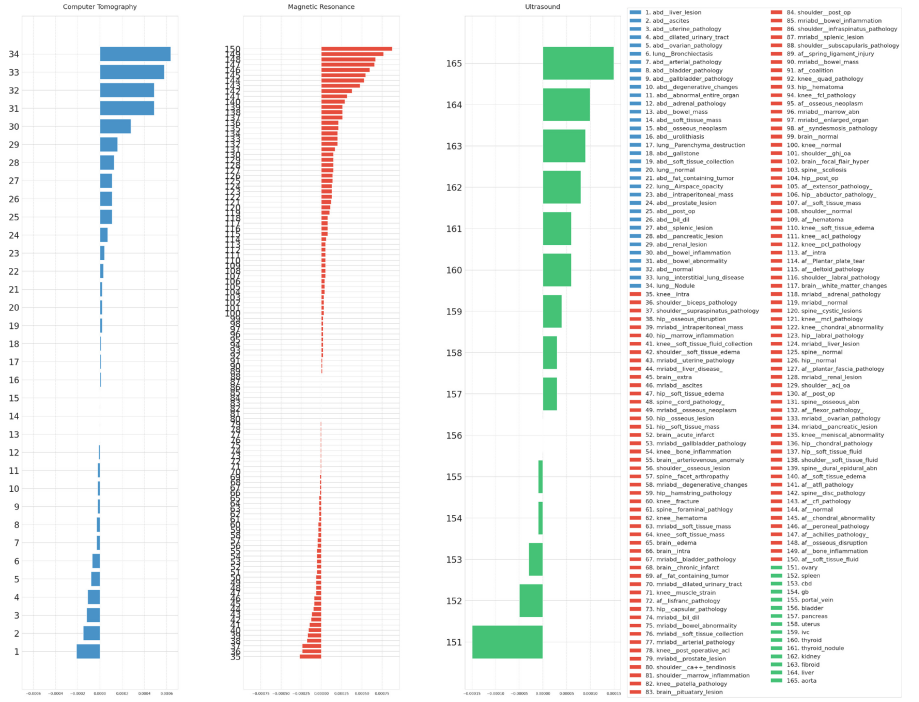


Fig. 7. Comparison of prediction differences between a Normal Model and a Gradient Policy Model across three medical imaging modalities: Computer Tomography (CT - blue), Magnetic Resonance Imaging (MRI - red), and Ultrasound (US - green). The horizontal bars represent the normalized difference in predictions for various anatomical structures and pathologies. Positive values indicate higher prediction rates by the Normal Model, while negative values show higher rates by the Gradient Policy Model. The red dashed line at zero serves as a reference point for equal performance between the two models.

Figure 7 presents a comprehensive analysis comparing the gradient policy trained model to the conventional training approach. The diagram illustrates the performance tendency towards Policy Gradient per-labels

The results indicate that the proposed gradient policy technique consistently improves the performance of both lightweight and larger models for medical image classification. This suggests that the gradient policy helps the models learn more robust and generalizable features. However, the performance on some unrelated datasets, such as dermatoscope images and tissue microscopy, remains relatively low. Because the low-level image features are closer to natural images by ImageNet.

Interestingly, even though the noise matrix size has an effect on our proposed model, the known distributions such as Gaussian or Uniform does not effected. The pure noise condition does not substantially impact the performance metrics

either. This could imply that the model is able to effectively handle different types of noise perturbations.

An important remark on the convergence is that the dataset level α and β get 'almost' uniform distribution shape after beta sampling operation almost always.

It is noteworthy that our experiments conducted on natural images sourced from the ImageNet-1K dataset did not yield superior accuracy compared to existing methods, despite our best efforts and the application of novel techniques.

5 Conclusion

In summary, the Gradient Policy technique has demonstrated its effectiveness in enhancing the performance and generalization capabilities of deep learning models in biomedical image analysis. The ablation study highlights the superiority of the our proposed training schema using Gradient Policy technique over the conventional training across all performance metrics. This is achieved by fine-tuning hyper-parameters such as grid size and Gaussian blurring parameters. Moreover, the technique's ability to improve the performance of larger models like Resnet-50 further underscores its versatility and scalability. On the other hand, the case analysis reveals that while variations in noise models such as using normal or uniform distribution or pure noise noise condition lead to minor performance differences, no statistically significant improvement is observed.

The model's generalization performance on unseen concepts, evaluated using the protocol proposed by Sariyildiz et al., demonstrates the consistent superiority of the model pretrained on RadImageNet using Gradient Policy over models pretrained on ImageNet-1K and RadImageNet across all downstream datasets. This finding emphasizes the technique's ability to enhance the model's capacity to adapt to novel concepts and domains.

Furthermore, the low-shot adaptation performance on unseen concepts showcases the remarkable ability of the model pretrained on RadImageNet using Gradient Policy to quickly adapt to new concepts with limited samples, consistently outperforming models pretrained on ImageNet-1K and RadImageNet. This adaptability is crucial in the medical domain, where data scarcity and concept generalization are common challenges. The Gradient Policy technique not only improves the model's overall accuracy but also enables it to focus on relevant features and adapt quickly to unseen concepts with limited samples.

A RadImageNet: Artifacts and The Refinement

The RadImageNet database contains a diverse collection of ultrasound images, many of which exhibit various artifacts including text annotations, markers, and colorful segmentation masks. To address this issue, we developed an algorithm that automatically processes the standard RadImageNet distribution by implementing stratified splitting and removing ultrasound artifacts. This refined

version of RadImageNet used in the paper, along with its corresponding code, is available on GitHub[3] (Fig. 8)

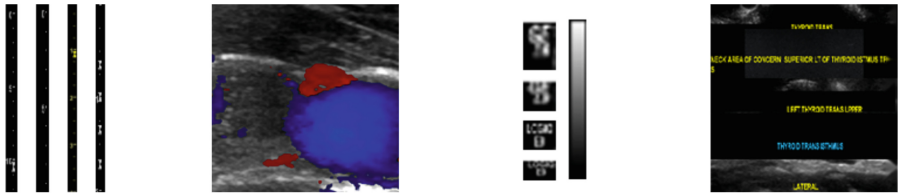


Fig. 8. Typical examples for the artifacts in ultrasound images: markers, segmentations, signs & colorbar and texts, respectively.

B Enhanced MedMNIST

MedMNIST is a comprehensive collection of standardized biomedical images designed for various analytical tasks in the medical field. This dataset has been expanded to include three new subsets, broadening the range of imaging modalities and classification challenges available to researchers. These additions complement the existing MedMNIST collections, offering a more diverse set of resources for developing and evaluating machine learning models across various medical imaging applications. The enhanced MedMNIST collection, including both existing and new datasets, is accessible on Hugging Face [4] (Table 5)

Table 5. The additional datasets for MedMNIST, the classification task including the number of classes (e.g., Multi-Class (MC) with 9 classes, Binary-Class (BC) with 2 classes), and the total number of samples contained within each dataset.

Name	Data Modality	Task (#Cls)	# Samples
Brain Tumor Dataset	Magnetic Resonance	MC (3)	3,064
Brain Dataset	Magnetic Resonance	MC (23)	1,600
Breast Cancer	Ultrasound	BC (2)	1,875

References

1. Marco, A., Wiering "Reinforcement learning algorithms for solving classification problems". In, et al.: IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL). IEEE. **2011**, 91–96 (2011)

[3] https://github.com/ogrenenmakine/Refined-RadImagenet.
[4] https://huggingface.co/datasets/ogrenenmakine/Enhanced-MedMNIST

2. Soumyendu Sarkar et al. “Rl-cam: Visual explanations for convolutional networks using reinforcement learning”. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023, pp. 3860–3868
3. Fadi AlMahamid and Katarina Grolinger. “Reinforcement learning algorithms: An overview and classification”. In: 2021 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE). IEEE. 2021, pp. 1–7
4. Masato Fujitake. “RL-LOGO: Deep Reinforcement Learning Localization for Logo Recognition”. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. 2024, pp. 2830–2834
5. Xueyan Mei et al. “RadImageNet: an open radiologic deep learning research dataset for effective transfer learning”. In: Radiology: Artificial Intelligence 4.5 (2022), e210315
6. CM Bishop. “Training with noise is equivalent to Tikhonov regularization”. In: Neural computation 7.1 (1995), pp. 108–116
7. Michał Koziarski and Bogusław Cyganek. “Image recognition with deep neural networks in presence of noise-dealing with and taking advantage of distortions”. In: Integrated Computer-Aided Engineering 24.4 (2017), pp. 337–349
8. M Eren Akbiyik. “Data augmentation in training CNNs: injecting noise to images”. In: arXiv preprint arXiv:2307.06855 (2023)
9. L Holmstrom and P Koistinen. “Using additive noise in back-propagation training”. In: IEEE Trans Neural Netw 3.1 (1992), pp. 24–38
10. Yin, S., et al.: Noisy training for deep neural networks in speech recognition. EURASIP Journal on Audio, Speech, and Music Processing **2015**, 1–14 (2015)
11. Martin Ferianc et al. “Navigating noise: A study of how noise influences generalisation and calibration of neural networks”. In: Transactions on Machine Learning Research (2024), pp. 1–44
12. S Gu and L Rigazio. “Towards deep neural network architectures robust to adversarial examples”. In: arXiv preprint arXiv:1412.5068 (2014)
13. N Srivastava et al. “Dropout: a simple way to prevent neural networks from overfitting”. In: The journal of machine learning research 15.1 (2014), pp. 1929–1958
14. T DeVries and GW Taylor. “Improved regularization of convolutional neural networks with cutout”. In: arXiv preprint arXiv:1708.04552 (2017)
15. Kaiyang Cheng et al. “Adversarial policy gradient for deep learning image augmentation”. In: Medical Image Computing and Computer Assisted Intervention-MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13-17, 2019, Proceedings, Part VI 22. Springer. 2019, pp. 450–458
16. Yang, J., Shi, R., Ni, B., “Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis”. In,: IEEE 18th International Symposium on Biomedical Imaging (ISBI). IEEE. **2021**, 191–195 (2021)
17. Mert Bulent Sariyildiz et al. “Concept Generalization in Visual Representation Learning”. In: International Conference on Computer Vision. 2021
18. Sangdoo Yun et al. “Cutmix: Regularization strategy to train strong classifiers with localizable features”. In: Proceedings of the IEEE/CVF international conference on computer vision. 2019, pp. 6023–6032
19. Dan Hendrycks et al. “Augmix: A simple data processing method to improve robustness and uncertainty”. In: arXiv preprint arXiv:1912.02781 (2019)
20. Ekin D Cubuk et al. “Autoaugment: Learning augmentation policies from data”. In: arXiv preprint arXiv:1805.09501 (2018)
21. Ekin D Cubuk et al. “Randaugment: Practical automated data augmentation with a reduced search space”. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2020, pp. 702–703

Improved Alzheimer's Disease Detection with Dynamic Attention Guided Multi-modal Fusion

Arkaprabha Basu[1], Sourav Raha[1], Avisek Gupta[1], and Swagatam Das[2(✉)]

[1] Institute for Advancing Intelligence, TCG Crest, Kolkata, India
[2] Electronics and Communication Sciences Unit (ECSU), Indian Statistical Institute, Kolkata, India
swagatam.das@isical.ac.in

Abstract. The early detection of neurodegenerative disorders such as Alzheimer's disease is crucial to providing effective healthcare for management and recovery. We address the task of ternary classification of healthy, mild cognitive impairment, and Alzheimer's disease categories from multiple data modalities of 3D MRIs, patient electronic health records, and genetic information. For this task, we propose a Dynamic Attention Guided Multi-modal Fusion (DAGMF) approach, broadly consisting of three deep network components. The first component independently performs feature extraction for all modalities and refines them using novel Per-Modality Attention blocks. Thereafter, the obtained modality representations are provided to a proposed Dynamic Attention Multi-modal Solver block, which models the dynamics of attention across learning iterations by a Neural Ordinary Differential Equation (NODE) solver to generate modality attention. The modality representations and attention are finally provided to a novel Attention-induced Multi-modal Fusion block, which uses the attention to perform late-fusion of the multiple modality representations by a second NODE solver, which models dynamics of the various modalities across learning iterations. Empirical studies on multi-modal datasets constructed from the ADNI collection show that the proposed DAGMF method provides better classification performance than state-of-the-art multi-modal deep learning approaches.

Keywords: Alzheimer's disease · Multi-modality · Attention modules · Neural ODE · Deep Learning

1 Introduction

The intricate nature of Alzheimer's Disease (AD) necessitates a multifaceted approach for early and accurate diagnosis. The consideration of multiple data

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-78195-7_29.

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15311, pp. 432–446, 2025.
https://doi.org/10.1007/978-3-031-78195-7_29

modalities may provide a more holistic view of the disease's impact on the brain by directly modeling relevant structural changes, functional anomalies, and genetic inclination [11,19]. In this study, we examine the potential of the following three data modalities to aid in diagnosing Alzheimer's disease: 3D Magnetic Resonance Imaging (MRI), Electronic Health Records (EHRs), and genetic information through Single-Nucleotide Polymorphisms (SNPs). Each modality provides partial and complementary information about a patient. The 3D MRI neuroimaging provides spatial information on brain atrophy and metabolic activity that may be indicative of AD [32] for binary classification. In comparison, EHRs containing patients' medical histories capture patterns in symptoms and progression rates, while SNPs [10] offer insight into hereditary risk factors by examining potential biomarkers for AD. When building statistical models to diagnose Alzheimer's disease, building such models on complementary data modalities representing more comprehensive profiles of patients may lead to more accurate diagnoses.

The success of deep learning towards the statistical modeling of input signals that can be high-dimensional has encouraged recent investigations on the diagnosis of Alzheimer's disease from specific patient modalities, most commonly neuroimages such as MRI and Positron Emission Tomography (PET). Recent investigations have also achieved success when considering multiple patient modalities, where usually along with neuroimages other information of that particular patient are included that are deemed important, such as EHRs or SNPs [2,3,29]. Apart from including multiple modalities, approaches that have generally been observed to improve the performances of deep networks have been towards designing suitable attention mechanisms [28]. The design of appropriate attention modules has been observed to lead to better Alzheimer's classification [14,16,33]. Also, we note that an essential development in deep learning research has been the Neural Ordinary Differential Equations (NODEs) [6,22], where deep networks can be designed to model the dynamics of variables involved in a system of differential equations. NODEs have been used successfully across a variety of fields [6], other information of that particular patient is included that is deemed necessary, such as EHRs or SNPs [7,29].

The above developments in deep learning and Alzheimer's disease classification have motivated us to propose a Dynamic Attention Guided Multi-modal Fusion (DAGMF) approach for the diagnosis of Alzheimer's disease from multiple patient modalities, which can be observed in Figure 1. We consider three patient modalities of 3D MRI, EHR, and SNP, combined in a *late-fusion* manner only towards the end of the network, to provide the network sufficient opportunity to extract information from each modality independently. Therefore, initially, the modalities undergo feature extraction separately, followed by applying novel Per-Modality Attention (PMA) blocks proposed to refine their representations further. The PMA blocks employ tanh instead of ReLU activations [21] to minimize early information loss. Similarly, the feature extractors avoid ReLU and employ leaky ReLU and Parameterized ReLU (PReLU) [15] activations instead. The refined representations undergo linear projections and are then forwarded

along two branches. The first branch of Dynamic Attention Multi-modal Solver (DAMS) produces attention based on the modality attention dynamics observed across several learning iterations. A NODE is utilized here to learn attention dynamics. The second branch of Attention Induced Multi-modal Fusion (AIMF) performs late-fusion of the multiple modalities based on the modalities' dynamics and the a fused dynamic attention based DAIF's attention. A second NODE is used as the solver for this late-fusion task. The fused representation is finally provided to a classifier to perform ternary classification to Cognitively Normal (CN), Mild Cognitive Impairment (MCI), or Alzheimer's Disease (AD) classes.

Thus, our original contributions are summarized as the following:

1. We design novel PMA blocks utilizing tanh activations to provide early attention on extracted modality representations to avoid loss of information that can occur when using ReLU activations. To minimize early loss of information, all initial feature extractors use leaky ReLU and PReLU activations instead of ReLU.
2. We propose a new DAMS block to generate attention for each modality. DAMS accomplishes this using a NODE solver by modeling the dynamics of attention across learning iterations.
3. A novel multi-modal late-fusion approach AIMF is designed to fuse the multiple modality representations by a NODE solver, which also takes as input the attention generated by DAMS.
4. Through extensive experiments, we show that the proposed DAGMF approach achieves improved ternary classification performances when compared with the recent state-of-the-art.
5. We establish a framework for creating multi-modal training and testing data from well-curated sources that can be followed to easily reproduce our results. All data creation, model training, and experiment source codes are available at https://github.com/Thecoder1012/DAGMF.

2 Related Works

Recent developments that have significantly motivated the development of the proposed DAGMF approach are discussed next.

Multi-modal classification of Alzheimer's Disease. The developments in Alzheimer's disease classification on single data modalities [9] encouraged investigations towards working with multiple data modalities. The primary consideration when working with multi-modal data, is to design when to combine the information from the multiple modalities of a data instance before deciding its class. A wide variety of possible approaches have been explored. Some examples of using classical Machine Learning (ML) methods for multi-modal fusion of patient data include Linear Discriminant Analysis (LDA) [19], Gaussian Discriminant Component Analysis (GDCA) [11], and Support Vector Machines (SVM) Bucholc et al. [3]. We observe the use of Multivariate BiLSTMs in [2]. With the development of Deep Learning (DL) approaches for Alzheimer's disease classification, the need to develop reliable multi-modal fusion techniques arose. One

approach has been to use the robust Random Forests (RF) classifiers along with deep networks [10,29]. As access to more capable hardware became easier, multi-modal information could simply be concatenated [32] to retain all information. Fusion mechahnisms for MRI and PET are shown in [17,26], following which a concatenation based approach was shown by Shukla et al. [25], which was based on a 3D multi-scale feature aggregation mechanism for the MRI and PET modality representations. A variant proposed by Goel et al. [13] involved extracting wavelet features extracted from ResNet50 features for MRI and PET modalities which underwent a fusion mechanism. To further refine the information we have by generating attention for it, in Golovanevsky et al. [14] we note the use of attention for modality fusion. In Ying et al. [31], the use of finetuned deep networks for the fusion of multi-modal data can be noted. A recent approach of interest by Teaima et al. [27] showed that high classification performances could also be achieved by training an ensemble of ML methods.

Attention in deep learning for Alzheimer's Disease classification. In [14] we observe the proposal of cross-modality attention to examine connections between different modality features and generate attention accordingly to accomplish multi-modal Alzheimer's disease diagnosis. Their success motivated subsequent works on designing appropriate attention modules to improve classification performances. The Enhanced Non-Local Attention block to capture global information and the Coordinate Attention modules to capture local information was designed in [16]. IDANet [33] allows further customization of attention that can automatically scale and refine structural MRI patches. Other attention mechanisms, such as self-attention, have been explored in [4], while [12] generates attention at multiple scales to identify local variations and correlations across long ranges.

NODEs in Medical Diagnosis. Across several medical problem domains, we observe the recent use of NODEs to capture the complex dynamics of system variables. [6] uses NODEs to describe the dynamic course of Alzheimer's disease, using longitudinal neuroimaging to document the disease's developments. NODEs were used by [30] on multi-modal data for the survival analysis of breast cancer patients. We also observe NODEs to be successfully employed in deep networks for lung disease diagnosis [22] and cardiovascular risk prediction [18].

3 Proposed Method

For the classification of Alzheimer's disease from multi-modal data, our proposed DAGMF approach is illustrated in Figure 1. For each input data modality, representations are obtained from the Per-Modality Representation Refinement block (PMRR), which are forwarded along two branches. The first branch of Dynamic Attention Multi-modal Solver (DAMS) employs an ODE solver to generate per-modality attentions, while the second branch of Attention Induced Multi-modal Fusion (AIMF) uses an ODE solver to fuse the modalities. On the fused modality representations, the final classification is performed.

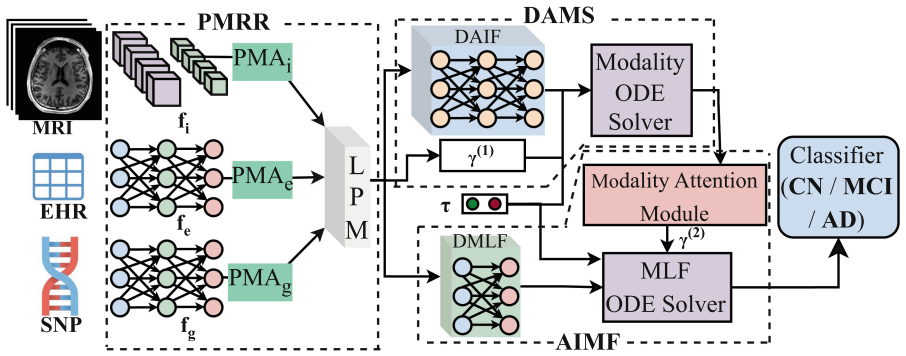


Fig. 1. The proposed DAGMF approach. In the initial PMRR block, three data modalities of 3D MRI, EHR, and SNP undergo feature extraction by networks f_i, f_e, and f_g, respectively, followed by targeted augmentation by Per Modality Attention (PMA) units. The obtained modality representations are passed through Linear Projection Modules (LPM), and forwarded along two branches. The first branch of DAMS integrates the modality representations by a Dynamic Attention Intermediate Fusion (DAIF) block, followed by a Modality ODE solver to generate the multi-modal attentions. This solver additionally requires an ODE state as input and a weight vector $\mathbf{w}$ which is also obtained from the LPM. In the second branch of AIMF, the modality representations are first refined and provided as input to a Multi-modal Late Fusion (MLF) ODE Solver. The MLF also takes as input an ODE state, and the output of the DAMS branch, which is provided additional attention by the Modality Attention Module. The output of MLF is finally provided to a ternary classifier to predict the CN/MCI/AD class.

3.1 Per-Modality Representation Refinement (PMRR)

Let us denote the three input data modalities as $X^{(i)}$, $X^{(e)}$, and $X^{(g)}$ for 3D MRI, EHR, and SNP respectively. Each modality is first propagated through separate feature extractor networks f_i, f_e, and f_g respectively, where f_i is a 3D Convolutional Neural Network (CNN), and f_e and f_g are 2D Multi-Layer Perceptrons (MLPs). This is followed by Per Modality Attention (PMA) blocks to selectively amplify modality components. The PMA blocks are two-layer MLPs constructed with two main motivations. The first motivation arises from the observation that in most attention blocks, ReLU activations are primarily used. ReLU sets to zero all negative preactivations, thereby preventing the related feature information from reaching later layers [8,20]. As the PMA blocks are used quite early in the proposed network, tanh activation is used in the first PMA layer to preserve more information that can be propagated to later layers. The second PMA layer uses a softmax activation to provide attention in the form of probabilities. Therefore, on each input modality, denoted in general as $X \in \mathbb{R}^{n \times d}$, with n number of features of dimension d, the mechanism producing attentions $\mathcal{A}(X)$ is,

$$\mathcal{A}(X) = \text{softmax}\left(\tanh(XW_1 + b_1)W_2 + b_2\right)^T X. \tag{1}$$

Here X is propagated through the first layer with parameters $W_1 \in \mathbb{R}^{d\times k}$, $b_1 \in \mathbb{R}^k$ and tanh activation, followed by the second layer with parameters $W_2 \in \mathbb{R}^{k\times 1}$, $b_2 \in \mathbb{R}$, and softmax activation. While the resulting $\mathcal{A}(X) \in R^{1\times d}$ provides scores for the feature dimensions, how the n features contributed to the scores is not evident. This provides the second motivation behind the design of the PMA block, where we associate the features with their attention scores to obtain the final attended features $\mathcal{P}$ as,

$$\mathcal{P} = \mathcal{A}(X) \odot X. \tag{2}$$

Here $\odot$ denotes the Hadamard product that is performed by broadcasting $\mathcal{A}(X)$ to the rows of X. In this manner, for all three modalities we obtain features $\mathcal{P}^{(i)}$, $\mathcal{P}^{(e)}$, and $\mathcal{P}^{(g)}$. These features are then passed through an MLP network called Linear Projection Module (LPM), to obtain per-modality features Z^i, Z^e, and Z^g, and are provided to the DAMS and AIMF branches.

3.2 Dynamic Attention Multi-modal Solver (DAMS)

To fuse the per-modality features, we consider the following approach involving a neural ODE solver which models the ongoing dynamics of the attention mechanism. Let f_A be the Dynamic Attention Intermediate Fusion (DAIF) block, which fuses the per-modality features from LPM as,

$$f_A(Z^{(i)}, Z^{(e)}, Z^{(g)}; W_A) = Z. \tag{3}$$

Here W_A denotes the parameters of the f_A network, and Z is the fused output of DAIF. The dynamics of the change of W_A over the number of iterations t of the learning algorithm is considered as,

$$\frac{d}{dt}W_A = f_A(W_A(t), t). \tag{4}$$

We use a Modality ODE Solver to approximate the solution of this differential equation. The solver is provided an initial $W_A(0)$ corresponding to iteration $t := 0$, and performs a numerical integration of the following up to $t := T$ to predict $W_A(t)$,

$$W_A(t) = W_A(0) + \int_0^T f_A(W_A(t), t)dt. \tag{5}$$

Additionally, here an ODE weight vector $\gamma^{(1)}(t)$ and an ODE time state $\tau(t)$ is required. The ODE weight vector is computed by an MLP from the output of LPM $\gamma^{(1)}(t) = \mathcal{N}_{\gamma^{(1)}}([Z^{(i)}, Z^{(e)}, Z^{(g)}])$. The overall operation of the Modality ODE Solver can be represented as a function g_A, which provides as output per-modality attentions $S_A(t)$, based on the predicted $W_A(t)$ and the fused representations Z,

$$S_A(t) = g_A(W_A(t), Z, \gamma^{(1)}(t), \tau(t)). \tag{6}$$

3.3 Attention Induced Multi-modal Fusion (AIMF)

The primary objective of the AIMF block is to use a neural ODE solver to fuse multi-modal representations based on the dynamic attention received from the DAMS block. Thus a Dynamic Modality Late Fusion (DMLF) block is considered, which consists of networks $f_M^{(i)}$, $f_M^{(e)}$, and $f_M^{(g)}$ to refine the representations obtained from LPM. Here for the first modality $Z^{(i)}$, $f_M^{(i)}$ with parameters $W_M^{(i)}$ provides,

$$\mathcal{D}^{(i)} = f_M^{(i)}(Z^{(i)}, W_M^{(i)}). \tag{7}$$

The definitions are similar for modalities $Z^{(e)}$ and $Z^{(g)}$. We note that unlike DAIF, DMLF does not fuse the multi-modal representations, and produces per-modality representations $\mathcal{D}^{(i)}$. Instead, the MLF ODE Solver is tasked with fusing the outputs from DMLF. The dynamics of DMLF are modeled as a differential equation similar to eqn. (4), and the MLF ODE solver predicts $W_M^{(i)}(t)$, $W_M^{(e)}(t)$, $W_M^{(g)}(t)$ by numerical integration, which for modality i is,

$$W_M^{(i)}(t) = W_M^{(i)}(0) + \int_0^T f_M(W_M^{(i)}(t), t)dt, \tag{8}$$

and is defined similarly for modalities e and g. Along with the initial $W_M(0)$, the solver requires an ODE time state $\tau(t)$ and an ODE weight vector $\gamma^{(2)}(t)$. This weight vector is obtained from the output $S_A(t)$ of the DAMS block as $\gamma^{(2)}(t) = \mathcal{N}_{\gamma^{(2)}}(S_A(t))$, where an MLP denoted as Modality Attention Module $\mathcal{N}_{\gamma^{(2)}}$ provides additional attention. The resulting fused output $S_M(t)$ from the MLF ODE Solver can be described in terms of a function g_M, which depends on the predicted weights of the DMLF network, and its output modality representations, thus,

$$S_M(t) = g_M([W_M^{(i)}(t), D^{(i)}], [W_M^{(e)}, D^{(e)}], [W_M^{(g)}, D^{(g)}], \gamma^{(2)}(t), \tau(t)). \tag{9}$$

3.4 Classification and Loss Function

The task of classification is carried out by considering the fused modality representations along with individual modalities. The fused $S_M(t)$ is propagated along a two-layer MLP $\mathcal{N}_c$ to obtain three-dimensional class probabilities $h(t) = \mathcal{N}_c(S_M(t))$. Additionally, modality representations from the initial feature extractors are also propagated along two-layer MLPs to obtain class-probabilities $h^{(i)}(X^{(i)}), h^{(e)}(X^{(e)}), h^{(g)}(X^{(g)})$ as,

$$h^{(i)} = \mathcal{N}_c^{(i)}(f'^{(i)}(X^{(i)})), \quad h^{(e)} = \mathcal{N}_c^{(e)}(f'^{(e)}(X^{(e)})), \quad h^{(g)} = \mathcal{N}_c^{(g)}(f'^{(g)}(X^{(g)})).$$

Each output is compared with ground-truth one-hot vector y in terms of cross-entropy loss $\ell(h, y) = -\sum_{j=1}^{3} y_j \log(h_j)$. The convex combination of the loss functions is carried out using non-negative weights β, $\beta^{(i)}$, $\beta^{(e)}$, and $\beta^{(g)}$, whose sum is one. Thus, the combined loss is defined as,

$$\mathcal{L} = \beta\ell(h(t), y) + \beta^{(i)}\ell(h^{(i)}, y) + \beta^{(e)}\ell(h^{(e)}, y) + \beta^{(g)}\ell(h^{(g)}, y). \tag{10}$$

The last three terms thus focus on directly refining modality representations. The first term guides the model to learn to fuse modalities while modeling modality and attention dynamics, so as to classify accurately.

4 Experiments and Results

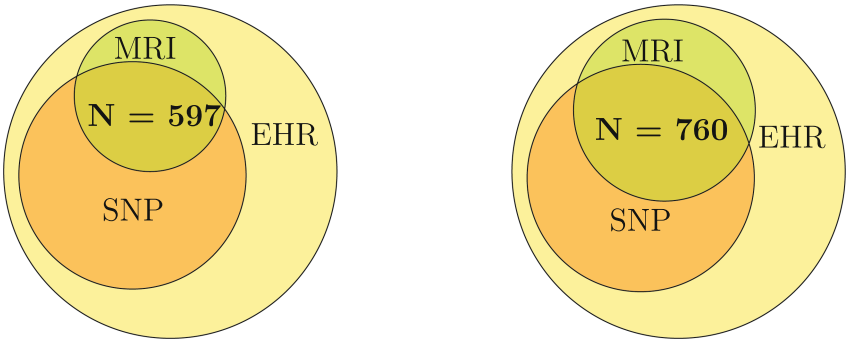
In this section, we evaluate the efficacy of the proposed DAGMF approach, by comparing their empirical performances with state-of-the-art approaches for Alzheimer's disease classification, on real-world data from well-established repositories of patients. The DAGMF network is created with the following feature extractors for each modality: f_i is a seven-layer 3D CNN, f_e is an eight-layer MLP, and f_g is a seven-layer MLP. For all feature extractors, the input layers have Leaky-ReLU activations, the output layers have PReLUs, and only certain hidden layers have ReLU [21] activations, which were determined empirically using validation data sets. The following experiment protocol was maintained. Datasets were split into 70%-30% training-test splits, and the models were trained under ten-fold cross-validation on the training split, using a weighted Adam optimizer for 65 iterations with learning rate 10^{-4}, batch size 32. The network architecture, training procedures, and performance evaluation are comprehensively provided in the supplementary document.

4.1 Creation of Multi-modal Datasets

The data for this study was obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) collection [23], which contains four datasets namely ADNI1, ADNI2, ADNI3, and ADNI-GO. From the available EHR data, for every patient we extracted 65 features that were deemed to have the potential to aid in the classification task. The SNP data for a subset of these patients was available; for them 500 random samples from six allele features were collected. Further, for a subset of these patients 3D MRI images were available. Figure 2 shows the two datasets that were created based on the availability of all modalities. We first created Dataset A containing data from 597 patients, which was later extended to form Dataset B containing data from 760 patients. All details on the procedure to reproduce the creation of both datasets are present in the supplementary material.

4.2 Empirical Evaluation

Here we report the classification performances that were achieved by the proposed DAGMF approach on Dataset A (hereafter referred to as DAGMF-A) and on Dataset B (referred to as DAGMF-B). The performances of DAGMF-A and DAGMF-B are compared with the performances reported by recent State-Of-The-Art (SOTA) multi-modal Alzheimer's disease classification methods on ADNI data. The SOTA methods investigated the performances possible when



(a) Dataset A (b) Dataset B

Fig. 2. Venn diagrams showing the intersection of the three main data modalities employed in our study for Dataset A and B. Based on the intersection of the more abundant EHR, comparatively less SNP, and the least available 3D MRI data, Dataset A was created with 597 patient data, and Dataset B with data from 760 patients.

considering the same modalities or even different ones, such as 2D MRI and PET images.

From the observed results noted in Table 1, we observe the highest test accuracies and F1-scores for DAGMF-A, closely followed by DAGMF-B. Among the competing methods, baseline performances are noted for the traditional approaches of LDA by Lin et al. [19], GDCA by Fang et al. [11], and the SVM-based fusion of Bucholc et al. [3]. Initial improvements over the baselines can be observed for the Multivariate BiLSTM by Abuhmed et al. [2], with further improvements noted for the finetuning (FT) of deep networks based approach of Ying et al. [31]. Among deep networks with RF-based fusion approaches, improved performances can be noted in Sappagh et al. [10] over that of Venugopalan et al. [29]. Among the modality concatenation-based approaches, all of which operate on 3D MRI and PET images, Song et al. [26] provides an initial baseline, with improvements noted in [13,17,32], and noteably in the approach of Shukla et al. [25]. One of the highest performances among recent methods are noted for Teaima et al. [27], which is closely followed in terms of the accuracy measure by the attention-based approach of Golovanevsky et al. [14]. Considering all recent and SOTA methods, the highest performances are observed to be achieved by the DAGMF approaches of DAGMF-A closely followed by DAGMF-B, indicating that modeling attention dynamics has contributed to improved classification performances. The efficacy of the proposed DAGMF approaches using NODE solver-based modeling of attention dynamics can be verified in terms of achieving the highest classification accuracies and F1-scores, among all results reported by recent and SOTA approaches. The higher performance of DAGMF-A indicates that the larger Dataset B contains problem complexities less captured in the smaller Dataset A, from which we conclude that future studies should also involve the creation of larger data sets.

Table 1. Comparison of Alzheimer's disease classification approaches between 14 recent and SOTA methods, and the proposed DAGMF networks. The proposed approaches are observed to outperform the SOTAs in terms of reported accuracies and F1-scores.

Method	Modalities					Accuracy(%)	Loss	F1-Score	Fusion Method
	2D-MRI	3D-MRI	EHR	GD	PET				
Lin et al. [19]	–	✓	–	✓	–	66.7	–	64.9	LDA
Fang et al. [11]	–	✓	–	–	✓	66.29	–	–	GDCA
Bucholc et al. [3]	✓	–	✓	–	✓	82.90	–	–	SVM
Abuhmed et al. [2]	✓	–	✓	–	✓	86.08	–	87.67	Multivariate BiLSTM
Ying et al. [31]	✓	–	–	✓	–	89.19	–	–	DL + FT
Venugopalan et al. [29]	–	✓	✓	–	–	78	–	78	DL + RF
Sappagh et al. [10]	–	–	✓	✓	–	92.62	–	92.56	DL + RF
Song et al. [26]	–	✓	–	–	✓	74.54	–	–	DL + Concat
Kong et al. [17]	–	✓	–	–	✓	87.67	–	–	DL + Concat
Zhang et al. [32]	–	✓	–	–	✓	95.68	–	–	DL + Concat
Goel et al. [13]	–	✓	–	–	✓	95.89	–	–	DL + Concat
Shukla et al. [25]	–	✓	–	–	✓	96.36	–	96.36	DL + Concat
Teaima et al. [27]	–	–	✓	✓	–	97.70	–	–	Ensemble ML
Golovanevsky et al. [14]	✓	–	✓	–	–	96.88	–	91.41	DL + Attention
DAGMF-A (Ours)	–	✓	✓	✓	–	**99.02** ±0.15	**0.1520** ±0.0124	**98.79**	**DL + DAGMF**
DAGMF-B (Ours)	–	✓	✓	✓	–	**98.27** ±0.18	**0.2147** ±0.0145	**97.31**	**DL + DAGMF**

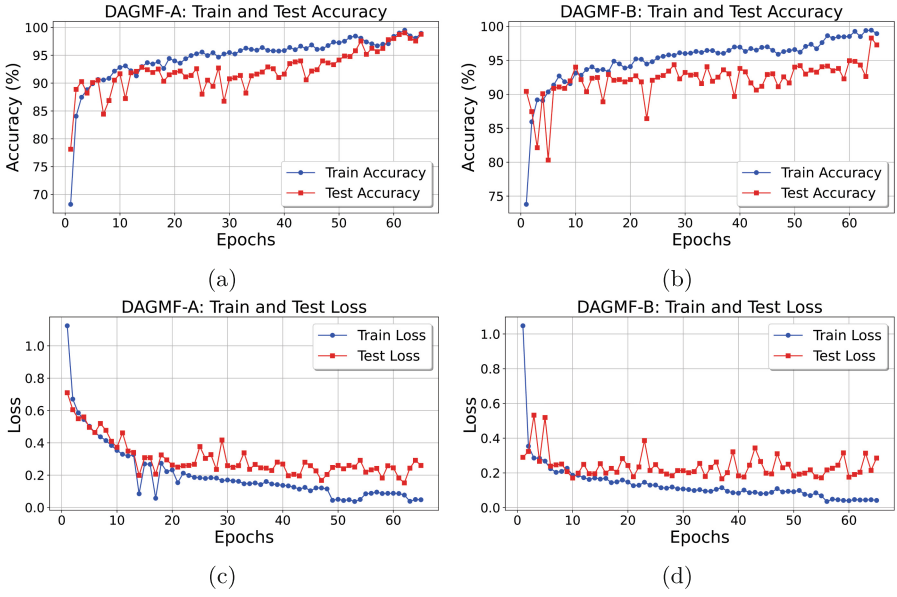


Fig. 3. Plots showing the changes in training and test accuracy with increasing number of epochs for (a) DAGMF-A and (b) DAGMF-B. The changes in training and test loss function values are also shown for (c) DAGMF-A and (d) DAGMF-B.

Therefore, DAGMF can be recommended for Alzheimer's disease classification, as it achieves the highest accuracies and F1-scores compared with the SOTA. To further examine the DAGMF approach, in Figure 3 we plot the changes in training and test accuracies achieved by DAGMF over training epochs, as well as the corresponding loss function values computed over training and test sets. For Dataset A, in Figure 3a we observe that with increasing epochs, the optimization of training loss results in the overall increase in training accuracy, and the eventual increase in test accuracy. In Figure 3c we also observe that optimizing the training loss results in its decrease, and also causes the loss on unseen test data to decrease. Similar observations can be made from Figures 3b and 3d for DAGMF on Dataset B. The changes in training and test F1-scores, Precision, and Recall, over the number of epochs exhibits a similar nature, and are provided in the supplementary document. Hence we conclude that the training procedure exhibits stability, while being reliable due to its resulting test set performances.

4.3 Ablation Study

In this section we perform ablation studies to verify that all proposed model components lead to improved classification performances, the results of which are shown in Table 2. The final DAGMF-A and DAGMF-B here are models V10 and V11 respectively. For feature extraction, we restate that the proposed feature extractors in V10 and V11 are $f := (f_e, f_g, f_i)$, where f_i is a seven-layer 3D CNN, f_e is an eight-layer MLP, and f_g is a seven-layer MLP, and PReLU and LeakyReLU activations are used. Now, we consider as a baseline the popular approach of using only ReLU-based feature extractors $f' := (f'_e, f'_g, f'_i)$ in models V1 to V7, which are trained on Dataset A. We also consider a network variant called DAGMF$_s$ in models V8 and V9 which has feature extractors $f'' := (f''_e, f''_g, f''_i)$ with fewer network parameters. Here f''_i is a four-layer 3D CNN, f''_e is a four-layer MLP, and f''_g is a six-layer MLP; we maintain the same usage of PReLU and LeakyReLU activations as in V10 and V11. Based on whether training and testing was done on Dataset A or B, we refer to model V8 as DAGMF$_s$-A, and similarly refer to model V9 as DAGMF$_s$-B. Further experiments on DAGMF$_s$ can be found in the supplementary material.

Other than the feature extractor, the following changes in model components are considered. In model V1, only cross-entropy loss ℓ is used instead of $\mathcal{L}$. In model V2, we remove both DAMS and AIMF blocks, and simply use an MLP to fuse the modality features and perform classification (NC'). Among variants of attention, the use of Multi-Head Attention (MHA) is introduced in model V3, followed by Modality Concatenation (MC). Cross-attention (CAttn) and Multi-Modal Fusion (MMF) were employed in V4, and in V5 PMA was introduced. In model V6, neural ODE was used only for modality fusion. In model V7, DAGMF was finally introduced to also model attention dynamics.

From the resulting accuracies in Table 2, we observe that models V1 to V7 attain lower classification accuracies compared to V10 (DAGMF-A) and V11 (DAGMF-B). Reducing the complexity of the feature extractor networks

in DAGMF$_s$-A and DAGMF$_s$-B has also led to a marginal reduction in accuracies. The highest classification accuracies were achieved by DAGMF-A and DAGMF-B, thus verifying the necessity of each network component that has been considered.

Table 2. Ablation study tracking performance improvements across *eleven* model iterations, while sequentially introducing the proposed components of DAGMF. Only ReLU-based feature extractors $f'{:=}(f'_e, f'_g, f'_i)$ in models V1 to V7 are compared with the proposed $f{:=}(f_e, f_g, f_i)$ in V10 and V11 using PReLU and LeakyReLU, as well as a variant with smaller feature extractors $f''{:=}(f''_e, f''_g, f''_i)$ in V8 and V9. As a variant of the classifier, NC' is considered. Variants in attention mechanism include MHA, PMA, and CAttn. The fusion strategies are MC, MMF, NODE, and the proposed DAGMF. For the loss, cross-entropy (ℓ) is compared with the proposed $\mathcal{L}$.

Model Version	Feature Extractors			Classifier		Modality Attentions			Fusion Strategies				Loss		Accuracy
	f'	f''	f	NC	NC'	MHA	PMA	CAttn	FC	MMF	NODE	DAGMF	ℓ	$\mathcal{L}$	
V1	✓			✓					✓				✓		91.82
V2	✓				✓									✓	89.29
V3	✓			✓		✓			✓					✓	92.95
V4	✓			✓			✓			✓				✓	94.83
V5	✓			✓			✓			✓				✓	95.16
V6	✓			✓			✓				✓			✓	96.23
V7	✓			✓			✓					✓		✓	98.42
V8 (DAGMF$_s$-A)		✓		✓			✓					✓		✓	98.93
V9 (DAGMF$_s$-B)		✓		✓			✓					✓		✓	97.35
V10 (DAGMF-A)		✓		✓			✓					✓		✓	**99.02**
V11 (DAGMF-B)		✓		✓			✓					✓		✓	**98.27**

5 Conclusion and Future Works

In this study, we proposed a new DAGMF approach to Alzheimer's disease classification. Three input data modalities are considered, on which early feature extraction is performed in a PMRR block using networks with leaky ReLU and PReLU activations instead of ReLU to avoid early loss of information. For similar reasons tanh activations are used in the proposed PMA blocks, which were designed to provide attention and refine modality representations. From the PMRR block, modality representations are provided to the DAMS block to generate modality attention based on the dynamics of attention across learning iterations. The outputs of the PMRR and DAMS are provided to the AIMF block to fuse the multiple modality representations, on which classification is performed. Experiments were conducted on datasets constructed from the ADNI collection, and the observed results show that the proposedd DAGMF network provides improved classification performances over those reported for the SOTAs.

These promising observations encourage further studies in Alzheimer's disease classification on multi-modal data with other modalities such as PET images

[1,32], as well as extensions towards Explainable AI [5,24] which should be explored under the guidance of expert clinicians to determine the commonalities and differences in decision variables compared to those that have been established in the medical literature. Finally, our observed success suggests the potential of viable investigations towards other possibly rarer neurodegenerative diseases.

Acknowledgements. We are grateful to the Machine Learning Research Group (MLRG), Electronics and Communication Sciences Unit, Indian Statistical Institute, Kolkata, for their unwavering support and valuable assistance towards the fulfillment of the work.

References

1. Abdelaziz, M., Elazab, A., Wang, T.: Alzheimer's disease diagnosis framework from incomplete multimodal data using convolutional neural networks. Journal of biomedical informatics p. 103863 (2021)
2. Abuhmed, T., El-Sappagh, S., Alonso, J.M.: Robust hybrid deep learning models for alzheimer's progression detection. Knowl.-Based Syst. **213**, 106688 (2021)
3. Bucholc, M., Ding, X., Wang, H., Glass, D.H., Wang, H., Prasad, G., Maguire, L.P., Bjourson, A.J., McClean, P.L., Todd, S., et al.: A practical computerized decision support system for predicting the severity of alzheimer's disease of an individual. Expert Syst. Appl. **130**, 157–171 (2019)
4. Carcagnì, P., Leo, M., Del Coco, M., Distante, C., De Salve, A.: Convolution Neural Networks and Self-Attention Learners for Alzheimer Dementia Diagnosis from Brain MRI. Sensors **23**(3) (2023)
5. Chataleff, R., Simonyan, K., Vedaldi, A.: Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 809–818 (2018)
6. Chen: Neural ordinary differential equations for disease progression modeling in alzheimer's disease. Nature Communications **11**(1), 456 (2020)
7. Cuomo, S., di Cola, V.S., Giampaolo, F., Rozza, G., Raissi, M., Piccialli, F.: Scientific machine learning through physics-informed neural networks: Where we are and what's next (2022)
8. Mohi ud din dar, G., Bhagat, A., Ansarullah, S.I., Othman, M.T.B., Hamid, Y., Alkahtani, H.K., Ullah, I., Hamam, H.: A novel framework for classification of different Alzheimer's disease stages using CNN model. Electronics **12**(2), 469 (2023)
9. Ebrahimi, A., Luo, S., Disease Neuroimaging Initiative, f.t.A.: Convolutional neural networks for Alzheimer's disease detection on MRI images. Journal of Medical Imaging **8**(2), 024503–024503 (2021)
10. El-Sappagh, S., Abuhmed, T., Islam, S.R., Kwak, K.S.: Multimodal multitask deep learning model for alzheimer's disease progression detection based on time series data. Neurocomputing **412**, 197–215 (2020)
11. Fang, C., Li, C., Forouzannezhad, P., Cabrerizo, M., Curiel, R.E., Loewenstein, D., Duara, R., Adjouadi, M., Initiative, A.D.N., et al.: Gaussian discriminative component analysis for early detection of alzheimer's disease: A supervised dimensionality reduction algorithm. J. Neurosci. Methods **344**, 108856 (2020)

12. Gao, X., Cai, H., Liu, M.: A hybrid multi-scale attention convolution and aging transformer network for alzheimer's disease diagnosis. IEEE J. Biomed. Health Inform. **27**(7), 3292–3301 (2023)
13. Goel, T., Sharma, R., Tanveer, M., Suganthan, P., Maji, K., Pilli, R.: Multimodal neuroimaging based Alzheimer's disease diagnosis using evolutionary RVFL classifier. IEEE Journal of Biomedical and Health Informatics (2023)
14. Golovanevsky, M., Eickhoff, C., Singh, R.: Multimodal attention-based deep learning for alzheimer's disease diagnosis. J. Am. Med. Inform. Assoc. **29**(12), 2014–2022 (2022)
15. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1026–1034 (2015)
16. Illakiya, T., Ramamurthy, K., Siddharth, M.V., Mishra, R., Udainiya, A.: AHANet: Adaptive Hybrid Attention Network for Alzheimer's Disease Classification Using Brain Magnetic Resonance Imaging. Bioengineering **10**(6) (2023)
17. Kong, Z., Zhang, M., Zhu, W., Yi, Y., Wang, T., Zhang, B.: Multi-modal data alzheimer's disease detection based on 3d convolution. Biomed. Signal Process. Control **75**, 103565 (2022)
18. Lee: Neural ordinary differential equations for respiratory signal analysis and lung disease diagnosis. IEEE Transactions on Biomedical Engineering **68**(7), 2153–2165 (2021)
19. Lin, W., Gao, Q., Du, M., Chen, W., Tong, T.: Multiclass diagnosis of stages of alzheimer's disease using linear discriminant analysis scoring for multimodal data. Comput. Biol. Med. **134**, 104478 (2021)
20. Lu, L.: Dying ReLU and Initialization: Theory and Numerical Examples. Communications in Computational Physics **28**(5), 1671–1706 (2020)
21. Nair, V., Hinton, G.E.: Rectified linear units improve neural network acoustic models. In: Proc. 27th International Conference on Machine Learning (ICML). pp. 807–814 (2010)
22. Pegolotti, L., Pfaller, M.R., Rubio, N.L., Ding, K., Brugarolas Brufau, R., Darve, E., Marsden, A.L.: Learning reduced-order models for cardiovascular simulations with graph neural networks. Comput. Biol. Med. **168**, 107676 (2024)
23. Petersen, R.C., Aisen, P.S., Beckett, L.A., Donohue, M.C., Gamst, A.C., Harvey, D.J., Jack, C., Jr., Jagust, W.J., Shaw, L.M., Toga, A.W., et al.: Alzheimer's disease Neuroimaging Initiative (ADNI) clinical characterization. Neurology **74**(3), 201–209 (2010)
24. Selvaraju, R., Vig, A., Parikh, A.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Weighted Class Activation Maps. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. p. 618–626 (2017)
25. Shukla, A., Tiwari, R., Tiwari, S.: Alzheimer's disease detection from fused pet and mri modalities using an ensemble classifier. Machine Learning and Knowledge Extraction **5**(2), 512–538 (2023)
26. Song, J., Zheng, J., Li, P., Lu, X., Zhu, G., Shen, P.: An effective multimodal image fusion method using mri and pet for alzheimer's disease diagnosis. Frontiers in digital health **3**, 637386 (2021)
27. Teaima, A.S., Wahed, M.A., Sami, S.A.: Multimodal-based classification paradigm for alzheimer disease. 2023 5th Novel Intelligent and Leading Emerging Sciences Conference (NILES) pp. 430–433 (2023)
28. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Polosukhin, I., Kaiser, L., Polosukhin, I., Kaiser, L.: Attention is all you need. In: Advances in neural information processing systems. pp. 599–609 (2017)

29. Venugopalan, J., Tong, L., Hassanzadeh, H.R., Wang, M.D.: Multimodal deep learning models for early detection of alzheimer's disease stage. Sci. Rep. **11**(1), 3254 (2021)
30. Xiang, J., Qi, B., Cerou, M., Zhao, W., Tang, Q.: Data-driven neural-ode modeling for breast cancer tumor dynamics and progression-free survival predictions. In: The Symbiosis of Deep Learning and Differential Equations III (2023)
31. Ying, Q., Xing, X., Liu, L., Lin, A.L., Jacobs, N., Liang, G.: Multi-modal data analysis for alzheimer's disease diagnosis: An ensemble model using imagery and genetic features. In: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). pp. 3586–3591 (2021)
32. Zhang, Y., Wang, S., Xia, K., Jiang, Y., Qian, P.: Alzheimer's disease multiclass diagnosis via multimodal neuroimaging embedding feature selection and fusion. Inf. Fusion **66**, 170–183 (2021)
33. Zhao, Q., Huang, G., Xu, P., Chen, Z., Li, W., Yuan, X., Zhong, G., Pun, C.M., Huang, Z.: IDA-Net: Inheritable Deformable Attention Network of structural MRI for Alzheimer's Disease Diagnosis. Biomed. Signal Process. Control **84**, 104787 (2023)

Transformers and CNNs in Neurodiagnostics: Handwriting Analysis for Alzheimer's Diagnosis

Gabriele Lozupone, Emanuele Nardone, Cesare Davide Pace, and Tiziana D'Alessandro(✉)

Department of Electrical and Information Engineering (DIEI), University of Cassino and Southern Lazio, Via G. Di Biasio 43, 03043 Cassino, FR, Italy
{gabriele.lozupone,emanuele.nardone,cesaredavide.pace, tiziana.dalessandro}@unicas.it

Abstract. Early diagnosis of Alzheimer's disease (AD) is critical for timely intervention and management. Handwriting analysis has been recognized as a promising diagnostic tool, as it is one of the first skills affected by AD. This research investigates the potential of using deep learning to assist AD diagnosis through handwriting image analysis. Unlike previous methods applied on this dataset, the model is trained on images of different handwriting tasks instead of training task-specific models. This approach improved the model's ability to identify AD-specific writing patterns and enhanced diagnostic performance. We conducted an extensive analysis of various convolutional neural networks and vision transformers, focusing on the ability of transformer models to transfer knowledge across different domains. Our contributions are as follows: (i) we conducted comprehensive experiments to identify the most effective deep learning models for the AD handwriting task, providing a benchmark for future research; (ii) we redesigned a pre-trained Optical Character Recognition (OCR) transformer model to assess its capability to transfer knowledge between OCR and classification tasks; and (iii) we developed a three-stage framework to increase performance and evaluate the impact of distinct handwriting tasks on the final AD diagnostic outcome. Our framework surpasses previous methodologies, achieving an accuracy of 87.99% and a sensitivity of 89.69%, demonstrating that handwriting-based deep learning-aided diagnostic systems hold significant promise as tools for early AD detection.

1 Introduction

The incidence of neurodegenerative diseases (NDs) has been steadily increasing in recent years, reflecting the complex challenges posed by the ageing population and changing lifestyles. This trend enhances the pressing need for continued research, increased awareness, and improved support systems. NDs, like Alzheimer's disease (AD), bring cognitive impairments such as memory loss and language difficulties and also affect motor skills. Early diagnosis is crucial for

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15311, pp. 447–463, 2025.
https://doi.org/10.1007/978-3-031-78195-7_30

effective intervention, yet AD currently lacks a cure. Research remains crucial in determining effective treatments, underscoring the significance of timely action.

The scientific community seeks innovative detection methods to identify the disease in its nascent stages. Among these, handwriting analysis has emerged as a promising avenue, offering insights into the early neurological changes characteristic of AD. Handwriting is one of the first skills to be altered by the onset of NDs symptoms [16,18,24,28] as it requires fine motor control, and its alterations can serve as early indicators. While there is a growing interest in utilizing handwriting analysis to support ND diagnosis [9,29], most research efforts have been centered around Parkinson's disease (PD), for which reference datasets have been collected over the years [12]. Concerning AD, the ongoing research usually focuses on privately collected datasets. Unbalanced and small datasets may make applying artificial intelligence (AI) techniques effectively and drawing comprehensive conclusions difficult. Many papers rely on handwriting tasks typically conducted on tablets, from which the authors extract features and evaluate the ability of several machine learning (ML) algorithms [17,25] or make a statistical analysis [30] to support the AD diagnosis. In contrast, this study is centered on images, delegating to architectures the task of focusing on automatically extracted features. Furthermore, this research proposes an approach based on offline images obtained by scanning the actual sheets of paper on which the task was performed, easing the data acquisition step.

Considering these factors, various handwriting tasks have been proposed to highlight the potential impairment in different abilities due to AD. In previous studies, we introduced a handwriting protocol to analyze the impact of various cognitive, memory, and motor skills on the performance of individuals with AD and healthy controls (HC) [3]. These handwriting tests were conducted with a tool to capture information on handwriting dynamics and store the paper sheets with the task performed. Different features were extracted after the acquisition step, and many classification schemes were implemented. It is worth noticing that most of our efforts were put into evaluating the system, considering one task at a time. After obtaining the performance for each task, we applied combining rules to subsets of tasks.

We initially utilized handwriting features and ML, in [4,7] we examined a subset of the dataset, using part of the total samples and tasks acquired. We computed static and dynamic features, applying ML algorithms for task-specific classification to distinguish between HC and AD. In [2], we considered all the tasks and their features, in-air, and on-paper, and we combined the predictions from a subset of tasks and features selected through a genetic algorithm. Another research path, exploited in [5,6], focused on generating synthetic images and the classification through Convolutional Neural Networks (CNNs), considering some acquired tasks. The study proposed in [8] shows the power of considering offline images to support AD diagnosis.

Moving from the past activity, this paper presents a deep learning approach to analyzing handwriting to support AD detection. The models are trained on images from all handwriting tasks, intending to capture the broader patterns

indicative of AD. Previous works exploited handwriting analysis task by task, each evaluated by its dedicated model. While these methods provided valuable insights, they inherently limited the scope of pattern recognition to within-task variations. Recognizing this limitation, the current study seeks to consolidate these tasks into a single dataset for a comprehensive analysis using a unique model. This approach is predicated on the hypothesis that a model trained on a more extensive and varied dataset can learn significant patterns attributable to AD, which may remain obscured in a more fragmented analysis framework.

The main contributions of this work are summarized as follows:

1. An in-depth comparison of different CNNs and Vision Transformers to identify the most effective models for the AD handwriting task, providing a benchmark for future research.
2. Investigation of a pre-trained Optical Character Recognition (OCR) transformer model to verify its capability to transfer knowledge between OCR and classification tasks.
3. Development of an experimental three-stage framework to improve diagnostic performance and evaluate the impact of distinct handwriting tasks on the final diagnostic outcome, thereby identifying the most informative tasks for AD detection.

The remainder of this work is organized as follows: Section 2 details the dataset and the deep architectures considered; Section 3 describes the experimental approach developed; whereas Section 4 reports and discusses the experimental results. Finally, Section 5 and Section 6 are devoted to the discussion and conclusions of the result, summarizing eventual future work inspired by the performance obtained in this study.

2 Materials and Methods

This section outlines the materials required for the proposed study, including the data and DL architectures employed. Section 2.1 delineates the data acquisition procedure and the processes done to obtain the offline images of handwriting samples. We selected this type of image to represent the handwriting because the preliminary results of our past activity [8] enhanced the advantages of using such a kind of data. Finally, Section 2.2 details the architectures evaluated.

2.1 Dataset

The rationale behind considering handwriting data to support AD diagnosis is exhaustively depicted in [9]. Data were acquired considering an experimental protocol [3] comprising 25 handwriting tasks, designed with the help of physicians to evaluate different abilities that could be affected by AD symptoms. The selected tasks belong to four categories: graphic, copy and reverse copy, memory, and dictation. Table 1 lists the performed tasks, comprising their enumeration, a

Table 1. List of tasks performed. Task categories are memory and dictation (M), graphic (G), and copy (C).

#	Description	Category
1	Signature drawing	M
2	Join two points with a horizontal line, continuously for four times	G
3	Join two points with a vertical line, continuously for four times	G
4	Retrace a circle (6 cm of diameter) continuously for four times	G
5	Retrace a circle (3 cm of diameter) continuously for four times	G
6	Copy the letters 'l', 'm' and 'p'	C
7	Copy the letters on the adjacent rows	C
8	Write recursively a sequence of four lowercase letters 'l'	C
9	Write recursively a sequence of four lowercase cursive bigram 'le'	C
10	Copy the word "foglio"	C
11	Copy the word "foglio" above a line	C
12	Copy the word "mamma"	C
13	Copy the word "mamma" above a line	C
14	Memorize the words "telefono", "cane", and "negozio" and rewrite them	M
15	Copy in reverse the word "bottiglia"	C
16	Copy in reverse the word "casa"	C
17	Copy words (regular/non-regular/non-words) in boxes	C
18	Write the name of the object shown in a picture (a chair)	M
19	Copy the fields of a postal order	C
20	Write a simple sentence under dictation	M
21	Retrace a complex form	G
22	Copy a telephone number	C
23	Write a telephone number under dictation	M
24	Clock Drawing Test	G
25	Copy a paragraph	C

description, and the belonging category. The protocol was administered following strict recruiting criteria, with the support of the geriatric ward's Alzheimer unit at the "Federico II" hospital in Naples. In detail, participants underwent clinical assessment and standard cognitive evaluations, such as the MMSE [13], FAB [15], and MoCA [23]. Participants in the study were carefully selected to ensure demographic and educational characteristics between the HC and AD groups were matched, as reported in Table 2. Individuals taking psychotropic drugs or substances that could influence cognitive abilities were excluded from

Table 2. Average demographic data of participants. Standard deviations are shown in parentheses.

	Age	Education	#Women	#Men
AD	71.5 (9.5)	10.8 (5.1)	46	44
HC	68.9 (12)	12.9 (4.4)	51	39

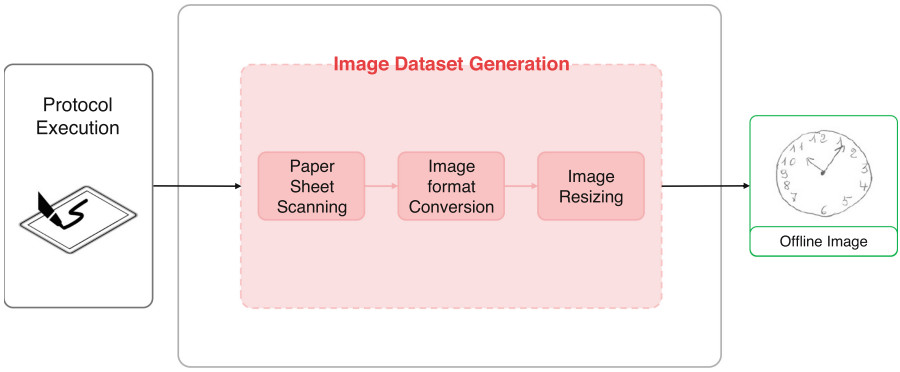


Fig. 1. Workflow diagram of offline image dataset generation process.

both groups. The final dataset consisted of 174 individuals, with 89 diagnosed with AD and 85 HC.

Handwriting samples were acquired using a WACOM Bamboo Folio graphic tablet, which allowed participants to write on standard A4 paper sheets fastened to the tablet's surface. The sensor-equipped pen recorded spatial coordinates (x, y), pressure (z), and timestamps at $200Hz$, capturing in-air movements within $3cm$ from the tablet surface. However, for this research, only offline images obtained from the paper sheets were utilized, disregarding the dynamic aspects of handwriting. Figure 1 shows the generation process of offline images. After the protocol execution, the paper sheets were scanned and saved as *.tif* files, with each frame representing a task. Thus, each frame was extracted and underwent a segmentation algorithm to isolate the participant's handwritten trace and saved as a *.png* file. The resulting images were resized to 299×299 pixels while ensuring the trace remained centred, minimizing information loss. The digitized images accurately reflect the participant's handwritten trace, with pixel values representing the natural grayscale shades of the ink on the paper, influenced by both the applied pressure and the dynamics of the movements. It is worth noting that each of the 179 participants executed each task only once, yielding in a dataset of 4475 images, 179 for each of the 25 tasks.

2.2 ImageNet-Pretrained Networks

This work verified and compared the ability of CNNs and Vision Transformers to learn AD writing patterns. This section briefly discusses the architecture families chosen for the analysis and their peculiarities. All the architectures described in this section employed transfer learning from the ImageNet dataset [10]. The **VGG** architecture family, developed in [26], represents a significant advance in deep learning for image recognition tasks. Characterized by its simplicity and depth, the architecture employs 3x3 convolutions, evenly stacked, to increase the depth of the network without complicating its structure. The **ResNet** [14]

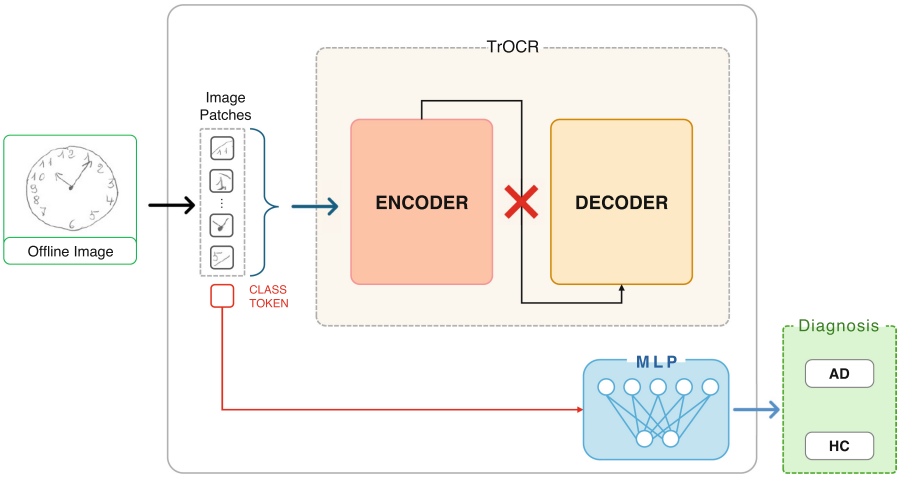


Fig. 2. Modified TrOCR architecture scheme. Encoder retained for feature extraction, decoder removed, class token and MLP added for classification.

family, inspired by VGG, introduced the residual block that enables learning the residual function for identity mapping that mitigates the vanishing gradient problem. The **EfficientNetV2** [27] architecture family pushes the boundaries of efficiency through innovative scaling strategies. By intelligently scaling model dimensions, they achieve remarkable accuracy and training speed enhancements, redefining the balance between performance and resource utilization. Transitioning from conventional convolutional designs, **ConvNeXt** introduces a paradigm shift inspired by Vision Transformers. In [21], the authors modified the ResNet architecture. They applied a macro design consisting of changes in the number of layers in each block. They patchified the input image with learnable convolution blocks characterized by stride increasing to simulate the ViT patch embedding [11]. These architectural choices show that purely convolutional architecture can compete with state-of-the-art Vision Transformers. The **ViT** [11] adapts the original Transformer architecture from Natural Language Processing to computer vision. The Transformer works with an input that consists of a sequence of words (or tokens). The authors of ViT generate a sequence by splitting the input image into non-overlapping patches with a fixed size of 16×16 pixels. Each patch is linearly projected into a fixed-sized space, and a class token is added to the sequence of embeddings. The class token enables the condensation of useful classification features through the self-attention mechanism and the direct connection with the head. The **Swin Transformer** family was proposed in [20], employing hierarchical structures and shifted window mechanisms. These architectural innovations enhance scalability and efficiency across a broad spectrum of vision tasks, demonstrating the potential of Transformer-based models in computer vision applications.

2.3 From OCR to Handwriting Alzheimer's Diagnosis

The chosen architecture is **TrOCR** presented in [19] and explicitly designed for OCR tasks. The motivation behind selecting TrOCR for this application stems from its performance, particularly on the Handwriting Dataset IAM[22]. Furthermore, its structure is prone to transfer learning for image classification tasks. TrOCR comprises two main components: the Transformer encoder and the Transformer decoder. The **Transformer Encoder** is responsible for processing the input image. As in ViT, the encoder processes a sequence of tokens representing patch embeddings extracted from the handwritten text image. The **Transformer Decoder** takes the output from the encoder and generates the text one character or token at a time. Therefore, the decoder part is unnecessary for a classification task, as in our case. As shown in Figure 2, we converted it to a ViT-like model to make the encoder usable for a classification task. The decoder part was detached because we only wanted to use the model's ability to produce feature representations from images. To enable classification, we adopted the same strategy as ViT. We added a class token to the input token sequence and attached an MLP responsible for the diagnosis part.

3 Experimental Approach

In this work, we developed a structured three-stage approach to evaluate the efficacy of deep learning models in identifying AD from task-specific text images derived from handwriting. The emphasis was placed on ensuring the comparability and reliability of the chosen convolutional and vision transformer models through a meticulous cross-validation process. Figure 3 presents an overview of the complete experimental framework, comprising three main phases: the First Stage, the Evaluation Stage, and the Combining Rule. Each of these phases is elaborated upon in detail within this section.

First stage: Model Training with Cross-validation The First Stage involved training the selected DL models on the entire dataset, which included offline images of different subjects engaged in specific writing tasks designed to reveal the presence of AD. The goal was to refine the models to produce robust feature representations that accurately mirrored AD-related patterns in handwriting. To enable a fair comparison between the models and enhance their ability to generalize, we implemented a 5-fold cross-validation technique. This approach entailed dividing the dataset into five distinct segments, ensuring each segment acted as a test set in one iteration and as part of the training/validation sets in the others. To prevent data leakage and preserve the integrity of our evaluation, we carefully assigned all writing tasks from the same subjects exclusively to one of the training, validation, or test sets. Thus, no subject's data was shared across the sets. The data was subjected to an 80-20 split within each fold for training/validation and testing, respectively. The 80% portion allocated for training/validation was divided using an 80-20 ratio. Consequently, 64% of the

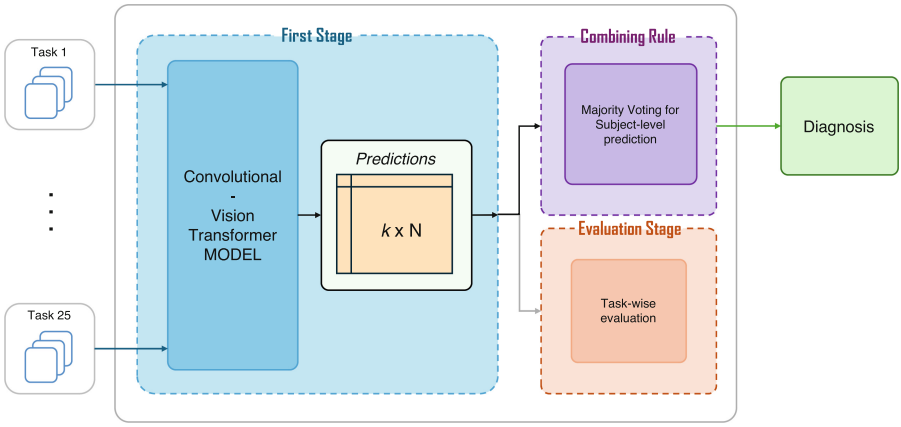


Fig. 3. Overview of the Experimental Approach for Diagnosing AD Using Deep Learning Models on Handwriting Images. This figure illustrates the three-step process, starting with the training of various models using a 5-fold cross-validation method (First Stage), followed by the evaluation of task-specific accuracy (Evaluation Stage), and concluding with the aggregation of task-specific outcomes for individual diagnoses through the majority voting (Combining Rule).

total dataset was used for training purposes, 16% for validation, and 20% for testing. This structured division ensured that the generalization ability of each model was thoroughly assessed.

Evaluation Stage: Task-wise Evaluation In the second phase, we evaluated the models trained in the first stage, emphasising task-specific predictions. We established a task-wise evaluation by structuring the predictions in a matrix format, with subjects delineated by rows and tasks by columns. This method facilitated the calculation of metrics for each task to the ground truth. Such an evaluation not only quantified the contributory value of individual tasks toward AD diagnosis but also allowed for an assessment of metrics uniformity across tasks. From this analysis, we derived two critical aspects: the unique contribution of each task to the AD diagnostic process and potential insights into whether combining tasks could lead to more accurate and robust AD diagnoses.

Combining rule: Majority Voting for subject-level prediction The final step involved aggregating the results from the individual tasks based on the test set predictions to produce a final diagnosis for each subject. By utilizing all the information available from all the tasks, we maximized the diagnostic accuracy of our framework. This integrated approach ensured that our conclusions were as comprehensive as possible, drawing on the full range of data available for each subject. It was important to note that steps 2 and 3 relied exclusively on the test set portions of each cross-validation fold, ensuring the model's ability to generalize to new, unseen data.

3.1 Experimental settings

Hyper-parameters We employed an AdamW optimizer with weight decay 10^{-2} and used a batch size 32. The images are resized to 224×224 to enhance transfer learning. We chose a learning rate of 10^{-3} and trained all the models for 100 epochs, using an early stopping strategy with the patience of 15 epochs. The loss function used was the binary cross entropy.

Performance Evaluation This study uses Accuracy, Sensitivity, Specificity, and Matthews Correlation Coefficient (MCC) as performance metrics. Accuracy is a classifier's most common performance measure, consisting of the percentage of correctly classified samples over the total. MCC [1] is a correlation coefficient between prediction and true label; it provides a more informative and truthful score than accuracy when evaluating binary classifications, allowing a more realistic interpretation of classifier performance, especially in the case of unbalanced datasets. Sensitivity and specificity serve as critical metrics for evaluating the diagnostic accuracy of our model, highlighting its ability to correctly identify AD cases and HC, respectively.

4 Results

In Table 3, we provide the results of the first stage of the experiment study. It can be observed that the architectures used reach an average Accuracy of approximately 74%, with the metrics calculated as the mean of a 5-fold cross-validation. Specifically, ConvNextSmall reaches 79.21% Accuracy and 82.32% Sensitivity. On the other hand, the VGG19 achieves 78.84% of Specificity, and ConvNextTiny obtains 57.88% of MCC. This achievement indicates that the architectures can discern the handwriting pattern of subjects with AD from those of HC.

Figure 4 and Figure 5 provide an analysis of the performance across the pool of tasks for CNN and Vision Transformer architectures, respectively. We can observe that in both graphs, there is a large variability in performance in terms of accuracy depending on the task. For instance, Tasks 8 and 9, which have higher performance, independently of the family of the architecture used, have higher accuracy and demand higher levels of cognitive processing and motor coordination. Instead, Tasks 1, 7, 20, and 25, which require minimal cognitive effort and fine motor skills, demonstrated significantly lower performance. These likely offer fewer distinctive features for the architectures to learn from, leading to a diminished capacity to differentiate between AD and HC. Table 4 illustrates the performance achieved by the architecture for each designated task, highlighting a distinct prevalence of ConvNext over other architectures. From the results, we see that ConvNextSmall reached 86.19% accuracy and 72.20% of MCC in task 9, and 86.69% specificity in task 10, while ConvNexTiny reached 89.69% of sensitivity in task 14.

Table 3. CNN and Transformer models performance on the first stage.

Model	Accuracy	Specificity	Sensitivity	MCC
VGG16	74.56 (4.35)	74.61 (9.34)	73.78 (5.76)	48.59 (9.70)
VGG19	75.48 (3.41)	**78.84 (7.18)**	71.83 (6.14)	50.90 (7.34)
ResNet50	70.74 (2.29)	78.04 (8.35)	64.08 (6.18)	42.72 (5.61)
ResNet101	71.45 (3.39)	77.46 (6.80)	65.35 (5.99)	43.19 (7.35)
EfficientNetV2M	72.50 (4.69)	74.35 (9.01)	70.28 (6.88)	44.86 (10.00)
EfficientNetV2S	73.05 (3.22)	77.41 (9.46)	68.52 (6.45)	46.43 (7.62)
ConvNextTiny	78.83 (2.65)	77.96 (7.18)	79.41 (8.57)	**57.88 (5.32)**
ConvNextSmall	**79.21 (3.39)**	74.66 (8.38)	**82.32 (7.47)**	57.82 (7.64)
ConvNextBase	78.03 (4.55)	76.80 (10.88)	78.01 (8.00)	55.42 (10.12)
VitB16	70.10 (3.82)	73.85 (9.60)	66.47 (9.53)	40.88 (7.95)
SwinV2B	77.61 (2.75)	79.62 (5.99)	75.70 (8.85)	55.69 (5.23)
SwinV2S	76.86 (2.95)	74.66 (11.44)	78.68 (9.18)	54.36 (5.34)
SwinV2T	77.13 (2.31)	74.92 (5.61)	78.44 (6.95)	53.82 (5.00)
TrOCR-Small-Hand	67.53 (2.73)	76.29 (6.01)	59.57 (5.22)	36.36 (5.20)
TrOCR-Base-Hand	66.02 (3.04)	77.98 (10.85)	55.47 (12.62)	35.10 (5.32)

Table 5, obtained through majority voting rule, shows that ConvNextSmall has the best overall performance with an Accuracy of 87.99% and MCC of 76.75%, showcasing a remarkably high Sensitivity of 89.69%, indicating its superior ability to identify AD subjects. Its performance is balanced, excelling in both sensitivity and MCC. This suggests that it can effectively handle both AD and HC classifications. ResNet50 exhibits the highest Specificity of 92.06%, indicating higher capability of identifying HC subjects. However, its lower Sensitivity 67.97%, and MCC 62.09% suggest that it may not be as effective at identifying AD subjects as other models. From the Vision Transformer architecture family, the SwinV2T performs well with an Accuracy of 85.54% and MCC of 71.53%. These models show a good balance between Sensitivity and Specificity, indicating their robustness in diagnosis but worse than ConvNextSmall. Variability across models is notable, with standard deviations indicating that some models (e.g., ConvNextBase, EfficientNetV2M) exhibit more variation in their performance metrics across folds. This variability might be due to differences in how models handle the intrinsic complexities and variations within the AD vs HC classification task. The TrOCR models show lower performance than other models in Table 5. Transfer learning from ImageNet has proven to be more effective. It can be assumed that for this specific analysis, a transfer of knowledge from task domains, e.g. classification task, is more valuable rather than input data domain, e.g. handwritten text images.

Table 4. Top model performance by task according to accuracy.

Task	Best Model	Accuracy	Specificity	Sensitivity	MCC
1	ConvNextBase	75.29 (7.57)	72.67 (17.78)	76.59 (11.94)	51.02 (15.09)
2	SwinV2S	82.55 (9.38)	78.96 (10.22)	85.83 (8.66)	65.04 (18.50)
3	ConvNextSmall	83.78 (6.43)	80.44 (7.64)	86.21 (9.11)	67.23 (13.32)
4	ConvNextSmall	83.17 (8.59)	79.65 (14.71)	85.91 (13.69)	66.37 (17.83)
5	SwinV2B	84.37 (7.44)	83.31 (10.30)	85.19 (8.34)	68.90 (15.55)
6	ConvNextTiny	81.94 (4.70)	80.13 (10.35)	82.89 (11.77)	64.16 (9.56)
7	ConvNextBase	70.48 (3.94)	71.77 (16.21)	67.43 (16.11)	40.63 (9.15)
8	ConvNextTiny	83.19 (6.68)	86.10 (9.46)	80.76 (13.41)	67.35 (12.65)
9	ConvNextSmall	**86.19 (6.47)**	85.79 (5.39)	86.32 (8.92)	**72.20 (13.05)**
10	ConvNextSmall	82.55 (3.81)	**86.69 (6.20)**	78.25 (8.98)	65.45 (7.51)
11	ConvNextTiny	82.57 (6.02)	85.23 (10.31)	79.58 (12.07)	65.54 (12.10)
12	ConvNextTiny	81.39 (8.11)	83.45 (8.60)	79.42 (21.25)	64.52 (14.72)
13	ConvNextTiny	81.39 (6.54)	81.09 (6.23)	81.94 (13.83)	63.38 (12.61)
14	ConvNextSmall	79.50 (5.05)	66.51 (19.05)	**89.69 (8.66)**	59.63 (11.46)
15	ConvNextBase	82.57 (7.70)	74.54 (14.57)	88.62 (8.97)	64.83 (17.03)
16	ConvNextBase	82.00 (7.68)	78.31 (4.73)	85.85 (13.13)	64.12 (15.26)
17	VGG19	80.20 (10.54)	78.96 (17.90)	82.24 (8.97)	61.39 (20.57)
18	ConvNextTiny	83.73 (3.50)	78.83 (10.73)	87.40 (2.96)	67.00 (7.92)
19	ConvNextSmall	82.53 (10.79)	77.13 (14.89)	86.75 (9.19)	64.27 (23.15)
20	VGG19	74.10 (3.40)	84.35 (8.98)	64.79 (6.95)	50.21 (7.59)
21	SwinV2T	80.71 (5.54)	74.31 (4.33)	85.76 (10.82)	61.44 (10.82)
22	ConvNextBase	80.73 (5.87)	78.18 (18.05)	81.37 (8.50)	60.79 (13.18)
23	ConvNextTiny	79.52 (3.94)	75.65 (12.96)	83.12 (8.86)	59.82 (8.02)
24	SwinV2B	80.71 (6.67)	81.73 (10.91)	80.12 (12.12)	62.46 (12.61)
25	ConvNextBase	72.87 (8.09)	70.81 (15.21)	73.85 (6.29)	44.78 (17.83)

Table 6 presents the performance metrics for the best models, freezing a subset of the model's layers during training, which can assist in preventing overfitting and enhancing performance. ConvNextSmall achieves its peak accuracy of 84.44% when employing 25% freezing. However, this performance falls short of the 87.99% accuracy it attains without implementing any freezing technique. Conversely, the SwinV2T model shows improved performance with 50% freezing, achieving an accuracy of 86.15% and an MCC of 72.30%. This is higher than its non-frozen performance (85.54% accuracy and 71.53% MCC), indicating that freezing half of the layers helps better adapt to the task. The TrOCR-Small-Hand model exhibits similar behaviour, gaining an accuracy improvement of 6.63%

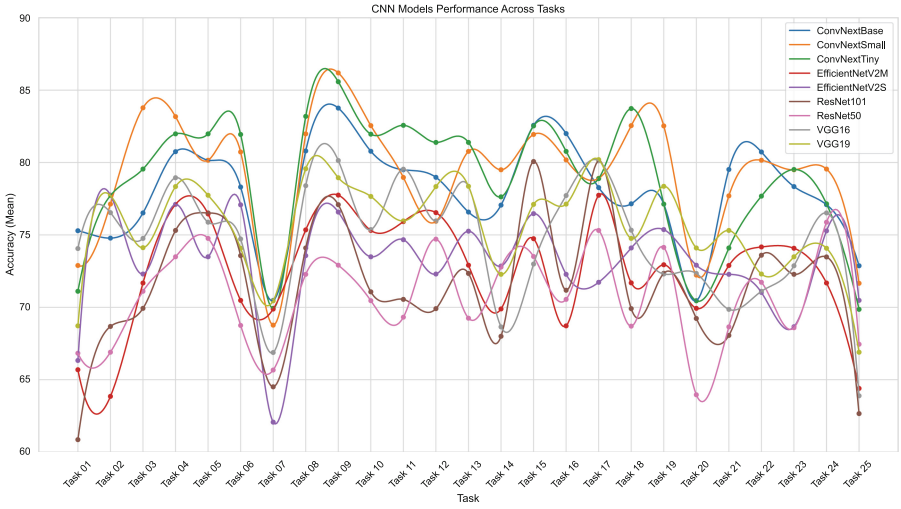


Fig. 4. Accuracy comparison of CNN models across 25 tasks.

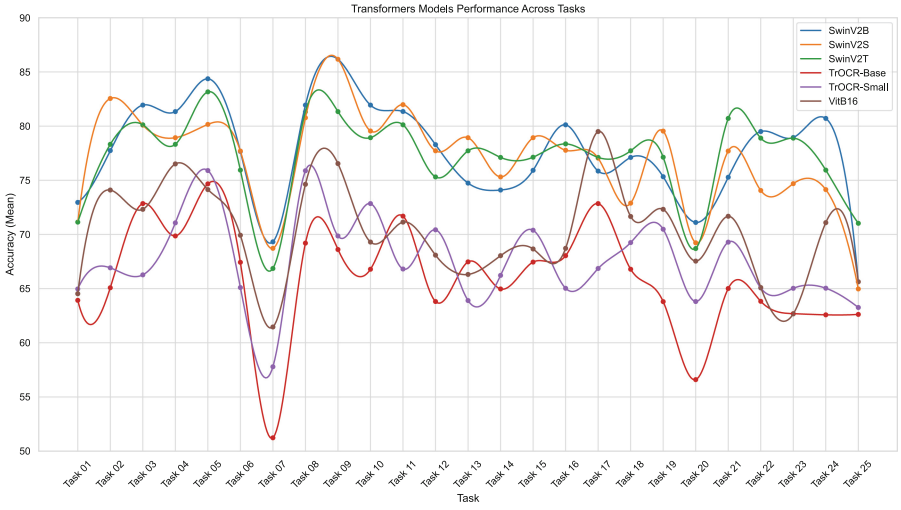


Fig. 5. Accuracy comparison of Transformers models across 25 tasks.

when it is subjected to 75% freezing than when it is not. The VitB16 model shows the best performance is with 25% freezing, but this is still lower than its non-frozen performance (81.34% vs 77.70% accuracy), suggesting that this model benefits from having more trainable parameters for this specific task. Overall, these results demonstrate that the impact of freezing varies across different model architectures. While some models like SwinV2T and TrOCR-Small-Hand benefit from freezing, others like ConvNextSmall and VitB16 perform better when more

Table 5. Performance metrics of CNN and Transformer models using majority voting.

Model	Accuracy	Specificity	Sensitivity	MCC
VGG16	82.57 (8.11)	85.02 (11.98)	79.69 (8.94)	65.01 (17.04)
VGG19	84.37 (3.35)	91.99 (5.24)	77.15 (9.23)	70.09 (5.56)
ResNet50	79.50 (2.39)	**92.06 (7.45)**	67.97 (8.41)	62.09 (4.64)
ResNet101	78.93 (7.61)	90.16 (10.74)	67.73 (11.99)	59.84 (15.39)
EfficientNetV2M	80.73 (8.89)	86.63 (13.87)	74.76 (10.46)	62.56 (18.45)
EfficientNetV2S	78.31 (5.89)	87.73 (10.86)	69.20 (9.48)	58.55 (12.37)
ConvNextTiny	87.38 (5.78)	88.46 (10.19)	86.59 (13.30)	76.38 (10.70)
ConvNextSmall	**87.99 (5.30)**	85.15 (9.90)	**89.69 (13.06)**	**76.75 (10.63)**
ConvNextBase	86.17 (8.65)	87.37 (13.44)	84.01 (12.28)	72.60 (18.09)
VitB16	77.70 (9.72)	86.63 (10.46)	69.42 (21.85)	58.40 (16.42)
SwinV2B	84.94 (5.07)	89.63 (8.20)	80.70 (11.98)	71.36 (9.51)
SwinV2S	83.74 (3.59)	83.08 (12.53)	84.30 (12.29)	69.26 (7.26)
SwinV2T	85.54 (1.18)	86.89 (7.01)	84.04 (6.86)	71.53 (2.75)
TrOCR-Small-Hand	71.66 (6.63)	87.95 (8.26)	57.23 (11.58)	47.86 (11.38)
TrOCR-Base-Hand	71.00 (11.54)	90.16 (11.36)	53.85 (20.86)	48.17 (19.91)

Table 6. Performance metrics of best models with varying percentages of layer freezing during training.

Model	Freezing	Accuracy	Specificity	Sensitivity	MCC
ConvNextSmall	25%	84.44 (9.30)	82.06 (7.95)	86.41 (13.11)	**69.05 (19.04)**
	50%	82.55 (5.13)	**89.48 (6.50)**	76.27 (12.90)	67.06 (9.13)
	75%	83.16 (4.82)	83.84 (13.00)	81.90 (11.53)	67.08 (11.06)
VitB16	25%	81.34 (6.71)	80.01 (10.94)	82.04 (10.70)	62.81 (14.39)
	50%	77.11 (5.60)	88.17 (9.65)	67.57 (13.57)	57.76 (11.07)
	75%	77.13 (5.51)	86.92 (11.19)	68.40 (11.93)	57.11 (10.60)
SwinV2T	25%	82.55 (6.41)	84.94 (14.67)	79.28 (17.37)	67.22 (12.94)
	50%	**86.15 (5.60)**	82.16 (14.88)	**88.54 (8.45)**	72.30 (11.91)
	75%	82.55 (9.23)	81.34 (12.02)	83.01 (12.76)	65.21 (19.20)
TrOCR-Small-Hand	25%	77.13 (8.19)	85.83 (16.42)	68.02 (12.14)	56.08 (17.52)
	50%	75.29 (7.03)	86.94 (17.19)	63.01 (12.62)	53.44 (15.87)
	75%	78.29 (7.58)	71.62 (7.92)	84.36 (11.35)	57.08 (16.40)

layers are trainable. It's also worth noting that even with freezing, ConvNextSmall and SwinV2T remain the top-performing models, consistent with the initial results.

5 Discussions

Our experimental framework results demonstrate the significance of every stage. Initially, we considered all data collectively, without isolating individual patient or task evaluations. This perspective revealed the effectiveness of our system, outperforming previous methodologies where tasks were analyzed separately for classification. Subsequently, in the second stage, we aggregated results by task, unveiling the varying significance of different tasks for our research objectives. Lastly, the third stage integrated a majority voting rule applied to predictions from the same patient, validating our hypothesis that combining classifications across multiple tasks significantly enhances patient-level classification accuracy. We evaluated and compared various architectures belonging to two distinct families: CNNs and Vision Transformers. Among the architecture families tested, ConvNext exhibited superior performance compared to other CNNs and Vision Transformers. Specifically, *ConvNextSmall* demonstrated better performance in Accuracy and Sensitivity, a crucial metric in the medical domain, indicating the proportion of correctly identified patients afflicted by a disease. Moreover, we explored the feasibility of transferring knowledge from OCR to AD recognition tasks. However, this approach achieved significantly lower performance compared to the best method, encountering challenges arising from domain disparities between OCR and AD recognition. It is important to note that the OCR approach used by these models is specifically designed to extract and interpret text from images. This means that TrOCR models focus primarily on what is written in the images rather than analyzing neuromotor characteristics that can indicate the presence of disease. Their inferior results may be attributed to the fact that text analysis is not directly relevant or effective for identifying subtle neuromotor differences between AD subjects and HC. From the second experiment, we underscored the variability in task performance within the handwriting protocol, independent of the architecture considered. Some tasks are more valuable than others for the problem in the exam, like tasks requiring cognitive and motor skills. On the contrary, simple copy tasks that do not require a particular cognitive effort obtain worse performance. This finding holds significant implications, suggesting the potential for protocol refinement or the consideration of a subset of tasks during the combination of predictions to enhance effectiveness, throwing the basis for future work.

6 Conclusions

This research proposes a non-invasive and low-cost approach that can easily reach anyone, allowing for broad screening and improving prevention. It interprets task performances and allows one to consider more samples from the same person, combining the predictions at the end.
Besides the encouraging results, every research study has room to be improved, and we mean developing new experiments and further analysis. In particular, as our system relies on offline data scanned from paper sheets, the dataset can

be easily augmented thanks to the great availability of on-paper handwritten material. As the models selected for this work are very data-hungry, we can expect an improvement in the performance by adding more data, especially for transformer-based architectures. Moreover, if we collect different handwriting samples, we can verify if the system can perform well independently of the choice of tasks. The developed approach holds promise not only for diagnosis but also for monitoring, thus facilitating the tailoring of personalized therapies. The approach has implications beyond the initial diagnosis. Continuously monitoring changes in writing style or patterns can provide valuable insights into the progression of certain conditions or disorders. This longitudinal tracking could enable clinicians to intervene proactively, adjusting treatment plans as needed and potentially improving patient outcomes.
Furthermore, our method utilizes scans of handwritten texts as its primary "offline" data source. This approach enables us to apply our analysis to historical documents recent or from many years ago–to identify indicators of AD. By examining these old text scans, we can investigate whether an individual showed signs of AD in their writings or assess their risk of developing the condition in the future.

Acknowledgements. Project ECS 0000024 "Ecosistema dell'innovazione - Rome Technopole" financed by EU in NextGenerationEU plan through MUR Decree n. 1051 23.06.2022 PNRR Missione 4 Componente 2 Investimento 1.5 - CUP H33C22000420001

References

1. Chicco, D., Jurman, G.: The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. BMC Genomics **21**, 1–13 (2020)
2. Cilia, N.D., De Stefano, C., Fontanella, F., Scotto di Freca, A.: Using genetic algorithms for the prediction of cognitive impairments. In: Lecture notes in computer science. vol. 12104, pp. 479–493. Springer (2020)
3. Cilia, N., De Stefano, C., Fontanella, F., Scotto di Freca, A.: An experimental protocol to support cognitive impairment diagnosis by using handwriting analysis. Procedia Computer Science **141**, 466 – 471 (2018)
4. Cilia, N., De Stefano, C., Fontanella, F., Scotto di Freca, A.: Using handwriting features to characterize cognitive impairment. Springer (2019)
5. Cilia, N.D., D'Alessandro, T., De Stefano, C., Fontanella, F.: Deep transfer learning algorithms applied to synthetic drawing images as a tool for supporting alzheimer's disease prediction. Mach. Vis. Appl. **33**(3), 49 (2022)
6. Cilia, N.D., D'Alessandro, T., De Stefano, C., Fontanella, F., Molinara, M.: From online handwriting to synthetic images for alzheimer's disease detection using a deep transfer learning approach. IEEE J. Biomed. Health Inform. **25**(12), 4243–4254 (2021). https://doi.org/10.1109/JBHI.2021.3101982
7. Cilia, N.D., De Stefano, C., Fontanella, F., Molinara, M., Scotto Di Freca, A.: Handwriting analysis to support alzheimer's disease diagnosis: A preliminary study. In: Vento, M., Percannella, G. (eds.) Computer Analysis of Images and Patterns. pp. 143–151. Springer International Publishing, Cham (2019)

8. Dalia Cilia, N., D'Alessandro, T., De Stefano, C., Fontanella, F.: Offline handwriting image analysis to predict alzheimer's disease via deep learning. In: 2022 26th International Conference on Pattern Recognition (ICPR). pp. 2807–2813 (2022). https://doi.org/10.1109/ICPR56361.2022.9956359
9. De Stefano, C., Fontanella, F., Impedovo, D., Pirlo, G., Scotto di Freca A, A.: Handwriting analysis to support neurodegenerative diseases diagnosis: a review. Pattern Recognition Letters **121**, 37–45 (2018)
10. Deng, J., Dong, W., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255. IEEE Computer Society (2009)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale (2021)
12. Drotár, P., Mekyska, J., Rektorová, I., Masarová, L., Smékal, Z., Faundez-Zanuy, M.: Evaluation of handwriting kinematics and pressure for differential diagnosis of parkinson's disease. Artif. Intell. Med. **67**, 39–46 (2016). https://doi.org/10.1016/j.artmed.2016.01.004
13. Folstein, M.F., Folstein, S.E., McHugh, P.R.: 'mini-mental state': A practical method for grading the cognitive state of patients for the clinician. J. Psychiatric Res. **12**(3), 189–198 (1975)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. pp. 770–778. IEEE Computer Society (2016). https://doi.org/10.1109/CVPR.2016.90
15. Iavarone, A., Ronga, B., Pellegrino, L., Loré, E., Vitaliano, S., Galeone, F., Carlomagno, S.: The frontal assessment battery (fab): Normative data from an italian sample and performances of patients with alzheimer's disease and frontotemporal dementia. Funct Neurol. **19**, 191–195 (7 2004)
16. Kandel, E.R., Schwartz, J.H., Jessell, T.M.: Principles of Neural Science. McGraw-Hill Medical, 4th edn. (Jul 2000)
17. Kobayashi, M., Yamada, Y., Shinkawa, K., Nemoto, M., Nemoto, K., Arai, T.: Automated early detection of alzheimer's disease by capturing impairments in multiple cognitive domains with multiple drawing tasks. Journal of Alzheimer's Disease **88**, 1–15 (06 2022). https://doi.org/10.3233/JAD-215714
18. Lambert, J., Giffard, B., Nore, F., de la Sayette, V., Pasquier, F., Eustache, F.: Central and peripheral agraphia in alzheimer's disease: From the case of auguste d. to a cognitive neuropsychology approach. Cortex **43**(7), 935–951 (2007)
19. Li, M., Lv, T., Lu, Y., Florencio, D., Zhang, C., Li, Z., Wei, F.: Trocr: Transformer-based optical character recognition with pre-trained models (2022)
20. Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., Wei, F., Guo, B.: Swin transformer v2: Scaling up capacity and resolution. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11999–12009 (2022). https://doi.org/10.1109/CVPR52688.2022.01170
21. Liu, Z., Mao, H., Wu, C., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. pp. 11966–11976. IEEE (2022). https://doi.org/10.1109/CVPR52688.2022.01167
22. Marti, U.V., Bunke, H.: The iam-database: an english sentence database for offline handwriting recognition. Int. J. Doc. Anal. Recogn. **5**(1), 39–46 (2002). https://doi.org/10.1007/s100320200071

23. Nasreddine, Z.S., Phillips, N.A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J.L., Chertkow, H.: The montreal cognitive assessment, moca: A brief screening tool for mild cognitive impairment. Journal of the American Geriatrics Society **53**(4), 695–699 (4 2005)
24. Neils-Strunjas, J., Groves-Wright, K., Mashima, P., Harnish, S.: Dysgraphia in Alzheimer's disease: a review for clinical and research purposes. J. Speech Lang. Hear. Res. **49**(6), 1313–30 (2006)
25. Qi, H., Zhang, R., Wei, Z., Zhang, C., Wang, L., Lang, Q., Zhang, K., Tian, X.: A study of auxiliary screening for alzheimer's disease based on handwriting characteristics. Frontiers in Aging Neuroscience **15** (2023). https://doi.org/10.3389/fnagi.2023.1117250
26. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR **abs/1409.1556** (2014)
27. Tan, M., Le, Q.V.: Efficientnetv2: Smaller models and faster training. arXiv preprint arXiv:2104.00298 (2021)
28. Vessio, G.: Dynamic handwriting analysis for neurodegenerative disease assessment: A literary review. Appl. Sci. **9**(21), 4666 (2019). https://doi.org/10.3390/app9214666
29. Werner, P., Rosenblum, S., Korczyn, A.: Handwriting process variables discriminating mild alzheimer's disease and mild cognitive impairment. Journal of Gerontology: PSYCHOLOGICAL SCIENCES **61**(4), 228–36 (2006)
30. Werner, P., Rosenblum, S., Bar-On, G., Heinik, J., Korczyn, A.: Handwriting process variables discriminating mild alzheimer's disease and mild cognitive impairment. The journals of gerontology. Series B, Psychological sciences and social sciences **61**, P228–36 (08 2006). https://doi.org/10.1093/geronb/61.4.P228

Author Index

A. Antonacopoulos et al. (Eds.): ICPR 2024, LNCS 15311, pp. 465–466, 2025.
https://doi.org/10.1007/978-3-031-78195-7