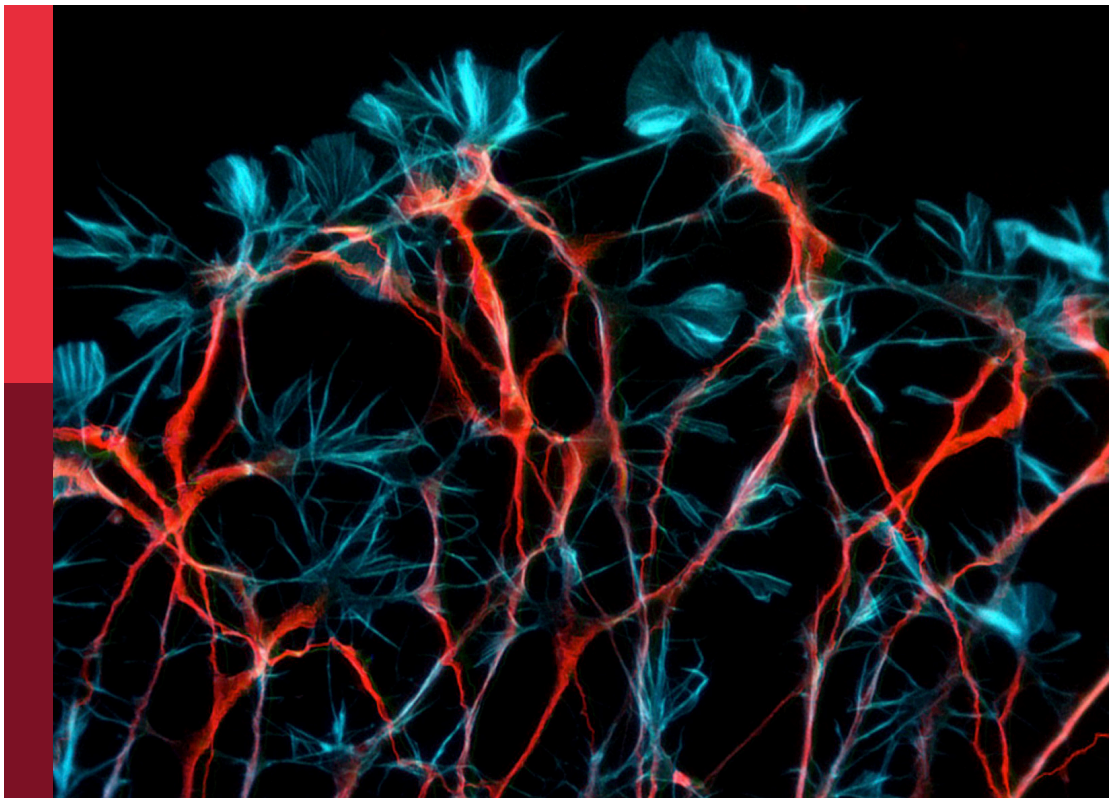
frontiers | Research Topics

Computational modeling and machine learning methods in neurodevelopment and neurodegeneration: from basic research to clinical applications

Edited by
Pablo Martinez-Cañada, Noemi Montobbio, Roberto Maffulli and Anees Abrol

Published in
Frontiers in Computational Neuroscience
Frontiers in Neuroscience



ISSN 1664-8714
ISBN 978-2-8325-5712-9
DOI 10.3389/978-2-8325-5712-9

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Computational modeling and machine learning methods in neurodevelopment and neurodegeneration: from basic research to clinical applications

Topic editors

Pablo Martinez-Cañada — University of Granada, Spain
Noemi Montobbio — University of Genoa, Italy
Roberto Maffulli — Italian Institute of Technology (IIT), Italy
Anees Abrol — Georgia State University, United States

Citation

Martinez-Cañada, P., Montobbio, N., Maffulli, R., Abrol, A., eds. (2024). *Computational modeling and machine learning methods in neurodevelopment and neurodegeneration: from basic research to clinical applications*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-5712-9

Table of contents

frontiers | Frontiers in Computational Neuroscience

TYPE Editorial
PUBLISHED 08 November 2024
DOI 10.3389/fncom.2024.1514220

OPEN ACCESS

EDITED AND REVIEWED BY
Si Wu,
Peking University, China

*CORRESPONDENCE
Pablo Martínez-Cañada
✉ pablomc@ugr.es

RECEIVED 20 October 2024
ACCEPTED 28 October 2024
PUBLISHED 08 November 2024

CITATION
Montobbio N, Maffulli R, Abrol A and Martínez-Cañada P (2024) Editorial: Computational modeling and machine learning methods in neurodevelopment and neurodegeneration: from basic research to clinical applications.
Front. Comput. Neurosci. 18:1514220.
doi: 10.3389/fncom.2024.1514220

Editorial: Computational modeling and machine learning methods in neurodevelopment and neurodegeneration: from basic research to clinical applications

Noemi Montobbio[1], Roberto Maffulli[2], Anees Abrol[3] and Pablo Martínez-Cañada[4,5]*

[1]Department of Health Sciences (DISSAL), University of Genoa, Genoa, Italy, [2]EXUS AI Labs, London, United Kingdom, [3]Center for Translational Research in Neuroimaging and Data Science, Georgia State University, Atlanta, GA, United States, [4]Research Centre for Information and Communications Technologies (CITIC), University of Granada, Granada, Spain, [5]Department of Computer Engineering, Automation and Robotics, University of Granada, Granada, Spain

KEYWORDS

computational modeling, machine learning, neurodevelopmental disorders, neurodegeneration, clinical

Editorial on the Research Topic
Computational modeling and machine learning methods in neurodevelopment and neurodegeneration: from basic research to clinical applications

Computational models and machine-learning methods are increasingly valuable for understanding how neural networks in the brain process information, and how this information influences decision-making and behavior. Abnormalities in these networks are linked to various brain disorders. Advances in brain simulation, machine learning, and neuroimaging have helped bridge different brain scales and uncover the processes underlying cognitive, motor and behavioral impairment in neurodevelopmental and neurodegenerative disorders.

The effective application of computational approaches still faces several challenges, including: the multiple spatial scales involved; the issue of interpretability of machine learning models, hampering transferability to clinical practice; and the lack of robust validation of non-invasive biomarkers of neural disorders. These challenges motivated us to edit the Research Topic "*Computational Modeling and Machine Learning Methods in Neurodevelopment and Neurodegeneration: from Basic Research to Clinical Applications*", culminating with the acceptance of 10 insightful papers that explore the subject from diverse perspectives using various innovative tools.

The contributions covered a variety of themes, including disease diagnosis (Ruppert-Junk et al., Turrisi et al., Fernández-Ruiz et al.), disease subtype or stage classification (Chen et al., Zheng et al.), predictors of disease progression (Zhang et al.), brain network simulation (Monteverdi et al., Moore et al.), lesion segmentation (Zaman et al.), and clustering methods in medicine (Poulakis and Westman). Deep learning was

widely present, and was explored from different perspectives, from the proposal of novel model architectures (Zaman et al.) to an analysis of validation and reproducibility issues (Turrisi et al.). Although most papers focused on MRI data, other neuroimaging modalities such as EEG (Zheng et al.) and speech (Fernández-Ruiz et al.) were explored as well.

Computer-aided diagnosis, as well as disease subtype or stage classification, are among the most frequently addressed tasks in machine learning studies in healthcare (Chan et al., 2020). In the present Research Topic, the study by Ruppert-Junk et al. investigated the use of [^{18}F]-FDG PET imaging to diagnose Parkinson's disease (PD) by focusing on midbrain metabolism, particularly in the substantia nigra. A machine learning model using random forest classification achieved high sensitivity and accuracy in distinguishing PD patients from healthy controls. Fernández-Ruiz et al. introduced a non-invasive method for identifying Smith–Magenis syndrome using machine learning techniques, focusing on cepstral peak prominence (CPP) from voice samples. The study significantly contributes to the theme of using computational methods for neurodevelopmental conditions by offering a potential clinical application for early diagnosis? Chen et al. explored the use of graph-based convolutional networks (GCNs) to classify multiple sclerosis (MS) clinical forms based on brain morphological connectivity from T1-weighted MRI data. The authors show how the approach outperforms state-of-the-art 3D Convolutional Neural Networks (CNNs) methods, offering insights into how computational models can help differentiate between MS subtypes.? Zheng et al. proposed a novel framework for epilepsy diagnosis using a complexity-based Graph Convolutional Neural Network (GCNN) to analyze multi-channel EEG signals across normal, acute, and chronic stages. By incorporating five complexity measures, their model achieved high accuracy in distinguishing between these phases, thus highlighting its potential in detecting chronic epilepsy for more effective intervention. Zhang et al. studied gray and white matter alterations in children affected by sensorineural hearing loss (SNHL) based on their auditory brainstem response. They identified independent predictive factors to study SNHL progression in children, highlighting the value of quantitative T1 assessments in specific regions of interest and tracking white matter and myelin volume and fraction parameters.

Automatic medical image segmentation tools are highly required by the medical community, and several deep learning techniques have been successfully applied in this field in recent years (Ramesh et al., 2021). Zaman et al. presented the Adaptive Feature Medical Segmentation Network (AFMS-Net) for 3D brain lesion segmentation. The network uses novel encoder-decoder structures for high-performance, computationally efficient segmentation, significantly advancing clinical imaging applications in scenarios requiring quick and efficient identification of key lesion areas.

Computational simulations of brain network alterations linked to neurological diseases can be a powerful and cost-effective tool to indicate new directions in clinical research (D'Angelo and Jirsa, 2022). Monteverdi et al. employed multiscale brain modeling using The Virtual Brain (TVB) with MRI data to simulate brain networks in patients with Alzheimer's disease (AD) and frontotemporal dementia (FTD). Their simulations revealed distinct disease-specific alterations in connectivity and synaptic transmission for each condition, which correlated with individual clinical profiles. These insights enhance our understanding of dementia mechanisms and may guide the development of personalized therapeutic strategies. Moore et al. proposed a novel deep learning approach to model neurodegeneration in the visual cortex through progressive lesioning of a convolutional neural network, also including a mechanism to simulate neuroplasticity by allowing the model to adapt to new information even after sustaining simulated damage. The authors show that incorporating neuroplasticity resulted in a smoother and slower decline in model performance, aligning with observed disease-related cognitive decline patterns. Overall, findings suggest that integrating neuroplasticity into deep learning models could enhance disease understanding and support testing rehabilitation approaches.

Finally, the issue of validation and reproducibility of computational techniques is raising growing interest (McDermott et al., 2021). Poulakis and Westman contributed with a letter elaborating on the applications and challenges of clustering for studying heterogeneity in psychiatric and neurological disorders. They emphasized the importance of careful methodological selection, validation, and expert involvement to address the limitations and improve the interpretation of clustering results in high-dimensional datasets. Turrisi et al. highlighted the importance of adhering to shared guidelines to ensure the reliability, robustness, and reproducibility of ML in healthcare. Using the challenging problem of Alzheimer's disease detection from MRI scans as a case study, the authors demonstrated best practices in data handling, model design, and assessment, while also revealing the susceptibility of prediction accuracy to modeling choices.

We believe that this Research Topic will provide readers with a stimulating overview of current themes in computational modeling and machine learning as applied to neurodevelopment and neurodegeneration. The contributions emphasize both the potential and the challenges of these approaches, offering insights that can inspire future research and ultimately support clinical advancements in diagnosing and treating brain disorders.

Author contributions

NM: Writing – original draft, Writing – review & editing. RM: Writing – original draft, Writing – review & editing. AA: Writing – original draft, Writing – review & editing. PM-C: Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. NM received funding from #NEXTGENERATIONEU (NGEU), Ministry of University and Research (MUR); and National Recovery and Resilience Plan (NRRP), MNESYS project (PE0000006; DN. 1553 11.10.2022). PM-C received funding from grants PID2022-139055OA-I00 and PID2022-137461NB-C31 funded by MCIN/AEI/10.13039/501100011033 and by "ERDF

A way of making Europe", and from "Junta de Andalucia" - Postdoctoral Fellowship Programme PAIDI 2021.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Chan, H. P., Hadjiiski, L. M., and Samala, R. K. (2020). Computer-aided diagnosis in the era of deep learning. *Med. Phys.* 47, e218–e227. doi: 10.1002/mp.13764

D'Angelo, E., and Jirsa, V. (2022). The quest for multiscale brain modeling. *Trends Neurosci.* 45, 777–790. doi: 10.1016/j.tins.2022.06.007

McDermott, M. B. A., Wang, S., Marinsek, N., Ranganath, R., Foschini, L., and Ghassemi, M. Reproducibility in machine learning for health research: still a ways to go. *Sci. Transl. Med.* 13:eabb1655. doi: 10.1126/scitranslmed.abb1655

Ramesh, K. K. D., Kumar, G. K., Swapna, K., Datta, D., and Rajest, S. S. (2021). A review of medical image segmentation algorithms. *EAI Endor. Trans. Pervasive Health Technol.*7, e6–e6. doi: 10.4108/eai.12-4-2021.169184

frontiers | Frontiers in Aging Neuroscience

TYPE Original Research
PUBLISHED 28 July 2023
DOI 10.3389/fnagi.2023.1204134

OPEN ACCESS

EDITED BY
Pablo Martinez-Cañada,
University of Granada, Spain

REVIEWED BY
Basabdatta Sen Bhattacharya,
Birla Institute of Technology and Science, India
Li Su,
The University of Sheffield, United Kingdom
Lorenzo Pini,
University of Padua, Italy

*CORRESPONDENCE
Anita Monteverdi
✉ anita.monteverdi01@universitadipavia.it
Egidio D'Angelo
✉ dangelo@unipv.it

RECEIVED 11 April 2023
ACCEPTED 10 July 2023
PUBLISHED 28 July 2023

CITATION
Monteverdi A, Palesi F, Schirner M, Argentino F, Merante M, Redolfi A, Conca F, Mazzocchi L, Cappa SF, Cotta Ramusino M, Costa A, Pichiecchio A, Farina LM, Jirsa V, Ritter P, Gandini Wheeler-Kingshott CAM and D'Angelo E (2023) Virtual brain simulations reveal network-specific parameters in neurodegenerative dementias.
Front. Aging Neurosci. 15:1204134.
doi: 10.3389/fnagi.2023.1204134

Virtual brain simulations reveal network-specific parameters in neurodegenerative dementias

Anita Monteverdi[1]*, Fulvia Palesi[2], Michael Schirner[3,4,5,6,7], Francesca Argentino[2], Mariateresa Merante[2], Alberto Redolfi[8], Francesca Conca[9], Laura Mazzocchi[10], Stefano F. Cappa[9,11], Matteo Cotta Ramusino[12], Alfredo Costa[2,12], Anna Pichiecchio[2,10], Lisa M. Farina[9], Viktor Jirsa[13], Petra Ritter[3,4,5,6,7], Claudia A. M. Gandini Wheeler-Kingshott[1,2,14] and Egidio D'Angelo[1,2]*

[1]Unit of Digital Neuroscience, IRCCS Mondino Foundation, Pavia, Italy, [2]Department of Brain and Behavioral Sciences, University of Pavia, Pavia, Italy, [3]Berlin Institute of Health, Charité – Universitätsmedizin Berlin, Berlin, Germany, [4]Department of Neurology with Experimental Neurology, Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt Universität zu Berlin, Berlin, Germany, [5]Bernstein Focus State Dependencies of Learning and Bernstein Center for Computational Neuroscience, Berlin, Germany, [6]Einstein Center for Neurosciences Berlin, Berlin, Germany, [7]Einstein Center Digital Future, Berlin, Germany, [8]Laboratory of Neuroinformatics, IRCCS Istituto Centro San Giovanni di Dio Fatebenefratelli, Brescia, Italy, [9]IRCCS Mondino Foundation, Pavia, Italy, [10]Advanced Imaging and Artificial Intelligence Center, IRCCS Mondino Foundation, Pavia, Italy, [11]University Institute of Advanced Studies (IUSS), Pavia, Italy, [12]Unit of Behavioral Neurology, IRCCS Mondino Foundation, Pavia, Italy, [13]Institut de Neurosciences des Systèmes, INSERM, INS, Aix Marseille University, Marseille, France, [14]NMR Research Unit, Queen Square Multiple Sclerosis Centre, Department of Neuroinflammation, UCL Queen Square Institute of Neurology, London, United Kingdom

Introduction: Neural circuit alterations lay at the core of brain physiopathology, and yet are hard to unveil in living subjects. The Virtual Brain (TVB) modeling, by exploiting structural and functional magnetic resonance imaging (MRI), yields mesoscopic parameters of connectivity and synaptic transmission.

Methods: We used TVB to simulate brain networks, which are key for human brain function, in Alzheimer's disease (AD) and frontotemporal dementia (FTD) patients, whose connectivity and synaptic parameters remain largely unknown; we then compared them to healthy controls, to reveal novel *in vivo* pathological hallmarks.

Results: The pattern of simulated parameter differed between AD and FTD, shedding light on disease-specific alterations in brain networks. Individual subjects displayed subtle differences in network parameter patterns that significantly correlated with their individual neuropsychological, clinical, and pharmacological profiles.

Discussion: These TVB simulations, by informing about a new personalized set of networks parameters, open new perspectives for understanding dementias mechanisms and design personalized therapeutic approaches.

KEYWORDS

virtual brain modeling, brain dynamics, excitatory/inhibitory balance, Alzheimer's disease, frontotemporal dementia, resting-state networks

Background

The advent of advanced in human *in vivo* recordings of brain signals from, e.g., magnetic resonance imaging (MRI), has led to the identification of brain networks that subtend specific functions (Smitha et al., 2017). The structural, metabolic and/or functional alteration of such networks eventually leads to the clinical manifestation of neurological diseases. In parallel, mathematical modeling of cellular and microcircuit functions are emerging, providing tools to link the micro- to the meso- and the macro-scale properties of brain signals (D'Angelo and Jirsa, 2022).

Neurodegenerative dementias include several neuropathological forms, primarily Alzheimer's disease (AD) and frontotemporal dementia (FTD). AD is associated with the accumulation of amyloid-β plaques and neurofibrillary tangles, which are widely recognized as typical biomarkers confirming the disease diagnosis. Most AD cases present the typical amnesic form, which reflects the accumulation of protein aggregates in medial temporal lobe structures and evolves in multidomain dementia. Dysfunctions outside the mesial temporal regions characterize atypical AD variants, which present predominant visual, language, executive, behavioral, or motor dysfunction (Graff-Radford et al., 2021). FTD is a heterogeneous neurodegenerative disorder, clinically characterized by behavioral abnormalities, language deficit and motor symptoms. Focal frontal and temporal atrophy are the main macroscopic evidence of FTD pathological changes and distinct atrophy patterns can be associated with different variants (Leyton and Hodges, 2010). Post-mortem histology and *in vivo* functional MRI (fMRI) studies have suggested a differential engagement of various brain networks in these diseases. However, a comprehensive assessment of functional connectivity (FC) changes in multiple networks *in vivo* to compare dementias subtypes has been rarely performed (Castellazzi et al., 2014), in favor of investigating specific networks, in particular the default mode network (DMN) specifically in AD (Hohenfeld et al., 2018). Increasing evidence underlines the need to expand the investigation beyond the DMN, considering that widespread increases and decreases in structural, functional and metabolic connectivity have been observed in different brain areas of AD patients (Arnemann et al., 2018; Stefanovski et al., 2021). Moreover, the development of *in vivo* imaging biomarkers of brain function becomes necessary to achieve efficient tailored diagnosis and personalized treatment, especially in less frequent and more heterogeneous conditions, such as atypical forms of AD or FTD variants (Graff-Radford et al., 2021).

Advanced recording techniques, such as MRI and/or high-density electroencephalography (hd-EEG), are mostly used to study structural and functional brain networks properties and their changes in pathological conditions, but they provide little information about cellular properties such as spatio-temporal dynamics of cellular communication, neuronal firing integrity or synaptic transmission. Proton magnetic resonance spectroscopy (MRS) provides a non-invasive technique to investigate the biochemical properties of the brain and detect metabolic alterations in dementia; aside the fact that acquiring MRS data would prolong the scan time for patients, who are already difficult to image, there is the consideration that most of the studies report extremely heterogeneous results, making clinical application of MRS in AD still limited (Maul et al., 2020). On the other hand, recent studies have addressed FC in FDG-PET data, highlighting the presence of specific metabolic patterns in neurodegenerative dementias, which requires individual subjects' analyses pipelines as appropriate for clinical settings (Titov et al., 2017).

Therefore, very little is known about the cellular and synaptic changes typical of different diseases, and even more so about whether changes that have cascaded from cells to networks are specific to individual patients.

Recent advances in multiscale brain modeling offer promising tools to study the whole brain temporal dynamics, integrating macroscopic information from structural and functional MRI with mathematical mesoscale representations of the underlying ensemble properties of cells and microcircuits. In particular, The Virtual Brain (TVB) modeling workflow allows the non-invasive investigation of brain features, such as network connectivity strength and excitatory/inhibitory (E/I) balance (Stefanovski et al., 2021; D'Angelo and Jirsa, 2022), which are relevant to brain disease and can be determined for each patient. The E/I balance, in turn, can be extracted at whole brain level or for specific brain networks from parameters measuring excitatory coupling, inhibitory coupling, and recurrent excitation inside network nodes (Deco et al., 2014). Importantly, all neurological conditions involve changes at multiple scales and can gain from the use of TVB for understanding the impact of cellular and microcircuit properties alterations on brain function. The promise for clinical use of TVB has been already suggested in epilepsy surgery (Jirsa et al., 2017), stroke (Falcon et al., 2016), brain tumors (Aerts et al., 2018), Multiple Sclerosis (Marti-Juan et al., 2022), and neurodegenerative conditions like dementia (Zimmermann et al., 2018; Stefanovski et al., 2019; Monteverdi et al., 2022; Triebkorn et al., 2022). Interestingly, the central position of an E/I imbalance in the cascade of pathophysiological events in AD is increasingly recognized (Maestú et al., 2021). However, very little is known on how such network neurophysiology acts in concert with structural and FC alterations to determine cognitive decline. Retrieving E/I information, even if summarized in mesoscale network parameters, is extremely important, as it will provide new insights in neurodegenerative mechanisms of disease that will eventually impact on finding effective treatments.

In this work, we applied TVB to enable the non-invasive investigation of connectivity strength and E/I balance in a heterogeneous cohort of dementia patients, including typical and atypical AD and FTD variants. We explored the relationship between neurophysiological parameters provided by TVB in multiple brain networks and neuropsychological scores recorded during patient examinations. TVB parameters differentiated AD from FTD and proved to be sensitive to profiles of cognitive performance and ongoing pharmacological treatment. In aggregate, this study shows how TVB analysis can be used to provide personalized fingerprints of dementia patients, opening new perspectives for differential diagnosis and for tailoring pharmacological and interventional workflows.

Abbreviations: expFC, experimental functional connectivity; expFCD, experimental dynamic functional connectivity; FC, functional connectivity; PCC, Pearson correlation coefficient; SC, structural connectivity; simFC, simulated functional connectivity; simFCD, simulated dynamic functional connectivity; TVB, The Virtual Brain; AD, Alzheimer's disease; FTD, frontotemporal dementia.

Materials and methods

Subjects

Twenty-three patients affected by neurodegenerative diseases were recruited at the IRCCS Mondino Foundation. The study was approved by the Local Ethical Committee and carried out in accordance with the Declaration of Helsinki. Written informed consent was obtained from all subjects. The protocol was approved by the Local Ethical Committee of the IRCCS Mondino Foundation. Patients underwent a complete diagnostic workup including clinical and neuropsychological assessment (see section below) MRI, and, when available, cerebrospinal fluid (CSF) biomarkers (amyloid-β and τ protein) assessment following the harmonized protocol of the RIN network [Italian Network of the Institutes (IRCCS) of Neuroscience and Neurorehabilitation] (Nigri et al., 2022). Subjects were classified into two main groups: 16 AD patients (13 females, 70 ± 8 years) and 7 FTD patients (1 female, 69 ± 5 years), further classified into distinct phenotypes. In particular, AD patients were additionally classified into: typical AD (10 subjects; Dubois et al., 2014); AD logopenic variant (2 subjects; Dubois et al., 2014); AD frontal variant (ADfv, 1 subject; Dubois et al., 2014); AD posterior cortical atrophy (ADpca, 1 subject; Dubois et al., 2014). One patient was classified as having corticobasal syndrome (CBS, 1 subject; Hassan et al., 2011), and one with dementia with Lewy bodies (DLB, 1 subject; McKeith et al., 2017). On the other hand, FTD patients were classified into: behavioral FTD (FTDbv, 5 subjects; Rascovsky et al., 2011); Primary Progressive Aphasia non-fluent variant (PPAnf, 1 subject; Gorno-Tempini et al., 2011), and Primary Progressive Aphasia semantic variant (PPAsv, 1 subject; Gorno-Tempini et al., 2011). Pharmacological therapy was also recorded.

Ten healthy controls (HC, 6 females, 67 ± 3 years) were enrolled on a voluntary basis as reference group. All HC underwent clinical assessment to exclude any cognitive impairment. For all subjects, exclusion criteria were: age >80 years, a diagnosis of significant medical, neurological and psychiatric disorder, pharmacologically treated delirium or hallucinations and secondary causes of cognitive decline (e.g., vascular metabolic, endocrine, toxic, and iatrogenic). **Supplementary Table 1** shows demographic, clinical, and neuropsychological data.

Neuropsychological assessment

All subjects underwent a neuropsychological examination based on a standardized battery of tests to assess their global cognitive status (Mini-Mental State Examination, MMSE) and different cognitive domains: memory (verbal: Rey's Auditory Verbal Learning Test, RAVLT; visuo-spatial: Rey–Osterrieth complex figure recall), phonemic and semantic fluency, visuo-constructional abilities (Rey–Osterrieth complex figure copy), attention (Trial Making Test part A, TMT-A) and executive functions (Frontal Assessment Battery, FAB; Trial Making Test part B and B-A; Stroop color-word test interference, time and errors; Raven's Colored Progressive Matrices, CPM47).

Raw scores were corrected for the effect of age, education, and sex according to the reference norms for the Italian population. Accordingly, corrected scores were classified into five Equivalent Scores (ES), from 0 to 4, with an ES of 0 reflecting a pathological performance, based on percentiles (Capitani and Laiacona, 1997). Domain scores, calculated by averaging the ES of the single tests, were obtained for memory, language-fluency, visuo-constructional abilities, attention, and executive functions, respectively.

MRI acquisitions

All subjects underwent MRI examination using a 3T Siemens Skyra scanner with a 32-channel head coil. The MRI protocol was harmonized within the RIN network including both diffusion weighted imaging (DWI) and resting-state fMRI (rs-fMRI) (Nigri et al., 2022). For DWI data a two-shell standard single-shot echo-planar imaging sequence (EPI) [voxel size = 2.5 mm × 2.5 mm × 2.5 mm, TR/TE = 8,400/93 ms, two shells with 30 isotropically distributed diffusion-weighted directions, diffusion weightings of 1,000 and 2,000 s/mm^2, 7 non-diffusion weighted b = 0 s/mm^2 images (b_0 images) interleaved with diffusion-weighted volumes] was implemented, and 3 non-diffusion weighted images with the reversed phase-encoding acquisition were additionally acquired for distortion correction. For the rs-fMRI data, GE-EPI sequence (voxel size = 3 mm × 3 mm × 3 mm, TR/TE = 2,400/30 ms, 200 volumes) was set. For anatomical reference, the protocol included a whole brain high-resolution 3D sagittal T1-weighted (3DT1) scan (TR/TE = 2,300/2.96 ms, TI = 900 ms, flip angle = 9°, voxel size = 1 mm × 1 mm × 1 mm).

Preprocessing of DWI and fMRI data

Preprocessing of diffusion and fMRI data was performed according to Monteverdi et al. (2022). Briefly, DWI data were denoised, and corrected for motion and eddy currents distortions (FMRIB Software Library and FSL)[1] (Andersson and Sotiropoulos, 2016), then white matter, gray matter (GM), subcortical GM and CSF were segmented from the co-registered 3DT1 volume (MRtrix3)[2] (Patenaude et al., 2011). 30 million streamlines whole-brain anatomically constrained tractography (Smith et al., 2012) was performed within MRtrix3, estimating fibers orientation distribution with multi-shell multi-tissue constrained spherical deconvolution (CSD) and using probabilistic streamline tractography (Tournier et al., 2012). fMRI preprocessing was carried out combining SPM12[3], FSL and MRtrix3 commands in a custom MATLABR2019b script. Marchenko–Pastur principal component analysis (MP-PCA) denoising (Ades-Aron et al., 2020) was firstly performed, followed by slice-timing correction, realignment, co-registration to the 3DT1 volume, polynomial detrending, nuisance regression of 24 motion parameters (Friston et al., 1996) and CSF temporal signal (Muschelli et al., 2014), and temporal band-pass filtering (0.008–0.09 Hz).

1 https://fsl.fmrib.ox.ac.uk/fsl/fslwiki
2 http://www.mrtrix.org
3 https://www.fil.ion.ucl.ac.uk/spm

Structural and functional connectivity

An *ad hoc* anatomical atlas in MNI (Montreal Neurological Institute) space was created combining 93 cerebral (AAL) (including cortical/subcortical structures) and 33 cerebellar (SUIT) labels (Diedrichsen et al., 2009). We then performed a mapping between our *ad hoc* atlas and the Buckner and Yeo (Buckner et al., 2011; Thomas Yeo et al., 2011) cerebral and cerebellar functional atlases to select the gray matter anatomical nodes of six networks known to support specific functions: (i) integrative networks: DMN, frontoparietal network (FPN), limbic network (LN), and attention network (AN); (ii) motor and sensory networks: visual network (VN) and somatomotor network (SMN) (**Figure 1**). For each subject, the gray matter parcellation of our combined anatomical atlas was applied to the whole-brain tractography to extract a whole-brain structural connectivity (SC) matrix, with the normalized number of streamlines as edges and cortical/subcortical/cerebellar areas as nodes. The subset of nodes defining each network and their connections were extracted from whole-brain SC obtaining specific network SC matrices, used as input to TVB (as detailed below). In addition, both static and dynamic experimental FC (expFC and expFCD, respectively) were reconstructed from rs-fMRI data for each of the six brain networks, to capture not only synchronous fluctuations of BOLD signals but also their spatiotemporal-dynamics during resting-state (Hansen et al., 2015). The expFC matrix was created by extracting the time-course of BOLD signals for each node and computing the Pearson's correlation coefficient (PCC) of the time-course of pairs of atlas-defined brain regions. Matrix elements were converted with a Fisher's z transformation and thresholded at 0.1206 (Palesi et al., 2020). FCD is the dynamic representation of FC over the time and reflects time-variant changes of resting state recordings. To obtain expFCD, expFC was computed over a sliding window of 40 s (expFCsw), shifted incrementally by 1 repetition time, which for our data it means to have 178 expFCsw (Battaglia et al., 2020). Then, each expFCsw was vectorized by considering the upper triangular entries and the vectorized expFCsw were correlated with each other generating the expFCD. Thus, expFCD was calculated as a time-versus-time matrix, containing the Pearson correlation between each expFCsw and all expFCsw, centered at all other time points along the total acquisition window, quantifying, therefore, time-evolving dynamics.

Virtual brain modeling

The TVB workflow [reported in Monteverdi et al. (2022) for the whole brain] was applied to each one of the six selected brain networks (**Figure 2**). The Wong-Wang neural mass model (Deco et al., 2014; **Supplementary Figure 1**), implemented with an optimized C code (Schirner et al., 2022), was chosen to simulate local microcircuits activity, resulting from two populations of interconnected excitatory and inhibitory neurons coupled through NMDA and GABA receptor types. In our TVB simulations, this neural mass model was associated to each node of the network, while the SC matrix was used for the nodes interconnection. A set of parameters had to be tuned globally for each network: the global coupling (G), which is a scaling factor that represents the connections strength, and three synaptic parameters, i.e., the excitatory (NMDA) synapses (J_{NMDA}), the inhibitory (GABA) synapses (J_i), and the recurrent excitation (w_+). The neural activity simulated with TVB was fed into the Balloon-Windkessel hemodynamic model (Stephan et al., 2007) to reconstruct resting-state BOLD fMRI time-courses over 8 min length and compute simulated FC (simFC) and FCD (simFCD). Parameters were adjusted iteratively using expFC and expFCD of each network as targets to optimize model fitness and the validity of the result was assessed by iterating the optimization using different initial conditions (**Supplementary Figure 2**; Good et al., 2022). For the simFC vs. expFC comparison, model parameters were tuned until the PCC between experimental and simulated data reached the highest value. For the simFCD vs. expFCD comparison, differences between experimental and simulated FCD were assessed using the Kolmogorov–Smirnov (KS) distance: lower KS values corresponded to a lower distance of frame-by-frame FCD properties, meaning that model and experimental matrices were closest to each other. Thus, to achieve the optimal TVB simulation it was necessary to find both the highest PCC and the lowest KS values. To this aim, an overall cost function was defined as (1 − PCC) + KS and lowest cost function values implied the best fit both to static and dynamic functional data (Kong et al., 2021).

Statistical analysis

Statistical tests were performed using SPSS software version 21. Optimal TVB parameters derived for each subject and for each network were tested for normality (Shapiro–Wilk) and then two control tests were performed to assess: (i) whether different networks presented a different E/I balance within the same clinical group (i.e., evaluation of the inter-network E/I balance); and (ii) whether inter-networks E/I balance changed in healthy vs. pathological subjects. Two statistical tests were used: (i) univariate general linear model followed by bias-corrected accelerated Bootstrap (Pek et al., 2018) to correct for age and gender differences in the groups and take into account non-Gaussian data distributions; and (ii) multivariate general linear model between the mean difference (i.e., the difference between the mean value) of TVB parameters in each network compared to the other networks in different clinical groups. Then, a multiple regression analysis was performed to investigate the relationship between individual scores of the 5 cognitive domains (memory, language-fluency, visuo-constructional abilities, attention, and executive functions) and the optimal TVB parameters. Neuropsychological scores in each cognitive domain were considered as dependent variables while model parameters derived for each network were used as predictors in a backward approach. The regression algorithm automatically removed one or more predictors to identify which of them significantly ($p < 0.05$) explained neuropsychological scores variance.

Meaningful TVB parameters were given as an input to clustering analysis. To avoid overfitting in the study design, the clustering algorithm first performed a feature selection reducing the number of TVB parameters (i) through a semi-supervised approach using LASSO regression model with TVB parameters as independent variables and the diagnostic class as dependent

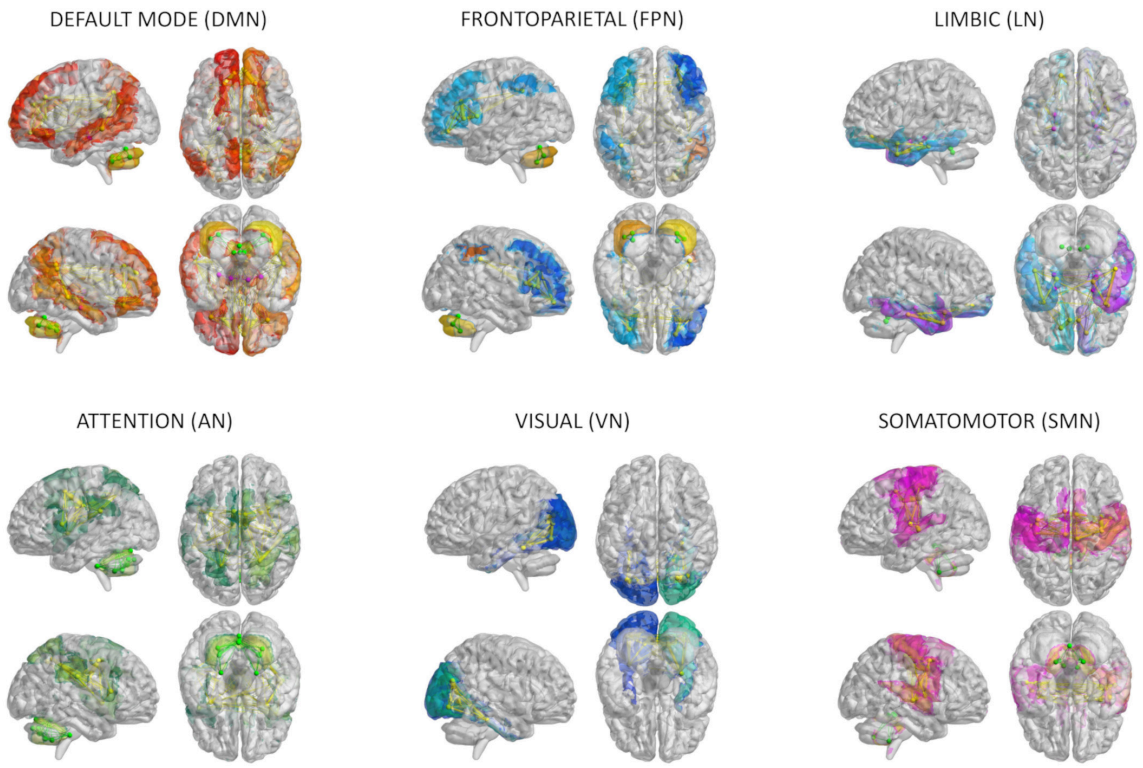


FIGURE 1
Brain networks. The six networks considered for modeling brain dynamics with The Virtual Brain (TVB): default-mode (DMN), frontoparietal (FPN), limbic (LN), attention (AN), visual (VN), and somatomotor (SMN) network. These networks were defined according to Buckner and Yeo atlases and extracted from whole-brain structural connectivity matrices of each subject, choosing a subset of nodes and connections from the whole brain parcellation. Nodes and edges considered for each network are differently colored.

variable; (ii) via PCC between the survived TVB parameters and the diagnostic class; (iii) through Variant Inflation Factors to find out just three meaningful but not correlated features. Then the number of clusters was derived using Gap statistics and the *K*-means algorithm was applied to label each subject into one cluster defining a personalized fingerprint (Redolfi et al., 2020).

Code and data accessibility

All codes used for this study are open source. The optimized TVB C code can be found at https://github.com/BrainModes/fast_tvb. The dataset will be made available at 10.5281/zenodo.811392.

Results

E/I balance in brain networks

Model optimization was performed in each of the six brain networks considered in this work. Global coupling (G) and mesoscopic network parameters (J_i, J_{NMDA}, and w_+) were adjusted iteratively to fit the experimental data. The reliability of the procedure was assessed by an extensive exploration of the parameter space and by iterating the optimization using different initial conditions (**Supplementary Figure 2**; Good et al., 2022). Model optimization yielded subject-specific sets of model parameters describing connectivity and E/I balance in each network. TVB parameters revealed differences between networks of healthy and pathological subjects (**Supplementary Figure 3** and **Supplementary Table 2**) that will be further analyzed and explained below.

Differences of E/I balance between pathological groups

The mean difference of each network compared to the others was computed in different clinical groups for all the TVB parameters (i.e., G, J_i, J_{NMDA}, and w_+). Significant mean difference changes were found both for the TVB parameters in several networks (**Figure 3**) with network changes summarized in **Figure 4**. In particular, both in AD and FTD, the connectivity strength (G) decreased in LN and increased in DMN compared to HC; in FTD, G of FPN was lower with respect to other networks. Considering mesoscale synaptic parameters, both FTD and AD showed lower excitatory coupling (J_{NMDA}) in SMN compared to HC; in FTD, J_{NMDA} was lower in VN and higher in FPN; in AD, J_{NMDA} in DMN was higher with respect to other networks. Both in AD and FTD, recurrent excitation (w_+) increased in SMN compared to HC; in FTD, w_+ was lower in FPN; in AD w_+ was lower in DMN with respect to other networks. In FTD, inhibitory coupling (J_i) was lower in FPN and higher in DMN; in AD, AN showed higher J_i and LN lower J_i with respect to other networks.

10.3389/fnagi.2023.1204134

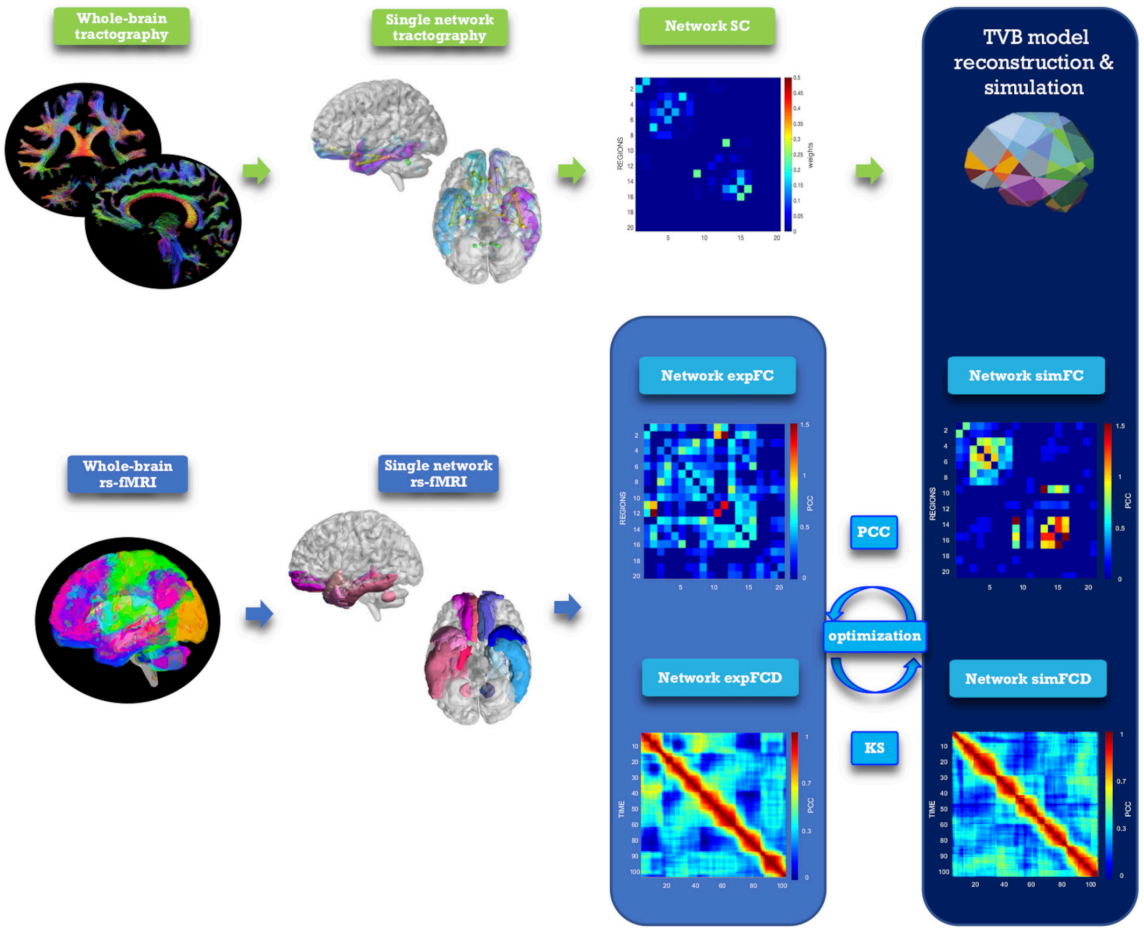


FIGURE 2
Analysis and modeling workflow. Schematic representation of MRI processing steps integrated in the modeling workflow. From top left, clockwise: diffusion-weighted images after preprocessing and tractography, extraction of a network, structural connectivity (SC) matrix reconstruction for the selected network, TVB simulation performed for the network, reconstruction of simulated static and dynamic (simFC and simFCD) functional connectivity matrices of the same network, optimization of the simulation using model inversion with the experimental FC and FCD (expFC and expFCD), derived from BOLD signals of nodes belonging to the network, as target. Optimal TVB simulation implies both the highest Pearson correlation coefficient (PCC) for static functional data and the lowest Kolmogorov–Smirnov (KS) distance for dynamic functional data.

Clinical relevance of TVB parameters

To assess the significance of the observed mean difference changes in TVB parameters, these were used in backward regression to explain the variation of scores associated to different neuropsychological domains assessed in patients. Network-specific levels of global coupling, excitatory coupling, inhibitory coupling, and recurrent excitation (predictors) significantly ($p < 0.05$) explained a percentage of variance in the cognitive domains, in which the network is involved (Table 1). The explained variance ranged from ∼20 to ∼45%. Therefore, the mean difference changes in TVB parameters were relevant to explain the neuropsychological performance of patients.

Patients' labeling according to network properties

The TVB parameters that significantly explained the neuropsychological performance were considered for patients' labeling using machine learning strategies. From the nineteen parameters identified with backward regression (Table 1) the LASSO algorithm allowed to reduce them to six. Then, G of FPN was excluded, presenting PCC <0.1, and after Variant Inflation Factors three independent and not correlated variables were considered as the most informative features to perform patient's labeling: J_i of AN, G of the LN and G of the DMN. Gap statistics identified that seven homogeneous classes would be appropriate and the *K*-means assigned each subject to one of the seven clusters. Each of the identified clusters was characterized by a specific composition of TVB network features (Figure 5A and Supplementary Figure 4). Considering the biophysical meaning of each parameter, they could be described as follows:

1. Cluster 0 and cluster 3 were mainly characterized by low connectivity strength of LN, high connectivity strength of DMN and hyperinhibition in AN;
2. Cluster 1 and cluster 4 were mainly characterized by high connectivity strength of LN, low connectivity strength of DMN and low inhibition in AN;

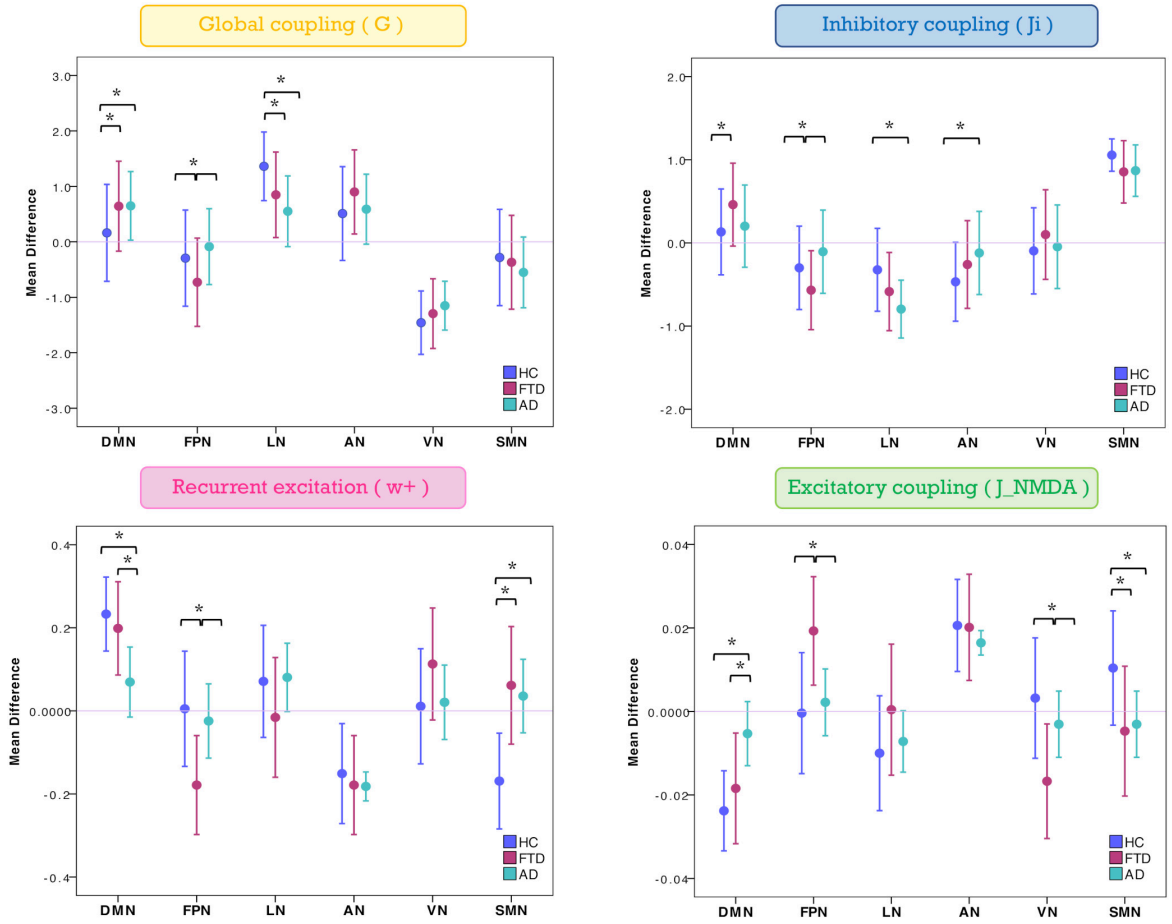


FIGURE 3
Changes of inter-network relationship. Mean difference of TVB parameters in each given network (DMN, FPN, LN, AN, VN, and SMN) against the others. Positive/negative values indicate a higher/lower TVB parameter mean in a network (on the *x*-axis) with respect to the TVB parameter mean in the others (line at mean difference 0). Asterisks indicate significant differences ($p < 0.05$) between clinical groups (HC, FTD, and AD).

3. Cluster 5 and cluster 6 were mainly characterized by high connectivity strength of LN, high connectivity strength of DMN and low inhibition in AN;
4. Cluster 2 was mainly characterized by low connectivity of LN, low connectivity strength of DMN and hyperinhibition in AN.

Clusters 0 and 3 were associated with the lowest mean MMSE values (20.39 ± 5.21 and 18.57 ± 8.28, respectively) while clusters 1 and 4 were associated with the highest mean values (29.08 ± 1.14 and 29.33 ± 1.16, respectively) (**Figure 5B** and **Table 2**). No HC was classified into clusters 0 or 3. Moreover, different disease phenotypes were distributed amongst the clusters (**Figure 5B**): typical AD subjects spread through clusters supporting a heterogeneous distribution of connectivity values in the LN and DMN networks and inhibition of the AN, but no AD patient was found in cluster 1 and the single AD patient belonging to cluster 4 presented a high MMSE score; on the other hand, cluster 0 contained the DLB phenotype, cluster 1 both the non-amnesic variants of AD (ADlv and ADpca), cluster 3 the logopenic variant and the CBS characterized by low MMSE values and cluster 5 contained the frontal variant. Considering the FTD group, FTDbv were heterogeneous and distributed amongst different clusters, but no FTDbv were found in cluster 3. On the other hand, cluster 0 contained PPAsv and cluster 5 PPAnf. Finally, pharmacological assessment of subjects belonging to different groups indicated that the majority of subjects following an antidepressant or anxiolytic treatment belonged to cluster 0 or 1 (**Table 2**). In particular, subjects belonging to cluster 0 were following an antidepressant therapy mainly with selective serotonin reuptake inhibitors (SSRIs), with the exception of one patient, treated with vortioxetine. Patients belonging to cluster 1, instead, were taking antidepressant drugs different from SSRIs, such as tricyclic antidepressants (e.g., amitriptyline) and serotonin-norepinephrine reuptake inhibitors (e.g., duloxetine), apart from one HC belonging to this group who was found to be on a SSRIs treatment.

Discussion

In this work we have generated virtual brain models of dementia patients and simulated neural dynamics of brain networks. The main result is the emergence of specific patterns of alteration in DMN, FPN, and LN, which allow to differentiate

FIGURE 4
Pathological impact on inter-network relationships. Inter-network (DMN, FPN, LN, AN, VN, and SMN) relationship patterns related to neurodegeneration are summarized in the tables. The increase (yellow) or decrease (blue) of network TVB parameters (G, global coupling; J_NMDA, excitatory coupling; w+, recurrent synaptic excitation; Ji, inhibitory coupling) is indicated with colored arrows.

TABLE 1 Backward regressions results.

Networks	Variable (neuropsychology)	Predictors (TVB-parameters)	Explained variance (%)	Significance
Visual	Memory	J_NMDA	21.3	0.027
	Language-fluency	w+, Ji	30.1	0.028
Somatomotor	Visuo-constructional	Ji	21.3	0.030
Attention	Memory	w+, J_NMDA, Ji	33.4	0.047
Limbic	Memory	Ji	21.5	0.026
	Attention	w+, G, J_NMDA	42.0	0.030
Frontoparietal	Visuo-constructional	w+, G, J_NMDA, Ji	45.7	0.027
	Executive function	J_NMDA	21.3	0.027
DMN	Language-fluency	G, J_NMDA, Ji	39.1	0.022

The variance explained by the parameters used in backward regressions is calculated with the R^2 index. Significant threshold is set at $p < 0.05$. For each cognitive domain a different combination of features significantly explains a percentage of the variance (ANOVA).

AD from FTD. Inter-subject differences, matching the individual neuropsychological profiles and pharmacological treatment, suggest that this approach can generate personalized fingerprints of the disease that could be used to set up future stratification and interventional strategies.

Average model parameters in brain networks of AD and FTD

In a first analysis, we compared AD and FTD for their average network model parameters. Model parameters markedly differentiated the mechanisms underlying brain networks dynamics in AD and FTD, with the most typical changes being concentrated in the DMN and LN of AD and in the FPN of FTD.

Integrative networks

Global coupling

In both pathologies, G increased in DMN and decreased in LN, while it decreased in FPN in FTD only. It is worth noting that, in these simulations, G represents the overall strength of connections between nodes inside a specific brain network. Moreover, G derives from dynamic TVB analysis and not from functional analysis on fMRI data (Deco et al., 2012), providing new insights into brain connectivity that do not necessarily compare to previously reported connectivity alterations.

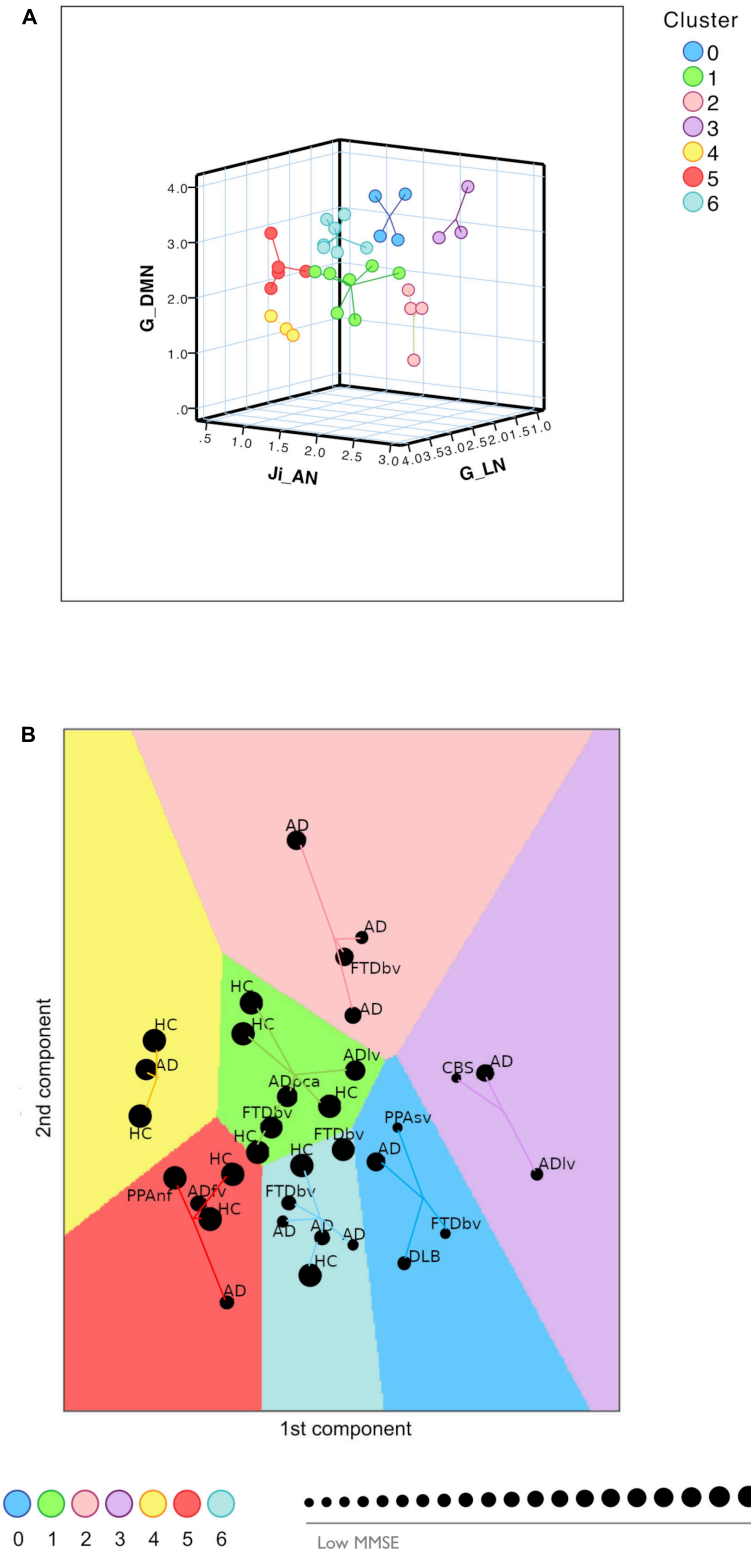


FIGURE 5
Clustering analysis. **(A)** Visual representation of the seven clusters (in different colors) identified with *K*-means analysis using the most meaningful TVB biophysical parameters as input variables. Cognitive network properties (Ji in AN, G in LN, and G in DMN) were considered as the most informative features to perform patient labeling and each of the identified clusters was characterized by a combination of low and high TVB-derived optimal parameters. Each dot represents a subject and lines connect subjects to their own cluster centroid. **(B)** Each subject was assigned to one of the seven clusters (HC, healthy control; AD, typical Alzheimer's disease; ADlv, AD logopenic variant; ADfv, AD frontal variant; ADpca, AD posterior cortical atrophy; CBS, corticobasal syndrome; DLB, dementia with Lewy bodies; FTD, frontotemporal dementia; FTDbv, FTD behavioral; PPAnf, primary progressive aphasia non-fluent variant; PPAsv, primary progressive aphasia semantic variant) identifying a personalized fingerprint based on cognitive network properties. Each dot represents a subject and lines connect subjects to their own cluster centroid. The dot dimension corresponds to the MMSE value.

TABLE 2 Mini-mental state examination (mean, SDs) and ongoing pharmacological treatment.

Group	Cluster	MMSE	Antidepressants	Anxiolytics
PPA sv	0	20.35 (5.21)	Vortioxetine + mirtazapine	
FTD bv			Fluoxetine	
AD				
DLB			Citalopram + quetiapine	Clonazepam
AD lv	1	29.08 (1.14)	Duloxetine	
FTD bv			Vortioxetine	
HC				
HC				
HC			Fluvoxamine	Lorazepam + lormetazepam + amisulpride
HC				
ADpca			Amitriptyline	
AD	2	25.04 (2.84)		
FTD bv				
AD			Venlafaxine	
AD				
AD lv	3	18.57 (8.28)		
AD				
AD cbs				Alprazolam
AD	4	29.33 (1.16)		
HC				
HC				
PPA nf	5	27.50 (3.43)		
HC				
HC				
AD			Bupropion + paroxetine	
AD fv				
FTD bv	6	24.40 (5.60)		
AD				
FTD bv				
HC				
HC				
AD				
AD				

In late onset AD there is meta-analytic evidence for a progressive decline of DMN FC, in particular in the posterior component (precuneus, posterior cingulate cortex) (Jones et al., 2016). Increased FC between the posterior DMN and high connectivity hubs, mainly located in the frontal lobes, has been reported in the prodromal stages (Jones et al., 2016). The present observation of increased G in DMN reflects hyper synchronicity, a state in which complexity is reduced along with mutual information transfer among the nodes (Borst and Theunissen, 1999). This concept, deriving from dynamic system theory, is clearly at odd with the common belief that stronger connectivity might represent compensation, leading to the conclusion that a phase-locked hypersynchronous network can perform very limited computations (Deco et al., 2012; Castellazzi et al., 2014). Consistent with this hypothesis is the finding of diffused increase of spectral power in the EEG delta band of AD patients (Babiloni et al., 2015).

Decreased FC inside LN and from LN nodes to neighboring regions has been associated with deterioration of memory and emotional functions (Cai et al., 2017). In FTDbv, a functional disconnection between frontal and limbic areas and an increased FC between DMN regions have been proposed as the probable correlates of apathy and stereotypic behavior (Zhou et al., 2010; Reyes et al., 2018). The decreased G within LN and FPN may be also very detrimental, leading to a reduction of computational states (Deco et al., 2012; Zimmermann et al., 2018).

Synaptic parameters

Another typical pattern differentiating AD from FTD emerged from synaptic parameters. Akin with neuropathology, the major AD changes were detected in DMN, while FTD changes mainly occurred in FPN. DMN showed increased excitatory coupling (J_{NMDA}) and reduced recurrent excitation (w_+) in AD, while it showed increased inhibitory coupling (J_i) in FTD. FPN showed no changes at all in AD but it showed a complex set of changes in FTD, including increased J_{NMDA}, reduced w_+ and reduced J_i. LN showed reduced J_i in AD. Therefore, the E/I balance, which remarkably impacts on brain dynamics (Deco et al., 2014), was altered in different brain networks, further differentiating AD and FTD.

We can just speculate about the meaning of these changes since information on synaptic parameters in AD and FTD pathologies is sparse. The increased J_{NMDA} in DMN may support the hyperexcitability supposed to explain cognitive impairment in AD (Palop and Mucke, 2016). Local hyperexcitability in the DMN was observed in previous studies, despite a net decrease in inhibitory and excitatory synaptic proteins (Lauterborn et al., 2021; Tok et al., 2021). The reduced J_i of the LN may support the limbic disinhibition reported in AD, which has been associated with a loss of GABAergic receptors (Jiménez-Balado and Eich, 2021). The reduced J_i of the FPN is consistent with the reduction of GABA concentration reported in FTD, which has been associated with behavioral disinhibition (Murley et al., 2021). Our simulations also predict overinhibition in the DMN of FTD, which provides a further differentiation with AD, where inhibition is not changed while excitation is enhanced. DMN has recently been suggested to take part in FTD pathophysiology (Pini et al., 2022). Therefore, the patterns of synaptic changes captured by our study prompts for further experimental and model analysis of synaptic alterations in microcircuits of the AD and FTD brain.

Motor and sensory networks

Both in AD and FTD, the SMN showed reduced J_{NMDA} and increased w_+. Although the impairment of GABAergic and glutamatergic systems in the motor and sensory networks still needs to be clarified, it should be noted that motor dysfunctions are known to occur in both AD and FTD (Burrell et al., 2011; Lorenzi et al., 2020). In AD, a reduced motor cortex excitability has been reported in mild cognitive impairment (Ferreri et al., 2021), suggesting that these parameters may change along the evolution of the disease. In FTD, motor circuit abnormalities have been suggested to depend on altered glutamatergic transmission (Benussi et al., 2020). Interestingly, in FTD abnormalities of oculomotor functions have been reported (Russell et al., 2021), which might be linked not only to SMN impairment, but also to a more extended involvement of VN, as supported by our results.

The relationship between network neurophysiology and neuropsychology

Model parameters for individual subjects were correlated with behavioral observations. Global coupling and synaptic parameters of each network significantly contributed to explain neuropsychological scores in specific cognitive domains: LN, AN, and VN with memory; DMN and VN with language-fluency; LN with attention; SMN and FPN with visuo-constructional performance; FPN with executive functions. This evidence is in line with several reports on the importance of motor regions in visuo-constructional performance (Chen et al., 2016), the contribution of AN and limbic areas in memory (Epelbaum et al., 2018), the relevance of frontoparietal areas for executive and visuo-constructional control (Melrose et al., 2013; Dixon et al., 2018), the role of DMN integration for semantic fluency (Jockwitz et al., 2017), and the involvement of visual structures in memory and language-fluency (Kucewicz et al., 2019; Vonk et al., 2019).

Thus, the relationship between neurophysiological parameters in brain networks and neuropsychological scores, which has not been investigated before, provides new cues for understanding the physiopathology of AD and FTD.

Toward personalized fingerprints of AD and FTD patients

The most meaningful model biomarkers for patient's labeling were G in DMN, G in LN, J_i in AN, consistent with known salient aspects of dementia affecting the ability of daydreaming (DMN), emotional control (LN) and attention (AN). Subjects were found to be distributed between seven different clusters revealing correspondence with their cognitive status (assessed with MMSE) and pharmacological treatment.

Patients with different MMSE scores tended to populate different clusters (see **Figure 5**), broadly separating patients from HC (MMSE >30), highlighting the importance of DMN, LN, and AN connectivity strength and E/I balance to ensure healthy cognitive function. Interestingly, high G between DMN nodes is associated with a worse performance, being hence disruptive and not compensatory. This analysis suggests that the heterogeneity of subject-specific TVB parameters is able to identify AD "subtypes" (Pini et al., 2021; Rauchmann et al., 2021) and FTD variants. Indeed, subjects belonging to atypical forms of AD and FTD variants were assigned to different clusters, capturing specific aspects of these pathologies and mostly mapping clinical severity assessed with MMSE. A finer grained analysis based on clinical phenotypes is not currently possible, given the limited sample size.

Patients' labeling based on TVB parameters correlated with pharmacological treatment. Most subjects belonging to clusters 0 and 1 were on antidepressant or anxiolytic treatment (cf. **Table 2**), which may influence the connectivity strength and the E/I balance of cognitive networks. The effect of SSRIs on LN and DMN FC is increasingly recognized (Van Wingen et al., 2014; Li et al., 2021), while the effect of antidepressant treatment with molecules different from SSRIs, such as vortioxetine, tricyclic molecules or SNRIs (Pérez et al., 2018), as well as the influence of antidepressants on GABA and glutamate levels needs further assessment (Spurny et al., 2021). Considering that patients treated with SSRIs belong to cluster 0 while patients treated with other antidepressant classes belong to cluster 1, our results pose a very intriguing question: is there an opposite impact on cognitive networks exerted by antidepressants with different mechanisms of action or does the cognitive networks profile determine pharmacological treatment response? Future work

should study TVB parameters longitudinally pre-post treatment to answer this important question with major potential clinical impact.

It should be noted that, in our cohort, patients were not treated with NMDA receptor antagonists (like memantine) (Robinson and Keating, 2006) or acetylcholinesterase inhibitors (like galantamine, rivastigmine, and donepezil) (Marucci et al., 2021), which are also known to act on AD pathophysiology. NMDA receptors are main triggers of synaptic plasticity, also affected by excitotoxicity and cholinergic receptors that, in turns, act on learning (Waxman and Lynch, 2005; Hasselmo, 2006). Since in the Wong–Wang neural mass model J_{NMDA} is mostly related to slow synaptic mechanisms driven by NMDA receptors (Deco et al., 2014) and receptor density can be remapped onto TVB through parameterization (Deco et al., 2021), an assessment of these receptor-dependent properties could be an important development in future studies.

Study considerations

The small sample size can be seen as a potential limitation in the present study. However, the main aim of this investigation was to assess the ability of TVB to provide a personalized fingerprint of patients, potentially beyond known diagnosis. TVB modeling provides a set of physiological features at single subject level, otherwise not available from standard signal/image acquisition and analysis. Thus, the small sample size does not impact on the TVB ability of uncovering subject-specific features of FC, and E/I profile. The high correlation of TVB parameters with both cognitive performance and pharmacological treatment reveals indeed its exquisite sensitivity to single-subject profiles and opens a broad range of prospective for clinical applications. On the other hand, the application of TVB to a larger cohort of patients bears the potential of improving disease classification of disease subtypes, critical for treatment stratification and for establishing intervention workflows.

Conclusion

The present study demonstrates that brain networks can be characterized in terms of a meaningful set of mesoscale parameters at the single-subject level in humans *in vivo*. The identification of network abnormalities in patients may be used to design neuromodulation, neuropharmacological, and neuropsychological paradigms capable of regulating circuit function and plasticity (Lin and Wang, 2018), while the high correlation of TVB parameters with both cognitive performance and pharmacological treatment reveals an exquisite sensitivity to single-subject features. As a corollary, it should be remembered that the small sample size does not impact significantly on the TVB capacity of uncovering subject-specific connectivity strength, and E/I profile. At present, it is unclear whether network properties in this study are influenced by therapy suggesting that future studies should systematically address this issue. In aggregate, TVB parameters are shedding light on the changes occurring inside the brain networks of AD and FTD patients opening new perspectives for understanding disease mechanisms and for designing personalized neuromodulation, neuropharmacological and neuropsychological paradigms.

Data availability statement

All codes used for this study are open source. The optimized TVB C code can be found at https://github.com/BrainModes/fast_tvb. The dataset is available at https://zenodo.org/record/8113922.

Ethics statement

The studies involving human participants were reviewed and approved by the Local Ethical Committee of the IRCCS Mondino Foundation. The patients/participants provided their written informed consent to participate in this study.

Author contributions

MC and AC: patients' recruitment and clinical assessment. LF, AP, and LM: MRI recordings. FC: neuropsychological testing. AR, AM, FP, and MS: data analysis. AM, FP, MM, FA, ED'A, and CG: TVB modeling and simulation. CG and FP: MRI theory and protocol design. MS, VJ, and PR: TVB support. SC, AC, and MC: neurological support. AM, ED'A, and FP: manuscript writing. ED'A, CG, and FP: work coordination and manuscript finalization. All authors had contributed to manuscript discussion and approved the final version of the manuscript.

Funding

The work performed at the IRCCS Mondino Foundation was supported by the Italian Ministry of Health (RC2022-RC2024). The work performed at the University of Pavia was supported by H2020 Research and Innovation Action Grants Human Brain Project 785907 and 945539 (SGA2 and SGA3) to ED'A, FP, and PR. Moreover, the project was supported by the MNL Project "Local Neuronal Microcircuits" of the Centro Fermi (Rome, Italy), #NEXTGENERATIONEU (NGEU) and funded by the Ministry of University and Research (MUR), the National Recovery and Resilience Plan (NRRP), project IR00011-EBRAINS-Italy to ED'A; Horizon2020 [Research and Innovation Action Grants Human Brain Project 945539 (SGA3)], BRC (#BRC704/CAP/CGW), MRC (#MR/S026088/1), Ataxia UK to CW-K; PR acknowledges Digital Europe Grant TEF-Health #101100700; H2020 Research and Innovation Action Grant Human Brain Project (ICEI 800858, EOSC VirtualBrainCloud 82642, AISN 101057655); H2020 Research Infrastructures Grant (EBRAINS-PREP 101079717, EBRAIN-Health 101058516); H2020 European Innovation Council (PHRASE 101058240); H2020 European Research Council Grant (ERC BrainModes 683049); JPND ERA PerMed PatternCog 2522FSB904; Berlin Institute of Health and Foundation Charité; Johanna Quandt Excellence Initiative; and German Research Foundation (SFB 1436, project ID 425899996;

SFB 1315, project ID 327654276; SFB 936, project ID 178316478; and SFB-TRR 295, project ID 424778381).

Acknowledgments

This manuscript has been released as a Pre-Print at BioRxiv (Monteverdi et al., 2023).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnagi.2023.1204134/full#supplementary-material

References

Ades-Aron, B., Lemberskiy, G., Veraart, J., Golfinos, J., Fieremans, E., Novikov, D. S., et al. (2020). Improved task-based functional MRI language mapping in patients with brain tumors through marchenko-pastur principal component analysis denoising. *Radiology* 298, 365–373. doi: 10.1148/RADIOL.2020200822

Aerts, H., Schirner, M., Jeurissen, B., Van Roost, D., Achten, E., Ritter, P., et al. (2018). Modeling brain dynamics in brain tumor patients using the virtual brain. *eNeuro* 5:ENEURO.0083-18.2018. doi: 10.1523/ENEURO.0083-18.2018

Andersson, J. L. R., and Sotiropoulos, S. N. (2016). An integrated approach to correction for off-resonance effects and subject movement in diffusion MR imaging. *Neuroimage* 125, 1063–1078. doi: 10.1016/j.neuroimage.2015.10.019

Arnemann, K. L., Stöber, F., Narayan, S., Rabinovici, G. D., and Jagust, W. J. (2018). Metabolic brain networks in aging and preclinical Alzheimer's disease. *NeuroImage Clin.* 17, 987–999. doi: 10.1016/j.nicl.2017.12.037

Babiloni, C., Del Percio, C., Boccardi, M., Lizio, R., Lopez, S., Filippo, C., et al. (2015). Occipital sources of resting state alpha rhythms subjects with amnesic mild cognitive impairment and Alzheimer's disease. *Neurobiol. Aging* 36, 556–570. doi: 10.1016/j.neurobiolaging.2014.09.011

Battaglia, D., Boudou, T., Hansen, E. C. A., Lombardo, D., Chettouf, S., Daffertshofer, A., et al. (2020). Dynamic functional connectivity between order and randomness and its evolution across the human adult lifespan. *Neuroimage* 222:117156. doi: 10.1016/j.neuroimage.2020.117156

Benussi, A., Dell'Era, V., Cantoni, V., Cotelli, M. S., Cosseddu, M., Spallazzi, M., et al. (2020). TMS for staging and predicting functional decline in frontotemporal dementia. *Brain Stimul.* 13, 386–392. doi: 10.1016/j.brs.2019.11.009

Borst, A., and Theunissen, F. (1999). Information theory and neural networks. *North Holl. Math. Libr.* 51, 307–340. doi: 10.1016/S0924-6509(08)70042-4

Buckner, R. L., Krienen, F. M., Castellanos, A., Diaz, J. C., and Thomas Yeo, B. T. (2011). The organization of the human cerebellum estimated by intrinsic functional connectivity. *J. Neurophysiol.* 106, 2322–2345. doi: 10.1152/jn.00339.2011

Burrell, J. R., Kiernan, M. C., Vucic, S., and Hodges, J. R. (2011). Motor neuron dysfunction in frontotemporal dementia. *Brain* 134, 2582–2594. doi: 10.1093/brain/awr195

Cai, S., Chong, T., Peng, Y., Shen, W., Li, J., von Deneen, K. M., et al. (2017). Altered functional brain networks in amnestic mild cognitive impairment: A resting-state fMRI study. *Brain Imaging Behav.* 11, 619–631. doi: 10.1007/s11682-016-9539-0

Capitani, E., and Laiacona, M. (1997). Composite neuropsychological batteries and demographic correction: Standardization based on equivalent scores, with a review of published data. *J. Clin. Exp. Neuropsychol.* 19, 795–809. doi: 10.1080/01688639708403761

Castellazzi, G., Palesi, F., Casali, S., Vitali, P., Wheeler-Kingshott, C. A. M., Sinforiani, E., et al. (2014). A comprehensive assessment of resting state networks: Bidirectional modification of functional integrity in cerebro-cerebellar networks in dementia. *Front. Neurosci.* 8:223. doi: 10.3389/fnins.2014.00223

Chen, H., Pan, X., Lau, J. K. L., Bickerton, W. L., Pradeep, B., Taheri, M., et al. (2016). Lesion-symptom mapping of a complex figure copy task: A large-scale PCA study of the BCoS trial. *NeuroImage Clin.* 11, 622–634. doi: 10.1016/j.nicl.2016.04.007

D'Angelo, E. D., and Jirsa, V. (2022). The quest for multiscale brain modeling. *Trends Neurosci.* 45, 777–790. doi: 10.1016/j.tins.2022.06.007

Deco, G., Jirsa, V., and Friston, K. J. (2012). "The dynamical and structural basis of brain activity," in *Principles of brain dynamics: Global state interactions*, eds M. I. Rabinovich, K. J. Friston, and P. Varona (Cambridge, MA: MIT Press), doi: 10.7551/mitpress/9108.003.0003

Deco, G., Kringelbach, M. L., Arnatkeviciute, A., Oldham, S., Sabaroedin, K., Rogasch, N. C., et al. (2021). Dynamical consequences of regional heterogeneity in the brain's transcriptional landscape. *Sci. Adv.* 7:eabf4752. doi: 10.1126/sciadv.abf4752

Deco, G., Ponce-Alvarez, A., Hagmann, P., Romani, G. L., Mantini, D., and Corbetta, M. (2014). How local excitation-inhibition ratio impacts the whole brain dynamics. *J. Neurosci.* 34, 7886–7898. doi: 10.1523/JNEUROSCI.5068-13.2014

Diedrichsen, J., Balsters, J. H., Flavell, J., Cussans, E., and Ramnani, N. (2009). A probabilistic MR atlas of the human cerebellum. *Neuroimage* 46, 39–46. doi: 10.1016/j.neuroimage.2009.01.045

Dixon, M. L., De La Vega, A., Mills, C., Andrews-Hanna, J., Spreng, R. N., Cole, M. W., et al. (2018). Heterogeneity within the frontoparietal control network and its relationship to the default and dorsal attention networks. *Proc. Natl. Acad. Sci. U.S.A.* 115, E1598–E1607. doi: 10.1073/pnas.1715766115

Dubois, B., Feldman, H. H., Jacova, C., Hampel, H., Molinuevo, J. L., Blennow, K., et al. (2014). Advancing research diagnostic criteria for Alzheimer's disease: The IWG-2 criteria. *Lancet Neurol.* 13, 614–629. doi: 10.1016/S1474-4422(14)70090-0

Epelbaum, S., Bouteloup, V., Mangin, J. F., La Corte, V., Migliaccio, R., Bertin, H., et al. (2018). Neural correlates of episodic memory in the Memento cohort. *Alzheimers Dement.* 4, 224–233. doi: 10.1016/j.trci.2018.03.010

Falcon, M. I., Riley, J. D., Jirsa, V., McIntosh, A. R., Chen, E. E., and Solodkin, A. (2016). Functional mechanisms of recovery after chronic stroke: Modeling with the virtual brain. *eNeuro* 3:ENEURO.0158-15.2016. doi: 10.1523/ENEURO.0158-15.2016

Ferreri, F., Guerra, A., Vollero, L., Ponzo, D., Määttä, S., Könönen, M., et al. (2021). TMS-EEG biomarkers of amnestic mild cognitive impairment due to Alzheimer's disease: A proof-of-concept six years prospective study. *Front. Aging Neurosci.* 13:737281. doi: 10.3389/fnagi.2021.737281

Friston, K. J., Williams, S., Howard, R., Frackowiak, R. S. J., and Turner, R. (1996). Movement-related effects in fMRI time-series. *Magn. Reson. Med.* 35, 346–355. doi: 10.1002/mrm.1910350312

Good, T., Schirner, M., Shen, K., Ritter, P., Mukherjee, P., Levine, B., et al. (2022). Personalized connectome-based modeling in patients with semi-acute phase TBI: Relationship to acute neuroimaging and 6 month follow-up. *eNeuro* 9:ENEURO.0075-21.2022. doi: 10.1523/ENEURO.0075-21.2022

Gorno-Tempini, M. L., Hillis, A. E., Weintraub, S., Kertesz, A., Mendez, M., Cappa, S. F., et al. (2011). Classification of primary progressive aphasia and its variants. *Neurology* 76, 1006–1014. doi: 10.1212/WNL.0b013e31821103e6

Graff-Radford, J., Yong, K. X. X., Apostolova, L. G., Bouwman, F. H., Carrillo, M., Dickerson, B. C., et al. (2021). New insights into atypical Alzheimer's disease in the era of biomarkers. *Lancet Neurol.* 20, 222–234. doi: 10.1016/S1474-4422(20)30440-3

Hansen, E. C. A., Battaglia, D., Spiegler, A., Deco, G., and Jirsa, V. K. (2015). Functional connectivity dynamics: Modeling the switching behavior of the resting state. *Neuroimage* 105, 525–535. doi: 10.1016/j.neuroimage.2014.11.001

Hassan, A., Whitwell, J. L., and Josephs, K. A. (2011). The corticobasal syndrome-Alzheimer's disease conundrum. *Expert. Rev. Neurother.* 11, 1569–1578. doi: 10.1586/ern.11.153

Hasselmo, M. E. (2006). The role of acetylcholine in learning and memory. *Curr. Opin. Neurobiol.* 16, 710–715. doi: 10.1016/j.conb.2006.09.002.The

Hohenfeld, C., Werner, C. J., and Reetz, K. (2018). Resting-state connectivity in neurodegenerative disorders: Is there potential for an imaging biomarker? *NeuroImage Clin.* 18, 849–870. doi: 10.1016/j.nicl.2018.03.013

Jiménez-Balado, J., and Eich, T. S. (2021). GABAergic dysfunction, neural network hyperactivity and memory impairments in human aging and Alzheimer's disease. *Semin. Cell Dev. Biol.* 116, 146–159. doi: 10.1016/j.semcdb.2021.01.005

Jirsa, V. K., Proix, T., Perdikis, D., Woodman, M. M., Wang, H., Bernard, C., et al. (2017). The Virtual Epileptic Patient: Individualized whole-brain models of epilepsy spread. *Neuroimage* 145, 377–388. doi: 10.1016/j.neuroimage.2016.04.049

Jockwitz, C., Caspers, S., Lux, S., Jütten, K., Schleicher, A., Eickhoff, S. B., et al. (2017). Age- and function-related regional changes in cortical folding of the default mode network in older adults. *Brain Struct. Funct.* 222, 83–99. doi: 10.1007/s00429-016-1202-4

Jones, D. T., Knopman, D. S., Gunter, J. L., Graff-Radford, J., Vemuri, P., Boeve, B. F., et al. (2016). Cascading network failure across the Alzheimer's disease spectrum. *Brain* 139, 547–562. doi: 10.1093/brain/awv338

Kong, X., Kong, R., Orban, C., Wang, P., Zhang, S., Anderson, K., et al. (2021). Sensory-motor cortices shape functional connectivity dynamics in the human brain. *Nat. Commun.* 12:6373. doi: 10.1038/s41467-021-26704-y

Kucewicz, M. T., Saboo, K., Berry, B. M., Kremen, V., Miller, L. R., Khadjevand, F., et al. (2019). Human verbal memory encoding is hierarchically distributed in a continuous processing stream. *eNeuro* 6:ENEURO.0214-18.2018. doi: 10.1523/ENEURO.0214-18.2018

Lauterborn, J. C., Scaduto, P., Cox, C. D., Schulmann, A., Lynch, G., Gall, C. M., et al. (2021). Increased excitatory to inhibitory synaptic ratio in parietal cortex samples from individuals with Alzheimer's disease. *Nat. Commun.* 12:2603. doi: 10.1038/s41467-021-22742-8

Leyton, C. E., and Hodges, J. R. (2010). Frontotemporal dementias: Recent advances and current controversies. *Ann. Indian Acad. Neurol.* 13, S74–S80. doi: 10.4103/0972-2327.74249

Li, L., Su, Y. A., Wu, Y. K., Castellanos, F. X., Li, K., Li, J. T., et al. (2021). Eight-week antidepressant treatment reduces functional connectivity in first-episode drug-naïve patients with major depressive disorder. *Hum. Brain Mapp.* 42, 2593–2605. doi: 10.1002/hbm.25391

Lin, Y. C., and Wang, Y. P. (2018). Status of noninvasive brain stimulation in the therapy of Alzheimer's disease. *Chin. Med. J.* 131, 2899–2903. doi: 10.4103/0366-6999.247217

Lorenzi, R. M., Palesi, F., Castellazzi, G., Vitali, P., Anzalone, N., Bernini, S., et al. (2020). Unsuspected involvement of spinal cord in Alzheimer disease. *Front. Cell. Neurosci.* 14:6. doi: 10.3389/fncel.2020.00006

Maestú, F., de Haan, W., Busche, M. A., and DeFelipe, J. (2021). Neuronal excitation/inhibition imbalance: Core element of a translational perspective on Alzheimer pathophysiology. *Ageing Res. Rev.* 69:101372. doi: 10.1016/j.arr.2021.101372

Marti-Juan, G., Sastre-Garriga, J., Vidal-Jordana, A., Llufriu, S., Martinez-Heras, E., Groppa, S., et al. (2022). Using the virtual brain to study the relationship between structural and functional connectivity in people with multiple sclerosis: A multicentre study. *Mult. Scler. J.* 28, 262–264.

Marucci, G., Buccioni, M., Ben, D. D., Lambertucci, C., Volpini, R., and Amenta, F. (2021). Efficacy of acetylcholinesterase inhibitors in Alzheimer's disease. *Neuropharmacology* 190:108352. doi: 10.1016/j.neuropharm.2020.108352

Maul, S., Giegling, I., and Rujescu, D. (2020). Proton magnetic resonance spectroscopy in common dementias–current status and perspectives. *Front. Psychiatry* 11:769. doi: 10.3389/fpsyt.2020.00769

McKeith, I. G., Boeve, B. F., Dickson, D. W., Halliday, G., Aarsland, D., Attems, J., et al. (2017). Diagnosis and management of dementia with Lewy bodies: Fourth consensus report of the DLB Consortium. *Neurology* 89, 88–100.

Melrose, R. J., Harwood, D., Khoo, T., Mandelkern, M., and Sultzer, D. (2013). Association between cerebral metabolism and Rey-Osterrieth Complex Figure Test performance in ALzheimer's disease. *J. Clin. Exp. Neuropsychol.* 35, 246–258. doi: 10.1080/13803395.2012.763113.Association

Monteverdi, A., Palesi, F., Costa, A., Vitali, P., Pichiecchio, A., Cotta Ramusino, M., et al. (2022). Subject-specific features of excitation/inhibition profiles in neurodegenerative diseases. *Front. Aging Neurosci.* 14:868342. doi: 10.3389/fnagi.2022.868342

Monteverdi, A., Palesi, F., Schirner, M., Argentino, F., Merante, M., Redolfi, A., et al. (2023). Virtual brain simulations reveal network-specific parameters in neurodegenerative dementias. *bioRxiv* [Preprint]. doi: 10.1101/2023.03.10.532087

Murley, A. G., Rouse, M. A., Simon Jones, P., Ye, R., Hezemans, F. H., O'Callaghan, C., et al. (2021). GABA and glutamate deficits from frontotemporal lobar degeneration are associated with disinhibition. *Brain* 143, 3449–3462. doi: 10.1093/BRAIN/AWAA305

Muschelli, J., Beth Nebel, M., Caffo, B. S., Barber, A. D., Pekar, J. J., and Mostofsky, S. H. (2014). Reduction of motion-related artifacts in resting state fMRI using aCompCor. *Neuroimage* 96, 22–35. doi: 10.1016/j.neuroimage.2014.03.028

Nigri, A., Ferraro, S., Gandini Wheeler-Kingshott, C. A. M., Tosetti, M., Redolfi, A., Forloni, G., et al. (2022). Quantitative MRI harmonization to maximize clinical impact: The RIN–neuroimaging network. *Front. Neurol.* 13:855125. doi: 10.3389/fneur.2022.855125

Palesi, F., Lorenzi, R. M., Casellato, C., Ritter, P., Jirsa, V., Gandini Wheeler-Kingshott, C. A. M., et al. (2020). The importance of cerebellar connectivity on simulated brain dynamics. *Front. Cell. Neurosci.* 14:240. doi: 10.3389/fncel.2020.00240

Palop, J. J., and Mucke, L. (2016). Network abnormalities and interneuron dysfunction in Alzheimer disease. *Nat. Rev. Neurosci.* 17, 777–792. doi: 10.1038/nrn.2016.141

Patenaude, B., Smith, S. M., Kennedy, D. N., and Jenkinson, M. (2011). A Bayesian model of shape and appearance for subcortical brain segmentation. *Neuroimage* 56, 907–922. doi: 10.1016/j.neuroimage.2011.02.046

Pek, J., Wong, O., and Wong, A. C. M. (2018). How to address non-normality: A taxonomy of approaches, reviewed, and illustrated. *Front. Psychol.* 9:2104. doi: 10.3389/fpsyg.2018.02104

Pérez, P. D., Ma, Z., Hamilton, C., Sánchez, C., Mørk, A., Pehrson, A. L., et al. (2018). Acute effects of vortioxetine and duloxetine on resting-state functional connectivity in the awake rat. *Neuropharmacology* 128, 379–387. doi: 10.1016/j.neuropharm.2017.10.038

Pini, L., Pizzini, F. B., Boscolo-Galazzo, I., Ferrari, C., Galluzzi, S., Cotelli, M., et al. (2022). Brain network modulation in Alzheimer's and frontotemporal dementia with transcranial electrical stimulation. *Neurobiol. Aging* 111, 24–34. doi: 10.1016/j.neurobiolaging.2021.11.005

Pini, L., Wennberg, A. M., Salvalaggio, A., Vallesi, A., Pievani, M., and Corbetta, M. (2021). Breakdown of specific functional brain networks in clinical variants of Alzheimer's disease. *Ageing Res. Rev.* 72:101482. doi: 10.1016/j.arr.2021.101482

Rascovsky, K., Hodges, J. R., Knopman, D., Mendez, M. F., Kramer, J. H., Neuhaus, J., et al. (2011). Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia. *Brain* 134, 2456–2477. doi: 10.1093/brain/awr179

Rauchmann, B. S., Ersoezlue, E., Stoecklein, S., Keeser, D., Brosseron, F., Buerger, K., et al. (2021). Resting-state network alterations differ between Alzheimer's disease atrophy subtypes. *Cereb. Cortex* 31, 4901–4915. doi: 10.1093/cercor/bhab130

Redolfi, A., De Francesco, S., Palesi, F., Galluzzi, S., Muscio, C., Castellazzi, G., et al. (2020). Medical informatics platform (MIP): A pilot study across clinical Italian cohorts. *Front. Neurol.* 11:1021. doi: 10.3389/fneur.2020.01021

Reyes, P., Ortega-Merchan, M. P., Rueda, A., Uriza, F., Santamaria-García, H., Rojas-Serrano, N., et al. (2018). Functional connectivity changes in behavioral, semantic, and nonfluent variants of frontotemporal dementia. *Behav. Neurol.* 2018:9684129. doi: 10.1155/2018/9684129

Robinson, D. M., and Keating, G. M. (2006). Memantine: a review of its use in Alzheimer's disease. *Drugs* 66, 1515–1534. doi: 10.2165/00003495-200666110-00015

Russell, L. L., Greaves, C. V., Convery, R. S., Bocchetta, M., Warren, J. D., Kaski, D., et al. (2021). Eye movements in frontotemporal dementia: Abnormalities of fixation, saccades and anti-saccades. *Alzheimers Dement.* 7:e12218. doi: 10.1002/trc2.12218

Schirner, M., Domide, L., Perdikis, D., Triebkorn, P., Stefanovski, L., Pai, R., et al. (2022). Brain simulation as a cloud service: The Virtual Brain on EBRAINS. *Neuroimage* 251:118973. doi: 10.1016/j.neuroimage.2022.118973

Smith, R. E., Tournier, J. D., Calamante, F., and Connelly, A. (2012). Anatomically-constrained tractography: Improved diffusion MRI streamlines tractography through effective use of anatomical information. *Neuroimage* 62, 1924–1938. doi: 10.1016/j.neuroimage.2012.06.005

Smitha, K. A., Akhil Raja, K., Arun, K. M., Rajesh, P. G., Thomas, B., Kapilamoorthy, T. R., et al. (2017). Resting state fMRI: A review on methods in resting state connectivity analysis and resting state networks. *Neuroradiol. J.* 30, 305–317. doi: 10.1177/1971400917697342

Spurny, B., Vanicek, T., Seiger, R., Reed, M. B., Klöbl, M., Ritter, V., et al. (2021). Effects of SSRI treatment on GABA and glutamate levels in an associative relearning paradigm. *Neuroimage* 232:117913. doi: 10.1016/j.neuroimage.2021.117913

Stefanovski, L., Meier, J. M., Pai, R. K., Triebkorn, P., Lett, T., Martin, L., et al. (2021). Bridging scales in Alzheimer's disease: Biological framework for brain simulation with the virtual brain. *Front. Neuroinform.* 15:630172. doi: 10.3389/fninf.2021.630172

Stefanovski, L., Triebkorn, P., Spiegler, A., Diaz-Cortes, M. A., Solodkin, A., Jirsa, V., et al. (2019). Linking molecular pathways and large-scale computational modeling to assess candidate disease mechanisms and pharmacodynamics in Alzheimer's disease. *Front. Comput. Neurosci.* 13:54. doi: 10.3389/fncom.2019.00054

Stephan, K. E., Weiskopf, N., Drysdale, P. M., Robinson, P. A., and Friston, K. J. (2007). Comparing hemodynamic models with DCM. *Neuroimage* 38, 387–401. doi: 10.1016/j.neuroimage.2007.07.040

Thomas Yeo, B. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., et al. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J. Neurophysiol.* 106, 1125–1165. doi: 10.1152/jn.00338.2011

Titov, D., Diehl-Schmid, J., Shi, K., Perneczky, R., Zou, N., Grimmer, T., et al. (2017). Metabolic connectivity for differential diagnosis of dementing disorders. *J. Cereb. Blood Flow Metab.* 37, 252–262. doi: 10.1177/0271678X15622465

Tok, S., Ahnaou, A., and Drinkenburg, W. (2021). Functional neurophysiological biomarkers of early-stage Alzheimer's disease: A perspective of network hyperexcitability in disease progression. *J. Alzheimer's Dis.* 88, 809–836. doi: 10.3233/jad-210397

Tournier, J. D., Calamante, F., and Connelly, A. (2012). MRtrix: Diffusion tractography in crossing fiber regions. *Int. J. Imaging Syst. Technol.* 22, 53–66. doi: 10.1002/ima.22005

Triebkorn, P., Stefanovski, L., Dhindsa, K., Diaz-Cortes, M., Bey, P., Bülau, K., et al. (2022). Brain simulation augments machine-learning–based classification of dementia. *Alzheimers Dement.* 8:e12303. doi: 10.1002/trc2.12303

Van Wingen, G. A., Tendolkar, I., Urner, M., van Marle, H. J., Denys, D., Verkes, R. J., et al. (2014). Short-term antidepressant administration reduces default mode and task-positive network connectivity in healthy individuals during rest. *Neuroimage* 88, 47–53. doi: 10.1016/j.neuroimage.2013.11.022

Vonk, J. M. J., Rizvi, B., Lao, P. J., Budge, M., Manly, J. J., Mayeux, R., et al. (2019). Letter and category fluency performance correlates with distinct patterns of cortical thickness in older adults. *Cereb. Cortex* 29, 2694–2700. doi: 10.1093/cercor/bhy138

Waxman, E. A., and Lynch, D. R. (2005). N-methyl-D-aspartate receptor subtypes: Multiple roles in excitotoxicity and neurological disease. *Neuroscientist* 11, 37–49. doi: 10.1177/1073858404269012

Zhou, J., Greicius, M. D., Gennatas, E. D., Growdon, M. E., Jang, J. Y., Rabinovici, G. D., et al. (2010). Divergent network connectivity changes in behavioural variant frontotemporal dementia and Alzheimer's disease. *Brain* 133, 1352–1367. doi: 10.1093/brain/awq075

Zimmermann, J., Perry, A., Breakspear, M., Schirner, M., Sachdev, P., Wen, W., et al. (2018). Differentiation of Alzheimer's disease based on local and global parameters in personalized Virtual Brain models. *NeuroImage Clin.* 19, 240–251. doi: 10.1016/j.nicl.2018.04.017

frontiers | Frontiers in Computational Neuroscience

TYPE Original Research
PUBLISHED 29 September 2023
DOI 10.3389/fncom.2023.1211096

OPEN ACCESS

EDITED BY
Noemi Montobbio,
University of Genoa, Italy

REVIEWED BY
Enrico Capobianco,
Jackson Laboratory, United States
Ilya Pyatnitskiy,
The University of Texas at Austin, United States

*CORRESPONDENCE
Xiaoling Peng
✉ xlpeng@uic.edu.cn

RECEIVED 24 April 2023
ACCEPTED 13 September 2023
PUBLISHED 29 September 2023

CITATION
Zheng S, Zhang X, Song P, Hu Y, Gong X and Peng X (2023) Complexity-based graph convolutional neural network for epilepsy diagnosis in normal, acute, and chronic stages. *Front. Comput. Neurosci.* 17:1211096. doi: 10.3389/fncom.2023.1211096

Complexity-based graph convolutional neural network for epilepsy diagnosis in normal, acute, and chronic stages

Shiming Zheng[1], Xiaopei Zhang[1], Panpan Song[2], Yue Hu[2], Xi Gong[1] and Xiaoling Peng[1]*

[1]Guangdong Provincial Key Laboratory of Interdisciplinary Research and Application for Data Science, BNU-HKBU United International College, Zhuhai, China, [2]Department of Neurology, Children's Hospital of Chongqing Medical University, Chongqing, China

Introduction: The automatic precision detection technology based on electroencephalography (EEG) is essential in epilepsy studies. It can provide objective proof for epilepsy diagnosis, treatment, and evaluation, thus helping doctors improve treatment efficiency. At present, the normal and acute phases of epilepsy can be well identified through EEG analysis, but distinguishing between the normal and chronic phases is still tricky.

Methods: In this paper, five popular complexity indicators of EEG signal, including approximate entropy, sample entropy, permutation entropy, fuzzy entropy and Kolmogorov complexity, are computed from rat hippocampi to characterize the normal, acute, and chronic phases during epileptogenesis. Results of one-way ANOVA and principal component analysis both show that utilizing complexity features, we are able to easily identify differences between normal, acute, and chronic phases. We also propose an innovative framework for epilepsy detection based on graph convolutional neural network (GCNN) using multi-channel EEG complexity as input.

Results: Combining information of five complexity measures at eight channels, our GCNN model demonstrate superior ability in recognizing the normal, acute, and chronic phases. Experiments results show that our GCNN model reached the high prediction accuracy above 98% and F1 score above 97% among these three phases for each individual rat.

Discussion: Our research practice based on real data shows that EEG complexity characteristics are of great significance for recognizing different stages of epilepsy.

KEYWORDS

EEG complexity measures, entropy, graph convolutional neural network, epilepsy diagnosis, chronic stage

1. Introduction

Epilepsy is a neurological disorder defined as a transient occurrence of clinical features produced by abnormal excessive or synchronous neuronal (Fisher et al., 2005). Worldwide, more than 50 million people have epilepsy, affecting humans of all ages, ethnicity, and society. It has been classified as one of the most highly challenging neural psychiatric diseases that the World Health Organization (WHO) focuses on prevention and treatment (Saxena and Li, 2017). Epilepsy is characterized by recurrent seizures caused by abnormal discharge of brain neurons and an ongoing predisposition to recurrent seizures. The patients with

epilepsy mainly include those with reflex seizures and those with more than one unprovoked seizure after 24 h. In particular, compared to the general population, the probability of having recurrent seizures in the next 10 years for epileptic patients who have had a single seizure is at least 60% (Fisher et al., 2005). Therefore, the diagnosis and treatment of epilepsy are of great significance for humans, while accurate prediction of epileptic seizures is crucial for achieving precision treatments in epilepsy. The rat pilocarpine (PILO) model of temporal lobe epilepsy (TLE) is an animal model in which central cholinergic receptors are activated to induce seizures by pilocarpine, a post-ganglionic cholinergic drug that can produce quasi-cholinergic effects by directly exciting M-cholinergic receptors (Song et al., 2016). Since the damage and indications of the rat PILO model are comparable to those of human TLE, it is a widely used animal epilepsy model of TLE. This model exhibits three important phases (Song et al., 2016): (1) the normal phase—1 day before status epilepticus (SE), (2) the acute phase—the duration of SE and 6–24 h after SE, and (3) the chronic phase—marked by occurrences of spontaneous recurrent seizures (SRS) after SE.

As one of the most potent and economical tools to record and monitor the brain's electrical activity, in recent years, electroencephalogram (EEG) analysis has become a hot topic in epilepsy diagnosis, and related studies for both doctors and researchers (Karlócai et al., 2011). Analyzing EEG recordings can provide an objective reference for diagnosing epilepsy-related diseases, such as the identification, prediction, focus location, or treatment evaluation of epilepsy (Karlócai et al., 2011). Various features extracted from EEG signals play essential roles in disease diagnosis as they can help researchers to describe the characteristics and mechanism of epileptic seizures. Basically, EEG signal features are divided into four categories. Time-domain features analyze how signal changes with time (Srinivasan et al., 2005; Sharmila and Geethanjali, 2018; Wei et al., 2019), frequency-domain features depict how signal lies within each frequency band (Srinivasan et al., 2005; Faust et al., 2010; Wen and Zhang, 2017), time-frequency domain features are characteristics consider both time and frequency domain (Tzallas et al., 2009; Wang et al., 2017), while nonlinear features regard the brain as a system to describe its complexity and the amount of information (Yuan et al., 2011; Li et al., 2017; Wang et al., 2017). Many previous studies have made significant progress in epilepsy detection based on one or more of these EEG signal features (Boonyakitanont et al., 2020). Since EEG signal shows non-stationary and nonlinear dynamic behavior when measuring the electrical activity of a brain (Natarajan et al., 2004), EEG signal features based on nonlinear dynamic properties may be better than the other three types of features in mining and detecting the regular changes of EEG in different stages of epileptogenesis. Recently, more and more researchers treated the dynamic changes of brain activity as a complex nonlinear system to study their complexity. Thus, some nonlinear complexity measures, especially various entropy indices, have attracted the great attention of researchers through their outperformance in characterizing EEG signals by quantifying the complexity and amount of information (Liang et al., 2015).

Most early studies achieved good performance for applying complexity measures and one or more classifiers to distinguish different stages of epilepsy by analyzing EEG signals. Sharma et al. (2014) built epileptic seizure detection models based on four complexity measures, including Shannon entropy, Renyi entropy, approximate entropy (ApEn), and sample entropy (SampEn), to classify the EEG signals during focal and non-focal epilepsy and achieved 87% accuracy by the least squares support vector machine (LS-SVM) classifier. To achieve auto-detection of focal and non-focal EEG recordings, Arunkumar et al. (2017) yielded the highest accuracy of 98% by feeding five different entropy features to the non-nested generalized exemplars (NNge) classifier after comparing with other four different classifiers, including naıve bayes classifier (NBC), radial basis function (RBF), support vector machines (SVM), and k nearest neighbor (KNN). Xiang et al. (2015) trained SVM using fuzzy entropy (FuzzEn) to detect epileptic seizures from normal groups and reached a detection rate of 98.31 and 100% on two different datasets, respectively.

However, most of these notable results were obtained from distinguishing epileptic EEG signals in the acute stage of epilepsy from normal. The study on EEG characteristics in the chronic stage has seldom been mentioned. Due to the fact that epilepsy patients are mostly in the chronic phase rather than the acute phase, identifying the chronic phase of epilepsy is particularly important for the timely diagnosis and treatment of epilepsy. It is beneficial to study and predict the chronic phase of epilepsy: (1) the pathophysiological mechanism of epilepsy and the effects and side effects of long-term medication in epileptic patients can be better understood; (2) and chronic seizures of epilepsy patients can be intervened and treated in advance. Hence, the primary motivation behind this work is to clarify the role of the complexity measures of EEG signals during acute and chronic seizures from normal groups. Further, it has been observed that most studies used traditional machine learning algorithms, such as SVM, Decision Tree, and KNN, to implement the classification tasks. Due to the simplistic structure of these conventional machine learning algorithms, only a single channel of EEG signals can be considered in the classification tasks. Nevertheless, multi-channel EEG is widely used for diagnosis and therapy in clinical practice because brain diseases are rarely limited to a specific region (Bullmore and Sporns, 2009). This prompted us to consider an advanced classifier that can integrate multi-channel EEG for epileptic detection.

Graph convolutional neural network (GCNN) is a deep neural network classification model capable of handling multichannel EEG signal analysis (Craley et al., 2022). It is an improvement of convolutional neural networks (CNN) and can preserve richer connection information than 2D or 3D matrices by considering EEG signals to be nodes in a topological graph and representing the relationships between them using edges (Lian et al., 2020). GCNN can describe the internal relationship between different graph's nodes, therefore providing a way to explore the relationship among multiple EEG channels in the EEG-based classification (Song et al., 2018). Thus, in recent years, GCNN has been applied and made an enormous impact on EEG-based recognition, including emotion recognition (Zhang et al., 2019), neurological disease diagnosis (Wagh and Varatharajah, 2020), sleep stage classification (Jia et al., 2020), epilepsy diagnosis (Covert et al., 2019; Li and Jung, 2021), and brain motor imagery (Hou et al., 2022).

In this paper, we developed an automatic epileptic detection system via GCNN using five complexity measures of EEG,

TABLE 1 Electrode coordinates for areas of interest in the rat PILO model of TLE during epileptogenesis.

Names of parts	Coordinates
Cornu ammonis 1 (CA1)	AP: 3.3–3.7 mm from Bregma, ML: 2.0–3.0 mm, and DV: 3.0–3.5 mm from the surface of neocortex
Cornu ammonis 3 (CA3)	AP: 3.3 mm, ML: 3.5–3.7 mm, and DV: 3.0–3.5 mm
The surface of neocortex of the bilateral parietal lobe (Reference Electrode)	AP: 7.0 mm, ML: 6.0 mm
Dentate gyrus (DG)	AP: 5.6 mm, ML: 4.0 mm, and DV: 6.0 mm

including approximate entropy, sample entropy, permutation entropy, fuzzy entropy, and Kolmogorov complexity to monitor dynamic changes and distinguish EEG recordings among normal, acute, and chronic stage of epilepsy. Statistically significant indicators are useful in indicating the difference between chronic and normal stages, prompting doctors to intervene in advance.

2. Materials and methods

2.1. EEG recordings

The experimental data used in this paper was from a previous study (Song et al., 2016), in which the rat PILO model of TLE is used in this experiment (Song et al., 2016). In particular, the subject rats were injected with pilocarpine to induce seizures and were stopped by utilizing diazepam. The EEG signals were recorded during the experiment by drilling holes in the skull at specific locations and implanting microelectrodes. The coordinates for particular sites of interest in the hippocampus in our study are shown in Table 1.

According to Song et al. (2016), each EEG recording has around 600,000 sampling points (10 min), and the original dataset could be mainly divided into six stages, including normal (1 day before SE), pre-seizure (30, 20, and 10 min before SE), acute [10 min after SE, 10 min before, and after utilizing diazepam (i.e., DZP injection)], stable (1, 2, and 3 h after the diazepam), latent (1, 3, and 7 days after SE), chronic (7, 14, and 28 days after SE) stages. Figure 1 describes and compares the 1 s waveforms (250–500 Hz) selected randomly from normal, acute, and chronic phases for representative rat (no.16) in channel CA1(R). Intuitively looking from Figure 1, the EEG of the acute phase is far from that of the normal and the chronic phases, with much wider amplitude and some typical waveform, while the difference between the normal phase and the chronic phase is not obvious.

2.2. Complexity measures

Five complexity metrics, including ApEn, SampEn, FuzzEn, PE, and KC, have been computed to quantify the dynamic changes of EEG signals during different stages of epileptogenesis. A brief introduction to these metrics is given in this section.

2.2.1. Approximate entropy

Approximate Entropy (ApEn) was proposed by Pincus et al. (1991) from the perspective of measuring the complexity of signal. It is a non-linear dynamic measure that quantifies the incidence of new information in the time series (Pincus et al., 1991). The higher the probability of a new pattern being generated in this time series, the higher the complexity of the sequence and the higher the corresponding ApEn value.

The calculation of ApEn is calculating the degree of self-similarity of a time series, that is, the difference between the probability of mutual approximation of m points adjacent to the sequence and the probability of mutual approximation of $m + 1$ points. Compared with the statistical characteristics such as mean and variance, ApEn can better reflect the characteristics of signal sequence in structural distribution.

2.2.2. Sample entropy

In order to reduce the estimation bias in the calculation of ApEn by comparing it to its own data segment, Sample Entropy (SampEn) was proposed by Richman and Moorman (2000). Different from ApEn, SampEn eliminates self-matches in the algorithm and computes the difference of logarithms of the probabilities. Therefore, SampEn is more accurate, more consistent, and not sensitive to the missing values.

2.2.3. Permutation entropy

Proposed by Bandt and Pompe (2002), Permutation Entropy (PE) provides a quantification measure of the complexity of a time series by capturing the order relations between reconstructed subsequences. Computed from the extracted probability distribution of the ordinal patterns (Henry and Judge, 2019), the value of PE may account for the temporal ordering structure (time causality) of a given time series. The PE approach is robust to noise, computationally efficient, and invariant with respect to non-linear monotonic transformations of the data.

2.2.4. Fuzzy entropy

Inspired by the concept of fuzzy set (Zadeh et al., 1996), Chen et al. (2007) proposed a new measure of complexity for time series in 2007, called Fuzzy Entropy (FuzzEn). Modified from ApEn and SampEn, but unlike them, FuzzEn measures the similarity of two vectors based on the idea of "fuzzy." That is, the similarity is no longer 1 or 0 determined by a single threshold but a fuzzy membership function, thereby blurring the similarity measure.

2.2.5. Kolmogorov complexity

As an early complexity measure, Kolmogorov Complexity (KC) was first proposed by Solomonoff (1960) and then developed by Chaitin (1977). According to Li and Vitányi (2008), for a given string or sequence, KC is defined as the size of the smallest program that is needed to generate that string. It was also known as "algorithmic complexity," "Kolmogorov-Chaitin complexity," "shortest program length," etc. Unlike Shannon's information

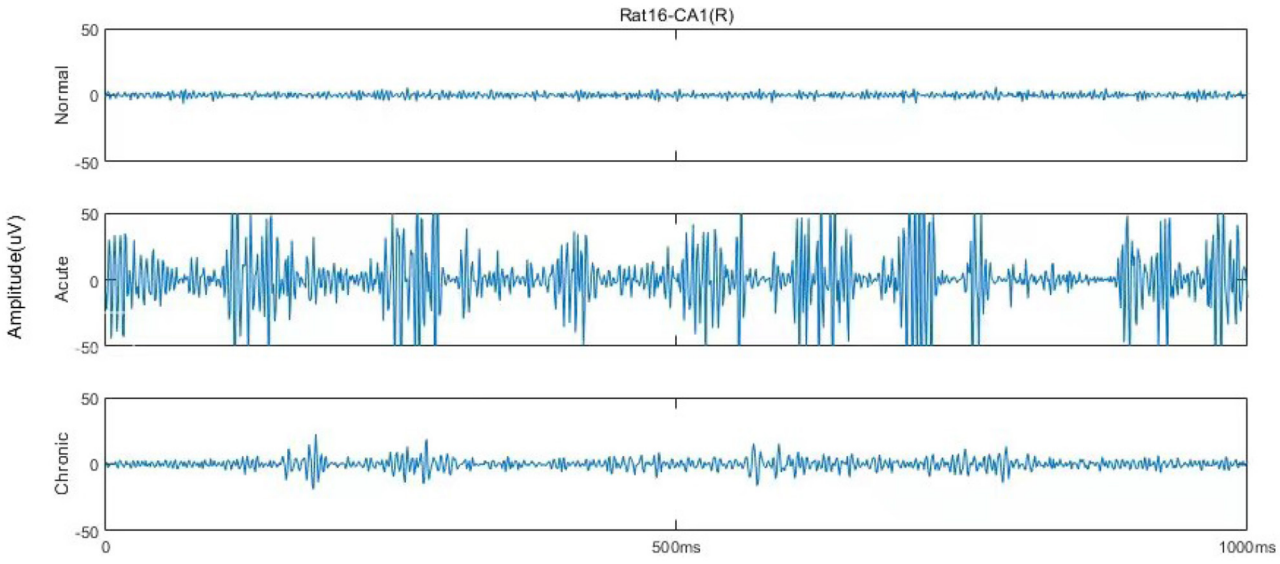


FIGURE 1
One second EEG waveforms in normal, acute, and chronic phases from rat no.16 in channel CA1(R), 250–500 Hz.

theory, KC is a measure of randomness or irregularity of individual objects rather than the average information of a random source.

2.3. Classification

In order to integrate all these complexity metrics at different channels, in this section, a GCNN-based classification framework is proposed and implemented to automatically identify and detect the acute and chronic stages of epilepsy.

2.3.1. Graph convolutional neural network (GCNN)

Our automatic epileptic detection system is built on GCNN proposed by Defferrard et al. (2016). GCNN is an extension framework that combines classical convolutional neural networks (CNN) and spectrum theory. Three main steps are involved to generalize CNNs to graphs, including designing the localized convolutional filters on graphs, clustering the similar vertices, and transforming spatial resolution for higher filter resolution (Defferrard et al., 2016). Thus, in addition to retaining the advantages of CNN, GCNN can deal with homogeneous and heterogeneous data (Such et al., 2017). In particular, it is capable of extracting features from unstructured data, such as graph representations, by performing convolutions on graph signals (Raeisi et al., 2022). Meanwhile, using graph as the input, GCNN provides a useful tool for processing signals from multiple channels simultaneously. Figure 2 shows a flow diagram of this automatic epileptic detection system for distinguishing EEG signals during the acute or chronic stage of epilepsy from normal.

2.3.1.1. Graph construction

As presented in Figure 2A, the inputs of our GCNN classifier are constructed on graphs with complexity measures. After collecting and preprocessing the 10-min 8-channel EEG as mentioned in Section 2.1, five complexity characteristics were extracted from each 1s-epoch EEG of each channel. To construct graphs, the sets of features are organized as a matrix. In particular, each feature matrix for a 1s-epoch EEG has eight rows and five columns, representing five extracted features at eight channels. Then, graphs representing five kinds of complexity at eight channels were generated and labeled with their specific stage (i.e., normal/acute/chronic). In this case, we notice that the connectivity pattern between channels may exist some kind of similarity in three stages of epilepsy. Therefore, to reduce potential interference due to this continuity between the three different stages, we construct each complete graph with eight nodes and all edges equal to 1, as the input to GCNN.

2.3.1.2. GCNN classification model

To achieve epileptic detection tasks, the constructed graphs were inputted to the classifier for training and validation to find the best GCNN model in identifying the specific stages (i.e., normal/acute/chronic) of current EEG fragments. As presented in Figure 2B, this GCNN network comprises two graph convolution blocks, two fully connected (FC) layers, and a softmax output layer. Each convolution block consists of a graph convolution layer, a max-pooling layer, and a Rectified Linear Unit (ReLU) active function. Specifically, the purpose of the convolution layer is to capture the features from the input graphs and learn the features that would be useful for the classification tasks. The max-pooling layer is a down-sample operation, which reduces the computation and avoids overfitting by decreasing the number of parameters to learn. Afterward, the ReLu layer will replace the input with zero if it is negative; otherwise, it will retain the original value. It is expressed by:

$$ReLu(x) = max(0, x), \quad (1)$$

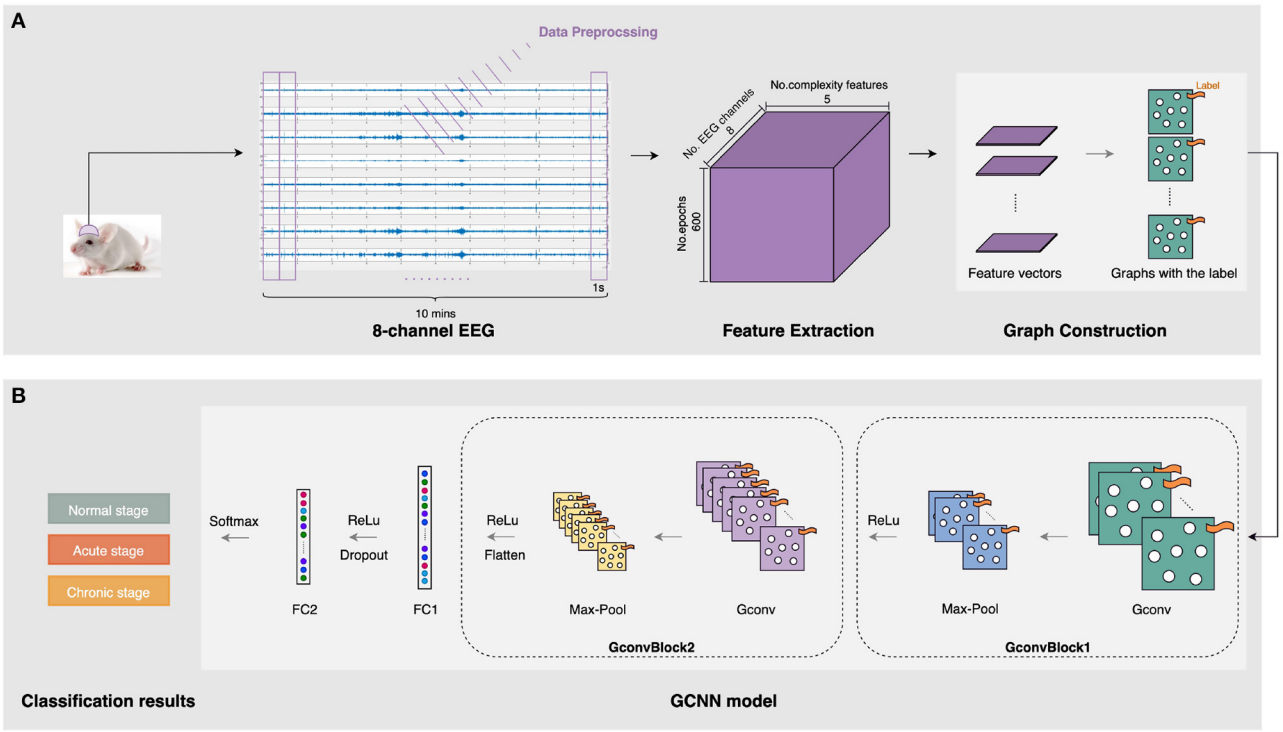


FIGURE 2
The architecture of the three-stage epileptic detection using complexity-GCNN classifier. **(A)** Graph construction for five complexity measures at eight channels. **(B)** epileptic detection GCNN classification model.

After a repeated graph convolution block, two FC layers followed. In particular, between these two FC layers, a ReLu layer was used, and a regularization technique called dropout was applied to avoid overfitting. Finally, the softmax activation function was used for three-stage epileptic detection tasks to obtain the result. The detailed configuration of this GCNN classification model is shown in Table 2.

TABLE 2 The configuration of the GCNN-based classifier.

	Layer	Output size (Tensor)
Input		[6*, 1, 8, 5]
GconvBlock1	Graph convolution	
	Pool	
	ReLU	[6, 10, 4, 3]
GconvBlock2	Graph convolution	
	Pool	
	ReLU	[6, 20, 2, 1]
	Flatten	[6, 40]
FC1	Fully connected	
	ReLU	[6, 15]
	Dropout	[6, 15]
FC2	Fully connected	[6, 3]
Prediction	Softmax	[6, 3]

*The batch size of training is six and the result will contain six training units.

2.3.2. Evaluation metrics

Three typical assessment methods: confusion matrix, accuracy and F1 score are employed to evaluate the classification performance of the GCNN model constructed on complexity measures.

2.3.2.1. Confusion matrix

It is a 3 × 3 matrix that tells us the rate of true positives and false positives when the sampled signal is from normal, acute, and chronic stages, respectively.

2.3.2.2. Accuracy

The overall accuracy is a classifier's ability to correctly predict the classes and is defined as:

$$Accuracy = \frac{Correct\ Predictions}{Total\ Predictions} \times 100\%. \quad (2)$$

2.3.2.3. F1 score

The F1 score refers to a balanced measure between two other metrics: precision and recall, where precision is the ability of the classifier to identify the positive class with accuracy, and recall is the ability of a model to predict each of the positive observations within a data set correctly. It is expressed as:

$$F1\ score = \frac{2 \times Precision \times Recall}{Precision + Recall} \times 100\%. \quad (3)$$

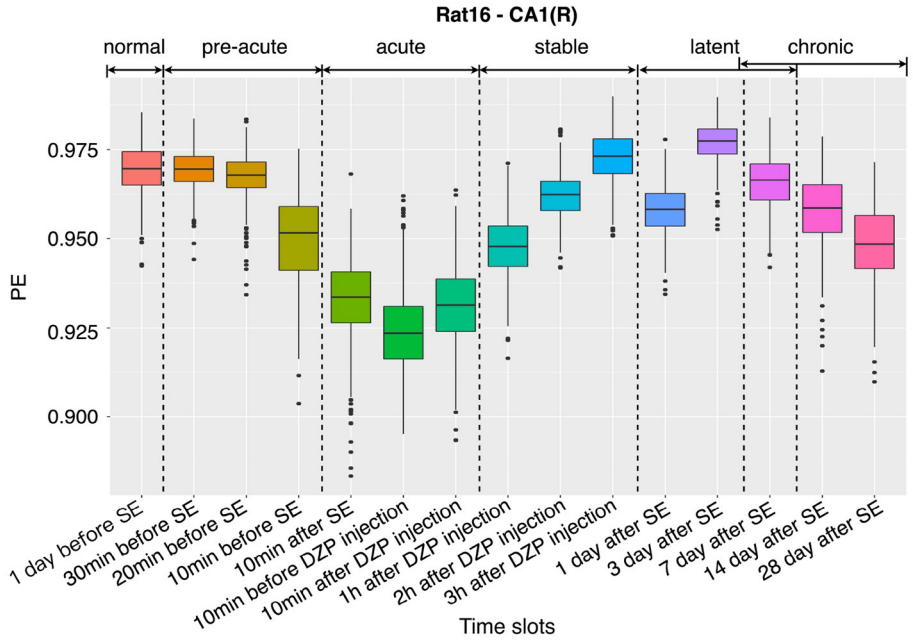


FIGURE 3
Dynamic changes of the complexity at 15 successive time slots of PILO modeling.

TABLE 3 Results of one-way ANOVA for distinguishing normal, acute, and chronic phases.

Complexity measure	Mean difference in multiple comparisons		*F*-test	*p*-value
	Normal-acute (*p*-value)	Normal-chronic (*p*-value)		
ApEn	0.7710 (5.1e-9)	0.0961 (5.4e-9)	1660.3	1.3e-215
SampEn	1.4203 (5.1e-9)	0.0858 (8.8e-8)	5502.9	0
PE	0.0310 (5.1e-9)	0.0200 (5.1e-9)	302.4	1.5e-85
FuzzEn	0.6199 (5.1e-9)	0.1011 (5.1e-9)	1711.6	2.2e-218
KC	0.1014 (5.1e-9)	0.0125 (8.2e-9)	1493.2	4.5e-206

3. Results and discussion

This section demonstrates the main results of EEG complexity analysis and three-stage epileptic detection.

The procedures of EEG processing and feature extraction were carried out using MATLAB R2022a. Statistical analyses were performed using SPSS 25.0, and the GCNN-based three-stage epileptic classification was conducted using Python 3.9.12.

During data processing, each 10-min EEG recording sample with 600,000 data points was divided into non-overlapping 1s epochs, resulting in 600 epochs and 1,000 data points in each epoch. Then, EEG signals were decomposed by wavelet transform based on the Haar wavelet and extracted a specific frequency band spanning 250–500 Hz (Fast Ripples). Following the data pre-processing, five complexity measures, including ApEn, SampEn, PE, FuzzEn, and KC, are calculated on each EEG epoch of the eight channels for further analysis.

3.1. Dynamic changes in complexity

To demonstrate the dynamic changes of the complexity for all stages mentioned in Section 2.1, a boxplot of the PE distributions at 15 successive stages of the channel CA1(L) of representative rat (no.16) is given in Figure 3. It was found that in the normal period (1 day before SE), the PE values are at a relatively high level, and the EEG shows a large randomness. The complexity starts to drop 30 min before SE, then continues to fall sharply until the DZP is injected. The decreasing of the complexity suggests that with the onset of epilepsy, EEG gradually presents some regular rhythms, which reduces the complexity. Afterward, from 10 min after DZP injection, PE values continue rising and recover to normal by 3 h after DZP injection. However, after the effect of DZP subsides, it is found that the values of PE begin to decline to a certain extent in the chronic stage. This indicates the appearance of SRSs. Using PE as a representative of EEG complexity clearly shows the dynamic changes of the brain's electrical activity before and after SE, in the process of seizure and DZP injection, and the chronic phase (Figure 3).

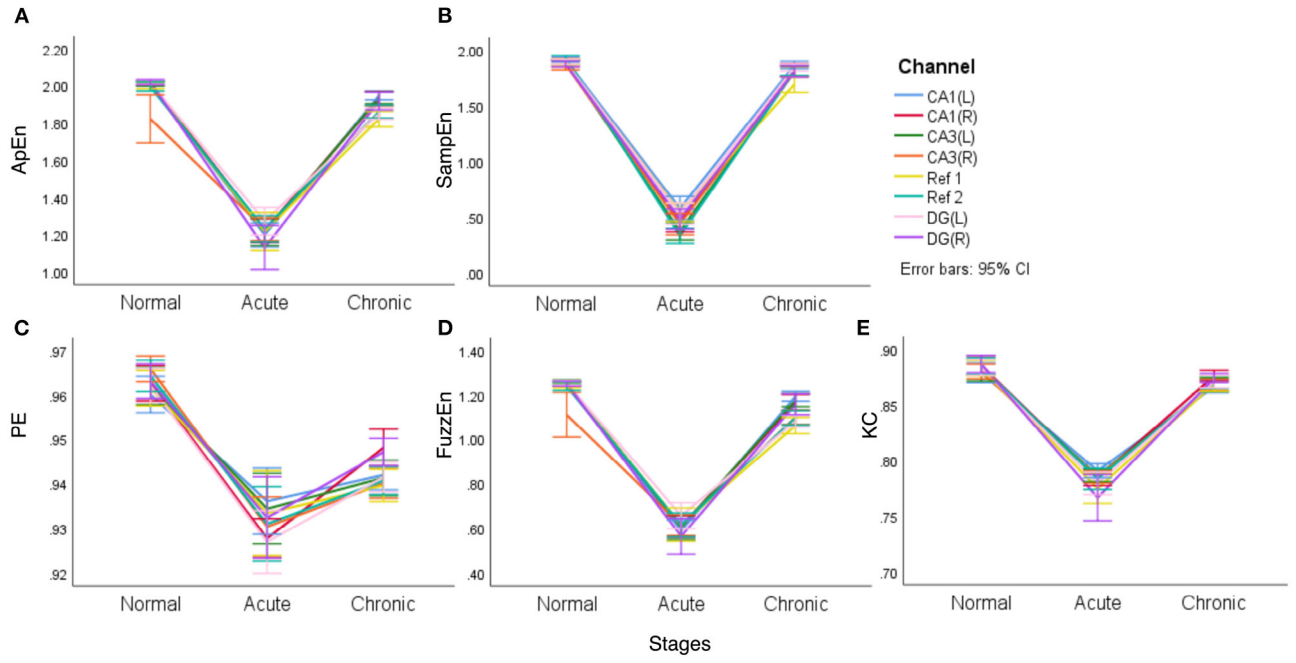


FIGURE 4
The line charts with error bars (95% CI) of each of the complexity measures in three epileptic stages for representative rat (no.16) **(A–E)**.

TABLE 4 **The hyperparameter settings of the GCNN-based classifier for classification.**

Hyperparameter	Values	
	For individual subject	Across all subjects
Learning rate	0.001	0.001
Epochs	3	50
Batch size (Train)	6	6
Batch size (Test)	2,175	540
Momentum	0.5	0.5
Log interval	10	10
Activation function	ReLU	ReLU

3.2. Statistical significance

EEG Complexity metrics at normal, acute, and chronic stages were compared through one-way ANOVA. The *F*-test statistics and the two-tailed *p*-values were presented in Table 3. Tukey's test was performed for pairwise comparison for the complexity between any two of the stages, and the mean differences (*p*-values) for normal and acute stages, normal and chronic stages were also given in Table 3. In this part, three 10-min EEG recordings, including "1 day before SE," "10 min before DZP injection," and "28 days after SE" were selected to represent normal, acute epilepsy, and chronic epilepsy, respectively. Each 10-min EEG recording was divided into 20 equal-length epochs. So, the number of each computed complexity measure for normal, acute, and chronic groups in one-way ANOVA is 160, including epochs from eight channels.

Through the results of one-way ANOVA, we found that using complexity as a feature can well reflect the differences between normal, acute, and chronic phases. Regardless of the type of complexity, the *p*-values of the *F*-tests are close to zero. In the pairwise comparisons using Tukey *post-hoc* testing, there is also a significant difference in complexity between normal and acute phases, as well as between normal and chronic phases, with *p*-values all below 10^{-7}. These results indicate that complexity measures are beneficial features in distinguishing different stages of epilepsy.

In fact, the difference between normal and chronic stages is rarely mentioned in literature. Song et al. (2016) tried to detect and quantify different phases of epileptogenesis by implementing average and peak spectral power of high-frequency oscillations (HFOs). They successfully found the dynamic changes between the acute and normal stages but failed to show statistical significance for differences between the chronic and normal stages using spectral power, the characteristic based on linear theories. Meanwhile, line charts of means and their 95% confidence intervals (CI) are presented to visualize the differences for all the five complexity measures in acute, normal, and chronic phases (Figure 4). Lines with eight colors represent eight EEG signal channels, including two reference channels (Ref 1 and Ref 2).

It is clear from Figure 4 that different complexity measures reflect similar laws, that is, the mean complexity of EEG is at a relatively high value in the normal period, while in the acute phase of epilepsy, the mean complexity has a significant decline, which confirms that during epilepsy, EEG will continue to appear some particular waveforms and become regular. In the chronic period, entropy will rise again, even returning to a level close to the normal

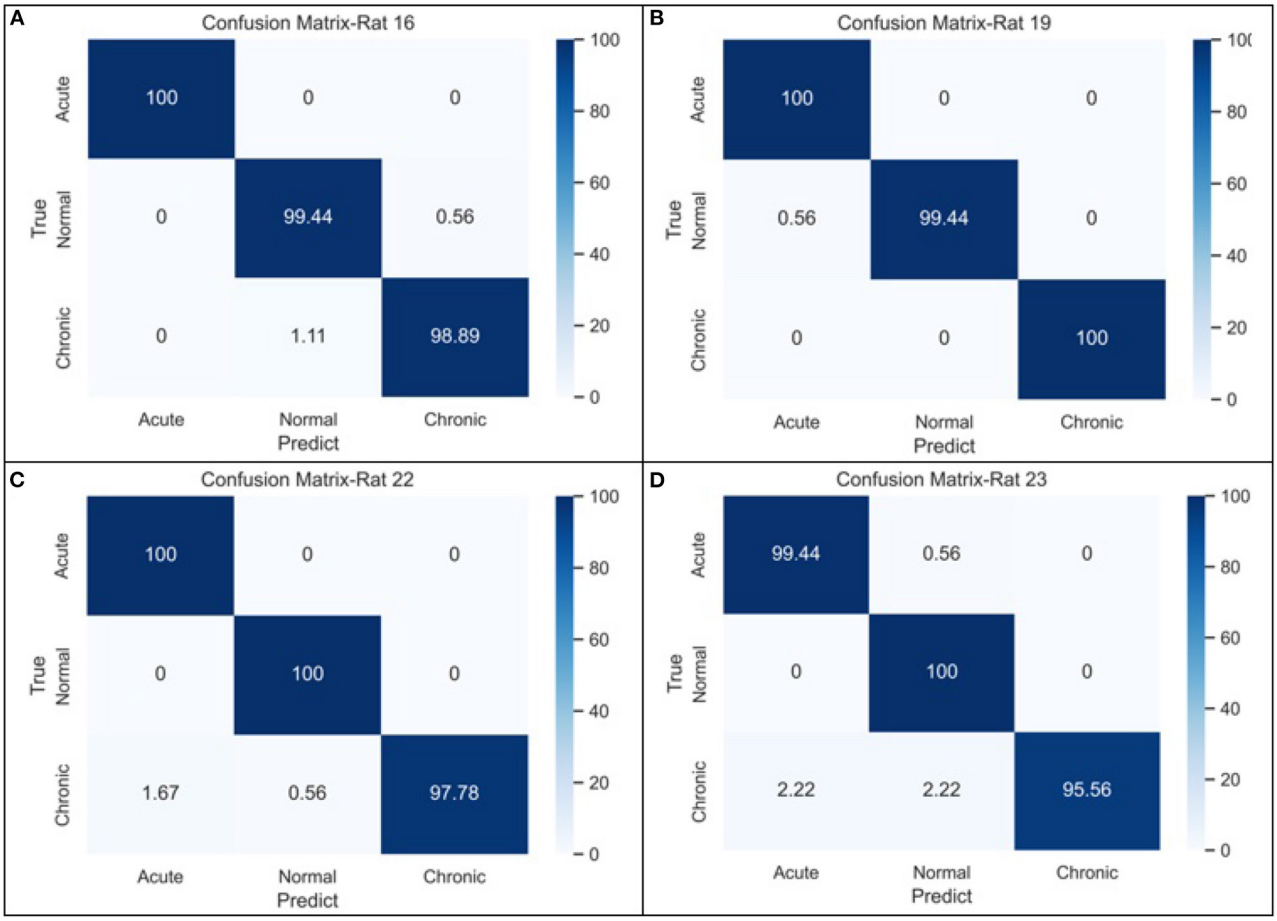


FIGURE 5
The confusion matrices of GCNN-based classifier for three-stage epileptic detection based on **(A)** rat no.16, **(B)** rat no.19, **(C)** rat no.22, and **(D)** rat no.23, respectively.

TABLE 5 **Classification performance of GCNN based on complexity measures.**

Subject	Accuracy (%)	F1 score (%)		
		Normal	Acute	Chronic
Rat no.16	0.9944	1.0000	0.9917	0.9916
Rat no.19	0.9981	0.9972	0.9972	1.0000
Rat no.22	0.9926	0.9917	0.9972	0.9888
Rat no.23	0.9833	0.9862	0.9863	0.9773
Combined	0.8782	0.9725	0.8603	0.7927

phase but slightly lower than the normal phase. In particular, for PE, the gap between the normal and chronic phases is relatively apparent. Another noteworthy point is that two reference channels (Ref 1 and Ref 2) are also included in this comparison. However, it is interesting to see from the line charts listed in Figure 4 that these two reference channels (Ref 1 and Ref 2) express similar complexity during the main stages of PILO modeling.

3.3. Classification performance

To evaluate the performance of complexity indicators in classifying the normal, acute, and chronic stages of epilepsy, we conduct GCNN-based classification with hyperparameter settings listed in Table 4 for each individual rat, and across all rats. The data was split into training, validation and testing sets, with a 50–20–30% partition. Figure 5 includes four confusion matrices obtained for four rats, where the detection rates of the three phases are calculated. Other useful evaluation indicators of model classification such as accuracy and F1 score are also listed in Table 5.

From the confusion matrices shown in Figure 5, the probability of being detected (i.e., sensitivity) for acute and normal phases is relatively high, reaching between 99.45 and 100%, while the detection rate of chronic phase is slightly lower, but still more than 95%. The classification performance across all subjects is shown in the last row of Table 5. It can be seen that when the measures of the four rats were merged, the effectiveness of classification decreased considerably due to the heterogeneity among individual rats.

To demonstrate the superiority of complexity metrics in differentiating chronic phases of epilepsy, we calculated two sets of EEG characteristics: one includes five complexity measures,

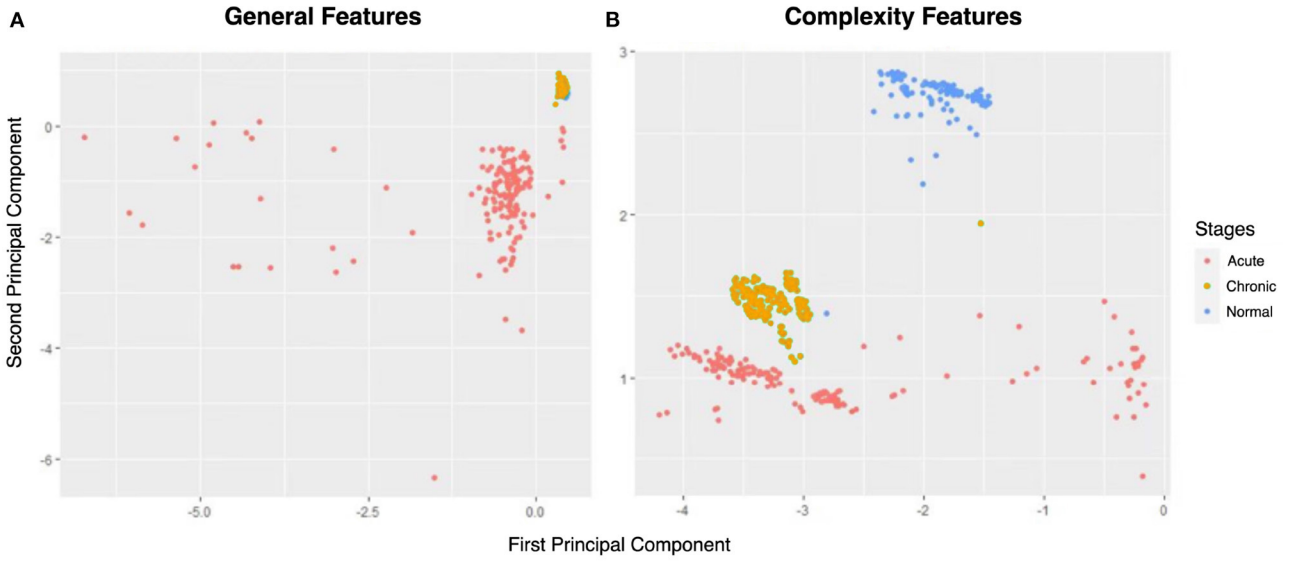


FIGURE 6
The principal component analysis (PCA) score plots of **(A)** general features and **(B)** complexity features.

and another has five general features: mean, variance, maximum, minimum, and skewness. Taking representative rat (no.16) as an example, the principal component (PC) method is applied to the two normalized five-dimensional characteristic data matrices to compress them to two-dimensional metrics. Figure 6 are 2-PC plots obtained from these two sets of features.

From Figure 6, the normal and acute phases can be well distinguished under either set of features. However, general indicators and complexity measures differ in their ability to distinguish normal and chronic phases. As shown in Figure 6A, there is a significant overlap between the yellow (i.e., chronic phase) and blue points (i.e., normal phase), so the general indicators mix these two phases. Nevertheless, the points of normal and chronic phases can be easily recognized using complexity measures (Figure 6B). Thus, the comparison in Figure 6 gives us a preliminary impression that complexity measurement can effectively identify the chronic phase of epilepsy.

4. Conclusion

In this paper, the differences in EEG between normal and chronic phases of epilepsy for rats were studied in depth for the first time. By calculating five commonly used complexity measures: ApEn, SampEn, PE, FuzzEn, and KC, the dynamic changes in brain waves during seizures can be perfectly displayed. Results of one-way ANOVA and PCA score plots show that complexity features can well reflect the differences between normal, acute, and chronic phases with extremely small *p*-values. In particular, among with these complexity metrics, PE exhibits the greatest discrepancy between normal and chronic stages. In order to integrate five complexity measures at eight channels, an automatic epileptic detection system via GCNN is developed. Our model reaches high performance in epilepsy detection that the recognition rate of each individual rat can achieve more than 98%, even 100%, including normal and chronic stages. In our case study, a comparison between modeling based on each individual subject and modeling across all subjects highlighted the non-negligible heterogeneity among individual rats. Modeling across all subjects may inadequately account for these individual differences, thus diminishing the model's fit to individual data. In contrast, modeling based on each individual subject can provide highly personalized models for each individual, significantly enhancing model accuracy, especially when the chronic phase is considered. This underscores the necessity of employing modeling based on each individual subject for personalized treatment recommendations in practical epilepsy management, ensuring better alignment with patients' unique needs.

While the above experiments yielded promising results in the classification of three epilepsy stages, our investigation was limited to the effectiveness of this framework solely in rat data and for just one type of epilepsy. In future work, we intend to extend the application of this framework to human EEG datasets. Concurrently, we will make adjustments to both graph representations and model parameters to elucidate the distinct characteristics of human EEG data, thus enhancing the model's generalization capabilities.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

SZ proposed the work, analyzed the data, and wrote the manuscript. XZ did the main computation work. PS did the experiment and generated the data. YH interpreted the results and provided guidance. XG analyzed the data and revised the manuscript. XP conducted the whole work. All authors contributed to the article and approved the submitted version.

Funding

This work was partially supported by the National Key R&D Program of China (Grant No. 2022YFC3600300), the National Natural Science Foundation of China (Grant No. 11971020), Guangdong Provincial Key Laboratory of Interdisciplinary Research and Application for Data Science (No. 2022B1212010006), and Chongqing Science and Health Joint Medical Research Project (No. 2023MSXM137).

Acknowledgments

We thank all participants who supported our study and the reviewers for constructive suggestions on the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Arunkumar, N., Ramkumar, K., Venkatraman, V., Abdulhay, E., Fernandes, S. L., Kadry, S., et al. (2017). Classification of focal and non-focal EEG using entropies. *Pattern Recogn. Lett*. 94, 112–117. doi: 10.1016/j.patrec.2017.05.007

Bandt, C., and Pompe, B. (2002). Permutation entropy: a natural complexity measure for time series. *Phys. Rev. Lett*. 88, 174102. doi: 10.1103/PhysRevLett.88.174102

Boonyakitanont, P., Lek-Uthai, A., Chomtho, K., and Songsiri, J. (2020). A review of feature extraction and performance evaluation in epileptic seizure detection using EEG. *Biomed. Signal Process. Control* 57, 101702. doi: 10.1016/j.bspc.2019.101702

Bullmore, E., and Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci*. 10, 186–198. doi: 10.1038/nrn2575

Chaitin, G. J. (1977). Algorithmic information theory. *IBM J. Res. Dev*. 21, 350359. doi: 10.1147/rd.214.0350

Chen, W., Wang, Z., Xie, H., and Yu, W. (2007). Characterization of surface EMG signal based on fuzzy entropy. *IEEE Trans. Neural Syst. Rehabil. Eng*. 15, 266–272. doi: 10.1109/TNSRE.2007.897025

Covert, I. C., Krishnan, B., Najm, I., Zhan, J., Shore, M., Hixson, J., et al. (2019). "Temporal graph convolutional networks for automatic seizure detection," in *Proceedings of Machine Learning Research 106 (PMLR)*, 160–180.

Craley, J., Jouny, C., Johnson, E., Hsu, D., Ahmed, R., and Venkataraman, A. (2022). Automated seizure activity tracking and onset zone localization from scalp EEG using deep neural networks. *PLoS ONE* 17, e0264537. doi: 10.1371/journal.pone.0264537

Defferrard, M., Bresson, X., and Vandergheynst, P. (2016). "Convolutional neural networks on graphs with fast localized spectral filtering," in *Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16* (Red Hook, NY: Curran Associates Inc.), 3844–3852

Faust, O., Acharya, U. R., Min, L. C., and Sputh, B. H. (2010). Automatic identification of epileptic and background EEG signals using frequency domain parameters. *Int. J. Neural Syst*. 20, 159–176. doi: 10.1142/S0129065710002334

Fisher, R. S., Boas, W. V. E., Blume, W., Elger, C., Genton, P., Lee, P., et al. (2005). Epileptic seizures and epilepsy: definitions proposed by the international league against epilepsy (ILAE) and the international bureau for epilepsy (IBE). *Epilepsia* 46, 470–472. doi: 10.1111/j.0013-9580.2005.66104.x

Henry, M., and Judge, G. (2019). Permutation entropy and information recovery in nonlinear dynamic economic time series. *Econometrics* 7, 10. doi: 10.3390/econometrics7010010

Hou, Y., Jia, S., Lun, X., Hao, Z., Shi, Y., Li, Y., et al. (2022). GCNs-Net: a graph convolutional neural network approach for decoding time-resolved EEG motor imagery signals. *IEEE Trans. Neural Netw. Learn. Syst*. 1–12. doi: 10.1109/TNNLS.2022.3202569

Jia, Z., Lin, Y., Wang, J., Zhou, R., Ning, X., He, Y., et al. (2020). "Graphsleepnet: adaptive spatial-temporal graph convolutional networks for sleep stage classification," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20), Vol. 2021*, 1324–1330. doi: 10.24963/ijcai.2020/184

Karlócai, M. R., Tóth, K., Watanabe, M., Ledent, C., Juhász, G., Freund, T. F., et al. (2011). Redistribution of cb1 cannabinoid receptors in the acute and chronic phases of pilocarpine-induced epilepsy. *PLoS ONE* 6, e27196. doi: 10.1371/journal.pone.0027196

Li, G., and Jung, J. J. (2021). Seizure detection from multi-channel EEG using entropy-based dynamic graph embedding. *Artif. Intell. Med*. 122, 102201. doi: 10.1016/j.artmed.2021.102201

Li, M., Chen, W., and Zhang, T. (2017). Automatic epileptic EEG detection using DT-CWT-based non-linear features. *Biomed. Signal Process. Control* 34, 114–125. doi: 10.1016/j.bspc.2017.01.010

Li, M., and Vitányi, P. (2008). *An Introduction to Kolmogorov Complexity and Its Applications, Vol. 3*. New York, NY: Springer.

Lian, Q., Qi, Y., Pan, G., and Wang, Y. (2020). Learning graph in graph convolutional neural networks for robust seizure prediction. *J. Neural Eng*. 17, 035004. doi: 10.1088/1741-2552/ab909d

Liang, Z., Wang, Y., Sun, X., Li, D., Voss, L. J., Sleigh, J. W., et al. (2015). EEG entropy measures in anesthesia. *Front. Comput. Neurosci*. 9, 16. doi: 10.3389/fncom.2015.00016

Natarajan, K., Acharya U, R., Alias, F., Tiboleng, T., Puthusserypady, S. K., et al. (2004). Nonlinear analysis of EEG signals at different mental states. *Biomed. Eng. Online* 3, 1–11. doi: 10.1186/1475-925X-3-7

Pincus, S. M., Gladstone, I. M., and Ehrenkranz, R. A. (1991). A regularity statistic for medical data analysis. *J. Clin. Monit*. 7, 335–345. doi: 10.1007/BF01619355

Raeisi, K., Khazaei, M., Croce, P., Tamburro, G., Comani, S., and Zappasodi, F. (2022). A graph convolutional neural network for the automated detection of seizures in the neonatal EEG. *Comput. Methods Prog. Biomed*. 2022, 106950. doi: 10.1016/j.cmpb.2022.106950

Richman, J. S., and Moorman, J. R. (2000). Physiological time-series analysis using approximate entropy and sample entropy. *Am. J. Physiol. Heart Circul. Physiol*. 278, H2039–H2049. doi: 10.1152/ajpheart.2000.278.6.H2039

Saxena, S., and Li, S. (2017). Defeating epilepsy: a global public health commitment. *Epileps. Open* 2, 153–155. doi: 10.1002/epi4.12010

Sharma, R., Pachori, R. B., and Acharya, U. R. (2014). Application of entropy measures on intrinsic mode functions for the automated identification of focal electroencephalogram signals. *Entropy* 17, 669–691. doi: 10.3390/e17020669

Sharmila, A., and Geethanjali, P. (2018). Effect of filtering with time domain features for the detection of epileptic seizure from EEG signals. *J. Med. Eng. Technol.* 42, 217–227. doi: 10.1080/03091902.2018.1464075

Solomonoff, R. J. (1960). *A Preliminary Report on a General Theory of Inductive Inference*. Citeseer.

Song, P., Xiang, J., Jiang, L., Chen, H., Liu, B., and Hu, Y. (2016). Dynamic changes in spectral and spatial signatures of high frequency oscillations in rat hippocampi during epileptogenesis in acute and chronic stages. *Front. Neurol.* 7, 204. doi: 10.3389/fneur.2016.00204

Song, T., Zheng, W., Song, P., and Cui, Z. (2018). EEG emotion recognition using dynamical graph convolutional neural networks. *IEEE Trans. Affect. Comput.* 11, 532–541. doi: 10.1109/TAFFC.2018.2817622

Srinivasan, V., Eswaran, C., and Sriram, N. (2005). Artificial neural network based epileptic detection using time-domain and frequency-domain features. *J. Med. Syst.* 29, 647–660. doi: 10.1007/s10916-005-6133-1

Such, F. P., Sah, S., Dominguez, M. A., Pillai, S., Zhang, C., Michael, A., et al. (2017). Robust spatial filtering with graph convolutional neural networks. *IEEE J. Select. Top. Signal Process.* 11, 884–896. doi: 10.1109/JSTSP.2017.2726981

Tzallas, A. T., Tsipouras, M. G., and Fotiadis, D. I. (2009). Epileptic seizure detection in EEGs using time-frequency analysis. *IEEE transactions on information technology in biomedicine,* 13, 703–710. doi: 10.1109/TITB.2009.2017939

Wagh, N., and Varatharajah, Y. (2020). "EEG-GCNN: augmenting electroencephalogram-based neurological disease diagnosis using a domain-guided graph convolutional neural network," in *Proceedings of Machine Learning Research 136 (PMLR)*, 367–378.

Wang, L., Xue, W., Li, Y., Luo, M., Huang, J., Cui, W., et al. (2017). Automatic epileptic seizure detection in EEG signals using multi-domain feature extraction and nonlinear analysis. *Entropy* 19, 222. doi: 10.3390/e19060222

Wei, Z., Zou, J., Zhang, J., and Xu, J. (2019). Automatic epileptic EEG detection using convolutional neural network with improvements in time-domain. *Biomed. Signal Process. Control* 53, 101551. doi: 10.1016/j.bspc.2019.04.028

Wen, T., and Zhang, Z. (2017). Effective and extensible feature extraction method using genetic algorithm-based frequency-domain feature search for epileptic EEG multiclassification. *Medicine* 96, 6879. doi: 10.1097/MD.0000000000006879

Xiang, J., Li, C., Li, H., Cao, R., Wang, B., Han, X., et al. (2015). The detection of epileptic seizure signals based on fuzzy entropy. *J. Neurosci. Methods* 243, 18–25. doi: 10.1016/j.jneumeth.2015.01.015

Yuan, Q., Zhou, W., Li, S., and Cai, D. (2011). Epileptic EEG classification based on extreme learning machine and nonlinear features. *Epilepsy Res.* 96, 29–38. doi: 10.1016/j.eplepsyres.2011.04.013

Zadeh, L. A., Klir, G. J., and Yuan, B. (1996). *Fuzzy Sets, Fuzzy Logic, and Fuzzy Systems: Selected Papers, Vol. 6*. World Scientific. doi: 10.1142/2895

Zhang, T., Wang, X., Xu, X., and Chen, C. P. (2019). GCB-Net: graph convolutional broad network and its application in emotion recognition. *IEEE Trans. Affect. Comput.* 13, 379–388. doi: 10.1109/TAFFC.2019.2937768

frontiers | Frontiers in Computational Neuroscience

TYPE Opinion
PUBLISHED 19 October 2023
DOI 10.3389/fncom.2023.1243092

OPEN ACCESS

EDITED BY
Noemi Montobbio,
University of Genoa, Italy

REVIEWED BY
Michael Thrun,
University of Marburg, Germany
Eduardo Castro,
University of New Mexico, United States
Andrea Chincarini,
National Institute of Nuclear Physics of Genoa, Italy

*CORRESPONDENCE
Konstantinos Poulakis
konstantinos.poulakis@ki.se

RECEIVED 20 June 2023
ACCEPTED 04 October 2023
PUBLISHED 19 October 2023

CITATION
Poulakis K and Westman E (2023) Clustering and disease subtyping in Neuroscience, toward better methodological adaptations. *Front. Comput. Neurosci.* 17:1243092. doi: 10.3389/fncom.2023.1243092

Clustering and disease subtyping in Neuroscience, toward better methodological adaptations

Konstantinos Poulakis[1]* and Eric Westman[1,2]

[1]Division of Clinical Geriatrics, Department of Neurobiology, Care Sciences and Society, Karolinska Institutet, Stockholm, Sweden, [2]Department of Neuroimaging, Centre for Neuroimaging Sciences, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom

KEYWORDS

clustering, unsupervised learning, Neuroscience, disease subtypes, Alzheimer's disease

The increasing interest in identifying disease biomarkers to understand psychiatric and neurological conditions has led to large patient registries and cohorts. Traditionally, clinically defined labels (e.g., disease vs. control group) were associated statistically with potential biomarkers to draw useful information about brain function related to a disease (supervised analysis) (Deo, 2015). However, the observed biomarker variability and the presence of clinical disease subtypes have sparked interest in quantitatively exploring heterogeneity (Feczko et al., 2019; Ferreira et al., 2020). The unsupervised[1] exploration of a disease population (without any clinical labels) through a selected sample is a demanding task that differs from supervised analysis by definition (Habes et al., 2020). However, in research the differences between the two are often overlooked. Therefore, we want to highlight the applications and challenges of clustering, where supervised analysis principles are sometimes misapplied. We also demonstrate how such practices can negatively impact clustering results.

Some common challenges in clustering methods include selecting relevant features to describe data heterogeneity, preprocessing to remove biases, choosing appropriate similarity measures to summarize critical information, selecting a suitable method for meaningful clustering, tuning clustering model parameters (such as cluster size) without ground truth, and validating clustering results (Halkidi et al., 2001; Hennig et al., 2015).

The most common clustering applications in medicine (Halkidi et al., 2001):

- Data reduction (Hennig et al., 2015). When dealing with large datasets, like genomics, proteomics, or medical imaging data, clustering can condense the information into representative vectors or filter out uninformative features.
- Generate new hypotheses. Discovering specific disease subtypes can lead to the development of new hypotheses, altering existing theories.
- Hypothesis testing (Thrun and Ultsch, 2021). Clustering can be used for hypothesis testing. For example, it can assess whether clinical observations align with biological data in diseases with known subtypes without forcing the association between biological data and clinical labels (supervised approach).
- Prediction in new patients (Wu et al., 2019). Clustering can identify disease subtypes and scientific theories that investigators can use to create supervised classification models for grouping new patients. This new classification is valuable for personalized medicine and future patient treatment, among other applications.

1 For the needs of this text, unsupervised analysis refers to clustering only, association analysis is not covered.

When working with unsupervised methods, it's crucial to understand their limitations and nuances. Clustering encompasses a wide range of techniques which handle population structures and characteristics differently. Understanding the idiosyncrasies of a dataset is essential for applying clustering successfully. Questions about how clustering results generalize to the disease population, which are the optimal model parameters, and why results change with slight dataset modifications often emerge during study design, model optimization, interpretation, and peer review. One intriguing approach that combines automatic machine learning with expert knowledge from the field is the 'human-in-the-loop' method (Holzinger, 2016). This approach is particularly effective in neurological applications and can help address the abovementioned questions.

Regarding cluster size and type, we may know in advance whether there is excess variation in a disease population, some heterogeneous disease features, and even subtype proportions. This knowledge is vital in the model selection process so that we can sort out methods that are wrong methodological fits for the population of interest. For example, k-means, one of the most popular clustering methods, tends to produce convex-shaped clusters (it tends to equalize the spatial variance) that are spherical and often become similar in size (Celebi et al., 2013). Therefore, if in a specific disease population, we are aware of rare disease subtypes that may also exist in our sample, we may want to avoid k-means. Instead, we should focus on clustering methods to identify outliers/outlier clusters (Campello et al., 2015). Further, the more variables we use in a clustering method, the more the dimensionality of the dataset increases. A good practice is to use methods that either pretreat data to reduce the dimensionality and then apply regular clustering to them or select a method that can cope with high dimensional datasets (Babu et al., 2011; Thrun, 2021). While the gold standard in machine learning, some studies fail to utilize suitable models for high-dimensional data (Noh et al., 2014; Hwang et al., 2016; Jeon et al., 2019; Levin et al., 2021), limiting our ability to assess the success of clustering.

Further, all clustering methods cannot cope with all types of data (ordinal/nominal categorical, numerical) (Halkidi et al., 2001). When we binarize continuous variables to utilize a clustering algorithm for binary data only, the reduction of information due to data transformation must be at least considered when interpreting the results (Zhang et al., 2016). Some algorithms use mixed data types and should be preferred when mixed data distributions are present (Szepannek, 2019). If not accounted for, data biases may render a clustering result misleading. For example, we may be interested in understanding the heterogeneity of a particular biological process during aging. Understanding and adjusting the data to consider the participants' age variability results in clusters of participants that are not driven by age differences but by differences in the biological process under investigation if those exist (given that other biases are not present). However, due to complex data/aging relationships, these effects may persist even after statistical accounting for aging. Other sampling features that can drive clustering results are sex, disease stage, comorbidities, medication exposure, and geographical position. For example, it is known that the disease stage may contribute to the observed heterogeneity in Alzheimer's disease (AD) (Ferreira et al., 2020), we have only recently started accounting for this or trying to assess its contribution (Young et al., 2017; Vogel et al., 2021; Yang et al., 2021; Poulakis et al., 2022) while in previous studies (Noh et al., 2014; Dong et al., 2016; Hwang et al., 2016; Zhang et al., 2016; Park et al., 2017; Poulakis et al., 2018; ten Kate et al., 2018) we did not assess or account for this effect.

Clustering results must generalize well to the population, which makes validation a central topic. Traditionally, cross-validation (CV), bootstrapping, external data testing (training, validating, and testing), and careful sample selection have been some of the most popular approaches in supervised analysis. However, validation in clustering is not straightforward since no ground truth exists. The adaptation of training and testing a clustering model using independent datasets can sometimes mislead us. For example, three subtypes are present in a hypothetical disease population N (s1, s2, and s3). One is the most prevalent (s1) (typical presentation), the second subtype (s2) has half of the prevalence of the first one ($n_{s2} = \frac{1}{2}n_{s1}$), and the third subtype has a low prevalence (one-tenth of the first subtype, $n_{s3} = \frac{1}{10}n_{s1}$) (s3). The disease population N equals $n_1 + n_2 + n_3$. A perfectly representative random sample of 100 patients from the disease population will include approximately 63 patients from s1, 31 from s2, and six from s3. A clustering model can then be trained on 70% (70 patients) and tested using 30% (30 patients). Suppose the data in the training set perfectly represent the population, a rare phenomenon, and clustering accurately identifies the subtypes. In that case, 44 patients will end up in Cluster 1, 22 in Cluster 2, and 4 in Cluster 3. The test set should have 19 patients in s1, 9 in s2, and 2 in s3. Clustering can then be applied to identify subtypes s1, s2, and s3. Since the actual data labels are unknown, which is what clustering should discover, the test set results will be compared to the training set. The problem arises with rare subtypes, such as the hypothetical s3 subtype (six patients in the sample, four in the training set, and two in the test set). Patients of such subtypes may end up in larger clusters when the overall dataset is split into small segments for the needs of the analysis. Unfortunately, the most interesting heterogeneous characteristics will enrich another cluster's greater information pool, especially in high-dimensional datasets. In the best-case scenario, those patients will be single outliers (if the algorithm can recognize outlier clusters) (Campello et al., 2015). Understanding their features is pivotal for the assessment of heterogeneity in the disease.

To the best of our knowledge, cross-validation has been successfully combined with clustering in two studies to assess the consistency of observations within the same cluster and to determine the optimal model solution (Varol et al., 2017; Yang et al., 2021). On the other hand, leave 10% of patients out-CV (a semi-supervised application where a control group is contrasted to a disease group) to decide the optimal clustering (Dong et al., 2016, 2017), may reveal the dominant patterns in the dataset. An interesting question is whether clusters of low/very low prevalence can survive this process. In AD, genetic mutations account for <1% of all AD (2020) cases, while early-onset AD accounts for 4%–6% (Mendez, 2017). Another evaluation approach is to compare clustering agreement after application of the same algorithm in different cohorts. We do not suggest that these results are wrong, but they may be misleading if different clustering findings in different cohorts are interpreted as a methodological

failure, while convergence of findings between cohorts is the aim (ten Kate et al., 2018; Vogel et al., 2021). Sometimes, it is a requirement that clustering should be repeated cohort-wise to prove model robustness (Poulakis et al., 2018, 2022). Instead of reducing data variability in clustering by splitting the available data into segments, we should acknowledge that cluster-cohort agreement-based evaluation criteria can potentially interrupt the discovery of rare data patterns. Another issue with the cohort-wise analysis is the potential sample imbalance between cohorts that may render one cohort solution less reliable than another. Of note, cohort-wise analysis is reasonable when cohorts have different feature sets or systematic differences (Marinescu et al., 2019; Tijms et al., 2020). Prior knowledge (subtype prevalence or number of subtypes) is essential when formulating a clustering experimental design (Halkidi et al., 2001, 2002). Another example, hypothetically, two separate clusters of patients may be formed because a clustering validation criterion gives marginally better scores instead of grouping the patients in one cluster. Field experts and not only clustering internal evaluation criteria should conclude whether differences between clusters are essential enough to suggest heterogeneity (Halkidi et al., 2002; Dolnicar and Leisch, 2010). It is also often observed that clustering algorithms optimally select two-cluster solutions. This finding may not provide any insight of the disease process when it only reveals biomarker severity differences of no clinical interest (Poulakis et al., 2021; Yang et al., 2021). Based on the above, we believe that as large datasets as possible should be used when training a clustering model. In contrast, datasets should not be divided for validation purposes if the focus is on revealing heterogeneity in a population.

Clustering is a valuable approach to understand heterogeneity in brain disorders and healthy aging. The machine learning community has invested a great deal of research in addressing the methodological issues discussed above. As with every statistical tool, these methods should be carefully applied, and understanding their properties and limitations is essential.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Funding

We would like to thank the Swedish Foundation for Strategic Research (SSF), the Swedish Research Council (VR), the Center for Innovative Medicine (CIMED), the Strategic Research Programme in Neuroscience at Karolinska Institutet (StratNeuro), Swedish Brain Power, the regional agreement on medical training and clinical research (ALF) between Stockholm County Council and Karolinska Institutet, Hjärnfonden, Alzheimerfonden, the Åke Wiberg Foundation, the King Gustaf V:s and Queen Victorias Foundation, and Birgitta och Sten Westerberg for additional financial support.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

AD (2020). 2020 Alzheimer's disease facts and figures. *Alzheimers Dement.* 16, 391–460. doi: 10.1002/alz.12068

Babu, B., Subash, C. N., and Gopal, T. V. (2011). Clustering algorithms for high dimensional data – a survey of issues and existing approaches. *Spec. Issue Int. J. Comput. Sci. Inform.* 2, 13. doi: 10.47893/IJCSI.2013.1108

Campello, R. J. G. B., Moulavi, D., Zimek, A., and Sander, J. (2015). Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Trans. Knowl. Discov. Data* 10, 1–51. doi: 10.1145/2733381

Celebi, M. E., Kingravi, H. A., and Vela, P. A. A. (2013). comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Syst. Appl.* 40, 200–210. doi: 10.1016/j.eswa.2012.07.021

Deo, R. C. (2015). Machine learning in medicine. *Circulation* 132, 1920–1930. doi: 10.1161/CIRCULATIONAHA.115.001593

Dolnicar, S., and Leisch, F. (2010). Evaluation of structure and reproducibility of cluster solutions using the bootstrap. *Mark. Lett.* 21, 83–101. doi: 10.1007/s11002-009-9083-4

Dong, A., Honnorat, N., Gaonkar, B., and Davatzikos, C. (2016). CHIMERA: clustering of heterogeneous disease effects via distribution matching of imaging patterns. *IEEE Trans. Med. Imaging* 35, 612–621. doi: 10.1109/TMI.2015.2487423

Dong, A., Toledo, J. B., Honnorat, N., Doshi, J., Varol, E., Sotiras, A., et al. (2017). Heterogeneity of neuroanatomical patterns in prodromal Alzheimer's disease: links to cognition, progression and biomarkers. *Brain* 140, 735–747. doi: 10.1093/brain/aww319

Feczko, E., Miranda-Dominguez, O., Marr, M., Graham, A. M., Nigg, J. T., and Fair, D. A. (2019). The heterogeneity problem: approaches to identify psychiatric subtypes. *Trends Cogn. Sci.* 23, 584–601. doi: 10.1016/j.tics.2019.03.009

Ferreira, D., Nordberg, A., and Westman, E. (2020). Biological subtypes of Alzheimer disease. *Neurology* 94, 436–448. doi: 10.1212/WNL.0000000000009058

Habes, M., Grothe, M. J., Tunc, B., McMillan, C., Wolk, D. A., and Davatzikos, C. (2020). Disentangling heterogeneity in Alzheimer's disease and related dementias using data-driven methods. *Biol. Psychiatry* 88, 70–82. doi: 10.1016/j.biopsych.2020.01.016

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2001). On clustering validation techniques. *J. Intell. Inf. Syst.* 17, 107–145. doi: 10.1023/A:1012801612483

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2002). Clustering validity checking methods. *ACM SIGMOD Rec.* 31, 19–27. doi: 10.1145/601858.601862

Hennig, C., Meila, M., Murtagh, F., Rocci, R. (2015). *Handbook of Cluster Analysis*. Boca Raton, FL: Chapman and Hall/CRC. doi: 10.1201/b19706

Holzinger, A. (2016). Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Inform.* 3, 119–131. doi: 10.1007/s40708-016-0042-6

Hwang, J., Kim, C. M., Jeon, S., Lee, J. M., Hong, Y. J., Roh, J. H., et al. (2016). Prediction of Alzheimer's disease pathophysiology based on cortical

thickness patterns. *Alzheimer's Dement. Diagnosis, Assess. Dis. Monit.* 2, 58–67. doi: 10.1016/j.dadm.2015.11.008

Jeon, S., Kang, J. M., Seo, S., Jeong, H. J., Funck, T., Lee, S.-Y., et al. (2019). Topographical heterogeneity of Alzheimer's disease based on MR imaging, tau PET, and amyloid PET. *Front. Aging Neurosci.* 11, 211. doi: 10.3389/fnagi.2019.00211

Levin, F., Ferreira, D., Lange, C., Dyrba, M., Westman, E., Buchert, R., et al. (2021). Data-driven FDG-PET subtypes of Alzheimer's disease-related neurodegeneration. *Alzheimers Res Ther.* 13, 49. doi: 10.1186/s13195-021-00785-9

Marinescu, R. V., Eshaghi, A., Lorenzi, M., Young, A. L., Oxtoby, N. P., Garbarino, S., et al. (2019). DIVE: a spatiotemporal progression model of brain pathology in neurodegenerative disorders. *Neuroimage* 192, 166–177. doi: 10.1016/j.neuroimage.2019.02.053

Mendez, M. F. (2017). Early-onset Alzheimer disease. *Neurol. Clin.* 35, 263–281. doi: 10.1016/j.ncl.2017.01.005

Noh, Y., Jeon, S., Lee, J. M., Seo, S. W., Kim, G. H., Cho, H., et al. (2014). Anatomical heterogeneity of Alzheimer disease based on cortical thickness on MRIs. *Neurology* 83, 1936–1944. doi: 10.1212/WNL.0000000000001003

Park, J.-Y., Na, H. K., Kim, S., Kim, H., Kim, H. J., Seo, S. W., et al. (2017). Robust Identification of Alzheimer's disease subtypes based on cortical atrophy patterns. *Sci. Rep.* 7, 43270. doi: 10.1038/srep43270

Poulakis, K., Pereira, J. B., Mecocci, P., Vellas, B., Tsolak, M., Kłoszewska, I., et al. (2018). Heterogeneous patterns of brain atrophy in Alzheimer's disease. *Neurobiol. Aging* 65, 98–108. doi: 10.1016/j.neurobiolaging.2018.01.009

Poulakis, K., Pereira, J. B., Muehlboeck, J.-S., Wahlund, L.-O., Smedby, Ö., Volpe, G., et al. (2022). Multi-cohort and longitudinal Bayesian clustering study of stage and subtype in Alzheimer's disease. *Nat. Commun.* 13, 4566. doi: 10.1038/s41467-022-32202-6

Poulakis, K., Reid, R. I., Przybelski, S. A., Knopman, D. S., Graff-Radford, J., Lowe, V. J., et al. (2021). Longitudinal deterioration of white-matter integrity: heterogeneity in the ageing population. *Brain Commun.* 3, fcaa238. doi: 10.1093/braincomms/fcaa238

Szepannek, G. (2019). clustMixType: user-friendly clustering of mixed-type data in R. *R J.* 10, 200. doi: 10.32614/RJ-2018-048

ten Kate, M., Dicks, E., Visser, P. J., van der Flier, W. M., Teunissen, C. E., Barkhof, F., et al. (2018). Atrophy subtypes in prodromal Alzheimer's disease are associated with cognitive decline. *Brain* 141, 3443–3456. doi: 10.1093/brain/awy264

Thrun, M. C. (2021). Distance-based clustering challenges for unbiased benchmarking studies. *Sci. Rep.* 11, 18988. doi: 10.1038/s41598-021-98126-1

Thrun, M. C., and Ultsch, A. (2021). Swarm intelligence for self-organized clustering. *Artif. Intell.* 290, 103237. doi: 10.1016/j.artint.2020.103237

Tijms, B. M., Gobom, J., Reus, L., Jansen, I., Hong, S., Dobricic, V., et al. (2020). Pathophysiological subtypes of Alzheimer's disease based on cerebrospinal fluid proteomics. *Brain* 143, 3776–3792. doi: 10.1093/brain/awaa325

Varol, E., Sotiras, A., and Davatzikos, C. (2017). HYDRA: revealing heterogeneity of imaging and genetic patterns through a multiple max-margin discriminative analysis framework. *Neuroimage* 145, 346–364. doi: 10.1016/j.neuroimage.2016.02.041

Vogel, J. W., Young, A. L., Oxtoby, N. P., Smith, R., Ossenkoppele, R., Strandberg, O. T., et al. (2021). Four distinct trajectories of tau deposition identified in Alzheimer's disease. *Nat. Med.* 27, 871–881. doi: 10.1038/s41591-021-01309-6

Wu, W., Bang, S., Bleecker, E. R., Castro, M., Denlinger, L., Erzurum, S. C., et al. (2019). Multiview cluster analysis identifies variable corticosteroid response phenotypes in severe asthma. *Am. J. Respir. Crit. Care Med.* 199, 1358–1367. doi: 10.1164/rccm.201808-1543OC

Yang, Z., Nasrallah, I., Shou, H., Wen, J., Doshi, J., Habes, M., et al. (2021). Disentangling brain heterogeneity via semi-supervised deep-learning and MRI: dimensional representations of Alzheimer's disease. *Alzheimers Dement.* 17*:* doi: 10.1002/alz.052735

Young, A. L., Marinescu, R. V., Oxtoby, N. P., Bocchetta, M., Yong, K., Firth, N. C., et al. (2017). Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with subtype and stage inference. *bioRxiv* [preprint]. doi: 10.1101/236604

Zhang, X., Mormino, E. C., Sun, N., Sperling, R. A., Sabuncu, M. R., Yeo, B. T. T., et al. (2016). Bayesian model reveals latent atrophy factors with dissociable cognitive trajectories in Alzheimer's disease. *Proc. Natl. Acad. Sci.* 113, E6535–E6544. doi: 10.1073/pnas.1611073113

frontiers | Frontiers in Computational Neuroscience

TYPE Original Research
PUBLISHED 01 December 2023
DOI 10.3389/fncom.2023.1274824

OPEN ACCESS

EDITED BY
Anees Abrol,
Georgia State University, United States

REVIEWED BY
MohammadMehdi Kafashan,
Washington University in St. Louis, United States
Yoonsuck Choe,
Texas A & M University, United States

*CORRESPONDENCE
Jasmine A. Moore
✉ jasmine.moore@ucalgary.ca

RECEIVED 08 August 2023
ACCEPTED 08 November 2023
PUBLISHED 01 December 2023

CITATION
Moore JA, Wilms M, Gutierrez A, Ismail Z, Fakhar K, Hadaeghi F, Hilgetag CC and Forkert ND (2023) Simulation of neuroplasticity in a CNN-based *in-silico* model of neurodegeneration of the visual system. *Front. Comput. Neurosci.* 17:1274824. doi: 10.3389/fncom.2023.1274824

Simulation of neuroplasticity in a CNN-based *in-silico* model of neurodegeneration of the visual system

Jasmine A. Moore[1,2,3]*, Matthias Wilms[1,2,4], Alejandro Gutierrez[1,2,3], Zahinoor Ismail[2,5], Kayson Fakhar[6], Fatemeh Hadaeghi[6], Claus C. Hilgetag[6,7] and Nils D. Forkert[1,2,4]

[1]Department of Radiology, University of Calgary, Calgary, AB, Canada, [2]Hotchkiss Brain Institute, University of Calgary, Calgary, AB, Canada, [3]Biomedical Engineering Program, University of Calgary, Calgary, AB, Canada, [4]Alberta Children's Hospital Research Institute, University of Calgary, Calgary, AB, Canada, [5]Department of Clinical Neurosciences, University of Calgary, Calgary, AB, Canada, [6]Institute of Computational Neuroscience, University Medical Center Hamburg-Eppendorf (UKE), Hamburg, Germany, [7]Department of Health Sciences, Boston University, Boston, MA, United States

The aim of this work was to enhance the biological feasibility of a deep convolutional neural network-based *in-silico* model of neurodegeneration of the visual system by equipping it with a mechanism to simulate neuroplasticity. Therefore, deep convolutional networks of multiple sizes were trained for object recognition tasks and progressively lesioned to simulate neurodegeneration of the visual cortex. More specifically, the injured parts of the network remained injured while we investigated how the added retraining steps were able to recover some of the model's object recognition baseline performance. The results showed with retraining, model object recognition abilities are subject to a smoother and more gradual decline with increasing injury levels than without retraining and, therefore, more similar to the longitudinal cognition impairments of patients diagnosed with Alzheimer's disease (AD). Moreover, with retraining, the injured model exhibits internal activation patterns similar to those of the healthy baseline model when compared to the injured model without retraining. Furthermore, we conducted this analysis on a network that had been extensively pruned, resulting in an optimized number of parameters or synapses. Our findings show that this network exhibited remarkably similar capability to recover task performance with decreasingly viable pathways through the network. In conclusion, adding a retraining step to the *in-silico* setup that simulates neuroplasticity improves the model's biological feasibility considerably and could prove valuable to test different rehabilitation approaches *in-silico*.

KEYWORDS

deep neural networks, neurodegeneration, Alzheimer's disease, *in-silico*, cognitive computational neuroscience

1 Introduction

Machine learning models have emerged as essential tools for solving complex data-driven classification and regression problems in various domains, and healthcare is no exception. Many machine learning models have been developed and evaluated in the past that, for example, aim to classify if patients have neurological diseases, or aim to predict disease progression and

outcomes based on clinical, imaging, and other assessment data (Lo Vercio et al., 2020; Pinto et al., 2020; James et al., 2021; Rajashekar et al., 2022). Despite their high value for computer-aided diagnosis, these machine learning models cannot be used naively as computational disease models, even when using approaches from the explainable artificial intelligence domain (Linardatos et al., 2020). However, in a more neuroscientific-inspired branch of research, deep learning models are being increasingly investigated as potential tools for modeling how the brain processes information (Kubilius et al., 2016; Yamins and DiCarlo, 2016; Lake et al., 2017; Richards et al., 2019; Lindsay, 2021). These deep neural networks are trained to mimic human behavior and function (Saxe et al., 2021). Although model architectures and training procedures are not identical to biological systems, for example by using backpropagation to learn, deep neural networks remain to be some of the best models of human-level cognition, which may provide a valuable basis for in-silico models of neurological diseases (Güçlü and van Gerven, 2014; Khaligh-Razavi and Kriegeskorte, 2014; Kaiser et al., 2017; Cichy and Kaiser, 2019; Perconti and Plebe, 2020). Establishing an *in-silico* model of neurological disease would, for example, allow us to obtain a better understanding of the effects of axonal and neuronal damage, and other pathological processes such as tau deposition on essential brain functions. Here, the term "*in-silico*" refers to the usage of computer methods for understanding biological processes in the living organism (Winder et al., 2021). Deep convolutional neural networks (CNNs), a deep learning model architecture specifically designed for solving computer vision problems such as object recognition, were originally inspired by the structure of neurons and synapses found in the mammalian visual cortex (Rawat and Wang, 2017). The concepts used to inspire CNNs date back to early models of the visual system, postulated by Hubel and Wiesel (1962, 1968). An emerging field that is gaining momentum recently involves using deep learning models as an abstraction of a healthy human brain, which can then be utilized as a basis for simulating neurodegenerative diseases (Tuladhar et al., 2021; Moore et al., 2022). Since CNNs were specifically designed for vision tasks and were modeled after information processing patterns in the mammalian brain, they can be used to model neuronal injuries that occur in the visual cortex, as for example the case in posterior cortical atrophy (PCA). PCA is characterized by the rapid deterioration and thinning of visual cortical areas such as V1, V2, V3, and V4, leading to a loss of visual recognition abilities in patients (Crutch et al., 2012; Maia da Silva et al., 2017). PCA is usually a variant of AD, caused by the same protienopathies. Previous research has established parallels between synaptic and neuronal pruning in CNNs and *in silico* models and the onset of posterior cortical atrophy (Moore et al., 2022). In this work, we compared the effects of applying either progressive neuronal or synaptic injury using an established CNN architecture (VGG19) as an initially cognitively healthy model. The CNN was trained to perform object recognition on 2D images, akin to the Boston Naming Test (BNT) or other similar neuropsychological assessments testing visual function (Williams et al., 1989). During the BNT, patients are presented with stimuli in the form of line drawings of items of 60 categories and are asked to identify the objects. Therefore, it may be possible to draw parallels between object recognition tasks of the CNN and cognitive assessments such as the BNT.

However, a shortcoming of this work was the method in which injury was applied to the network, which was not biologically realistic. Specifically, injury was progressively and statically imposed, without allowing the model to update weights or be exposed to any new training data. Thus, the aim of the present study was to expand upon and improve Moore et al.'s (2022) work by adding the crucial mechanism of simulated neuroplasticity via retraining as shown in Figure 1. In the present study, synapses are specifically set to zero to simulate full synaptic death in the visual cortex. While other pathological mechanisms may precede synaptic death and lead to a functional decline in synapses over time, synaptic death is the ultimate effect of any dementia disease. The ability of the human brain to develop new synapses is very limited in adults so that the remaining synapses need to be retrained to account for the loss and as a means of neuroplasticity. Thus, in this study, we froze the injured weights to prevent them from being subjected to the retraining process to simulate disease effects in humans where dead synapses cannot be simply replaced by new ones (John and Reddy, 2021). Furthermore, one could argue that a standard VGG19 network is overparameterized, and thus has too much reserve capacity as compared to human cognitive reserve, to be a biologically realistic in-silico model when studying injuries. Therefore, in the present study, we investigate two different models as a baseline for cognitively healthy object recognition. The analysis was performed using a full VGG19 model as well as using a highly pruned version of the VGG19 to examine the effects of plasticity as a function of imposed injury and number of model parameters or synapses.

While model compression and pruning are active branches of deep learning research, our paradigm of 'injury' does not follow typical pruning methods, which aim to reduce the number of parameters in a model while retaining full function (Choudhary et al., 2020). In contrast, we use progressive random pruning followed by retraining to simulate the cognitive effects of a neurodegenerative disease as a function of abnormal levels of atrophy. We find that by adding iterative retraining with every pruning step of synaptic ablation, the decline of visual cognition is much smoother and more similar to what is seen in patients with Alzheimer's disease (AD) (Mattsson et al., 2017).

2 Materials and methods

2.1 Models and data

The basis for the cognitively healthy object recognition model is a VGG19-like model with batch normalization trained on the CIFAR10 dataset (Russakovsky et al., 2015). CIFAR10 is a commonly used dataset for computer vision research, which consists of 60,000 32×32 natural color images. The dataset consists of 10 classes: plane, car, bird, cat, deer, dog, frog, horse, ship, and truck, with 6,000 images in each class. The train/test split used is 50,000 and 10,000 images, respectively. This dataset was chosen due to the relative simplicity and ease of computational load. The model architecture used in this work is comprised of five convolutional blocks, each followed by a batch normalization layer, ending with four max-pooling layers, and finally, a Softmax activation with 10 nodes corresponding to the 10 classes in the dataset. Our model was pretrained on ImageNet and fine-tuned on CIFAR10 for 100 epochs with a learning rate of 0.001, using a batch size of 128, and a stochastic gradient descent optimizer with momentum 0.9. After training, the full model achieves an accuracy of 93.74% on the test set of images. A VGG19 model was chosen for this research as it has been to have high correlation with mammalian

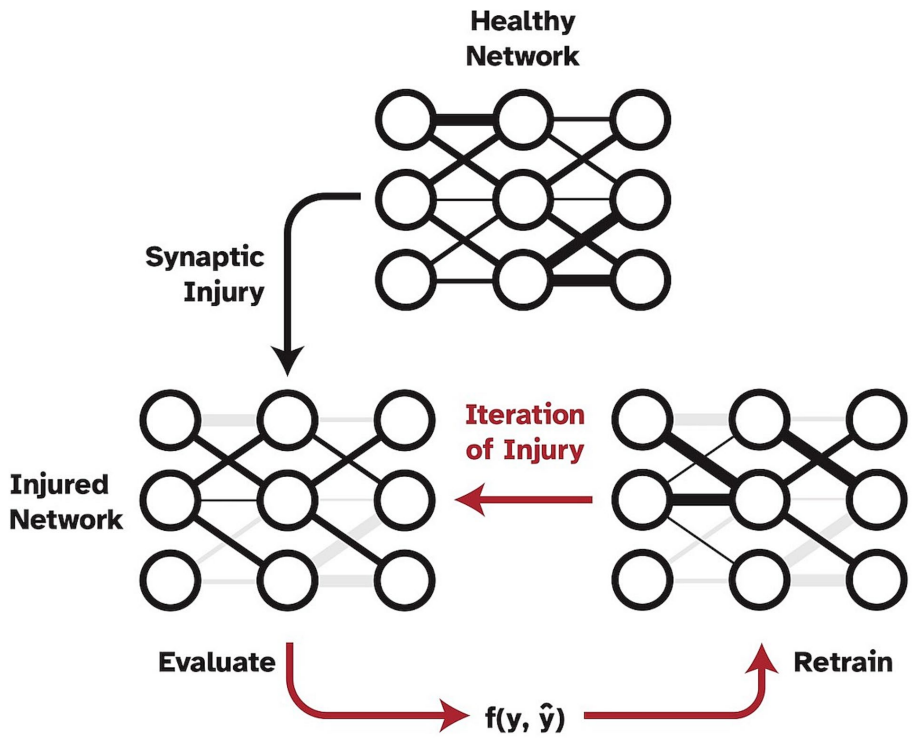


FIGURE 1
Pipeline of progressive synaptic injury with the added mechanism of neuroplasticity. After each iteration of synaptic damage, the model is retrained on the original training split of data and evaluated. Ablated synaptic weights are shown in grey.

neuronal activation data and is widely accepted as a SOTA baseline model in computer vision tasks.

Previous research has shown that VGG19 models may be largely overparameterized, especially for classifying CIFAR10, due to their retention of high levels of accuracy when subjected to optimized pruning techniques (Frankle and Carbin, 2018; Ayinde et al., 2019). More specifically, they likely have learned unnecessary or redundant pathways due to the enormous number of synapses and neurons they are equipped with. The brain has also been shown to be overparameterized, but is likely much more constrained by energy usage and physical space (Drachman, 2005; Mizusaki and O'Donnell, 2021). Thus, to investigate potential spurious results that are driven by overparameterization, rather than model plasticity abilities, and to perform experiments in a more physically constrained setting, we also investigated a considerably more optimized compressed model. To this end, we performed structured model pruning on the trained full VGG19 model. Model compression was informed by graph dependencies using methods developed and described in Li et al. (2016) and Fang et al. (2023). Filters and associated weights were removed simultaneously based on their L1 norm until the model inference speed, in terms of floating-point operations (FLOP), was increased by a user-defined amount. To probe the amount of structured pruning the model could tolerate before significant declines in accuracy, we performed model compression multiple times. The compression resulted in models that had been sped up 2x, 3x, and 4x from the original inference speed while maintaining similar, high accuracies. We found that increasing FLOP by three times with respect to the original VGG19 model resulted in a compressed model with only 8.54% of the original weights. Despite this considerable reduction of weights, this compressed model retained an accuracy of 93.3% on the test set. All model injury and retraining experiments described below were performed on both the full VGG19 and this compressed version. Experiments were conducted using Pytorch 1.13 on an NVIDIA GeForce RTX 3090 GPU with 24GB of memory.

2.2 Synaptic ablation and retraining

Synaptic ablation was imposed on the network in a uniformly disperse and progressive manner as originally proposed in Moore et al. (2022). This 'injury' type was implemented by setting weights from convolutional layers and dense layers in the network to zero, effectively severing the connections between nodes. This approach is akin to progression of synaptic damage seen in neurological diseases that accelerate atrophy rates in the brain, such as posterior cortical atrophy. It should be noted that this synaptic injury and retraining is not the same as optimized model pruning, and thus is more biologically reasonable as an in-silico paradigm. We imposed random synaptic ablation at a step rate of $1-(1-\gamma)^n$ where γ is the relative fraction of weights being ablated to the remaining uninjured weights in the network, and n is the number of iterations of injury. Once a synapse is ablated, it can no longer be used by the model and is excluded from the retraining process.

In our experiments, γ was set to 0.2 (20% of weights ablated) and n was set to 15 iterations as this was found to be representative of injury resolution while maintaining reasonable computational requirements. Following each iteration of injury, we retrained the model on the training split of data using the same initial training

parameters for three epochs to investigate how model performance could be regained. With each retraining step, the optimizer was reinitialized while the injured weights remained set to zero so that the model had to find alternative pathways to regain test performance. We performed this analysis ten times to reduce the risk that biasing effects related to the order in which synapses were randomly ablated are affecting the results.

2.3 Representational dissimilarity matrices

Representational dissimilarity matrices (RDMs) were computed to examine the changes in internal activations and representations of categorized data of both the injured and retrained networks when compared to the respective baseline, healthy networks. RDMs are routinely used to quantitatively correlate brain-activity, behavioral measurement, and computational modeling (Kriegeskorte et al., 2008; Khaligh-Razavi and Kriegeskorte, 2014; Mehrer et al., 2020). RDMs measure the representational distance between two sets of model activations given different inputs and can be used to visualize representational space. RDMs were generated by pairwise comparison between activations of the network's penultimate layer for all test set images using Pearson's correlation coefficient. We constructed RDMs for each iteration of both network ablation and retraining, and then compared them to the healthy network's RDM using Kendall's tau correlation coefficient. This approach effectively enabled us to quantify the effects of both injury and retraining on internal activations of the networks. Comparison of RDMs of the model as it is progressively injured and retrained allows for the examination of how the relative structure of representational space is affected and reconstructed with injury and retraining.

2.4 Brain-score

The Brain-Score is a widely used metric that has been developed to analyze how the CNN model activations are correlated and predictive of mammalian neural activation data (Schrimpf et al., 2018, 2020). Within this context, VGG19 has been found to be a relatively highly ranked model in terms of neural predictivity. We wanted to investigate how imposing injury to a 'healthy' VGG19 model affected its Brain-Score. Therefore, we created our baseline Brain-Scores by following the methods outlined by Schrimpf et al. (2018) and used the publicly available neural recording benchmarks for visual areas V1, V2, V4, and IT (Freeman et al., 2013; Majaj et al., 2015). The neural recordings dataset contains macaque monkey neural responses to 2,560 naturalistic images. More detail on this data and the methods we used to calculate Brain-Score can be found in the publicly available code from Schrimpf et al. We used the neural benchmarks to establish how well the internal representations of our CNN models matched internal representations of mammals. In computing the Brain-Score, we compute a composite measure of neural predictivity scores for all aforementioned visual areas. Neural predictivity is evaluated on how well the responses, or internal activity in our CNNs predicted the neural activity in the biological neural recordings. Consistent with the literature, we performed this analysis using principal components analysis to reduce the dimensionality of model activations to 1,000 components, and then used partial least squares regression with 25 components to correlate the CNN model activation to mammalian neural activations. Correlation coefficients were calculated for each of the publicly available benchmarks and then averaged to calculate an average Brain-Score. Brain-Score values were analyzed for progressive iterations of injury and retraining in the models. It should be noted that only publicly available, neural benchmark datasets were used in our Brain-Score calculations so there were disparities between our 'healthy' VGG19 Brain-Score and that reported on the Brain-Score leaderboard.

3 Results

3.1 Accuracy

Baseline model accuracies were 93.7 and 93.3% for the full VGG19 and compressed VGG19, respectively. The results showed that model performance was immediately affected by the first application of random synaptic injury (20% of synapses randomly deleted), leading to a large drop in object recognition accuracy in both the full and compressed models across all classes of the test split of the CIFAR10 benchmark dataset. Quantitatively, after the first iteration of injury, the full model's accuracy on the test set suffered a drop from 93.7% to a mere 10.0%, essentially chance level. Interestingly, the accuracy was substantially restored to 92.4% ± 0.001% after three epochs of the retraining iteration. With each iteration of injury, this pattern of large accuracy drops continued to repeat, with the model again tending to perform only at chance level (10% accuracy). However, retraining continued to improve model accuracy by a large margin, even until 96.5% of initial synapses had been removed. After this point, the model could only regain accuracy levels of 77.6% ± 0.010% with retraining. These effects are shown in Figure 2A. Similar to the full model, the compressed model also proved to be largely affected by introducing plasticity to the injury paradigm. With each iteration of synaptic injury, the compressed model accuracy plummeted to chance level accuracy. Remarkably, even with the initial healthy network containing a mere 8.54% of the size of the full VGG19, on average the compressed model was able to regain high levels of object recognition accuracy after retraining. Even after 48.8% of synapses were injured, the compressed model recovered 89.6% ± 0.005% accuracy with the retraining iteration. At injury levels of 83% and higher, the compressed model exhibited large standard deviations in accuracy over the ten trials that were performed (~ ± 20%) (Figure 2B). This may be due to the extreme synaptic sparsity that is associated with high injury levels. If a highly salient pathway in the model with limited synapses is ablated, the model may not be able to recover accuracy with retraining.

To further investigate the relationship between model size and recovery with retraining, we examined accuracy levels based on the total number of parameters within the full and compressed models. We inspected the point in the injury progression where, after retraining, the two models displayed close to identical levels of accuracy, and after which the retrained accuracies no longer displayed similar levels of accuracy. This point was found after the full model (20.04 million parameters) had been injured so that 83.2% of synapses had been removed, leaving the model with 3.36 million parameters. After retraining, the full model showed accuracy levels of 87.4% ± 0.003% on the test set. The point at which the compressed model showed similar levels of accuracy (87.0% ± 0.011%) was after 67.2% of its original synapses had been injured. This level of injury left

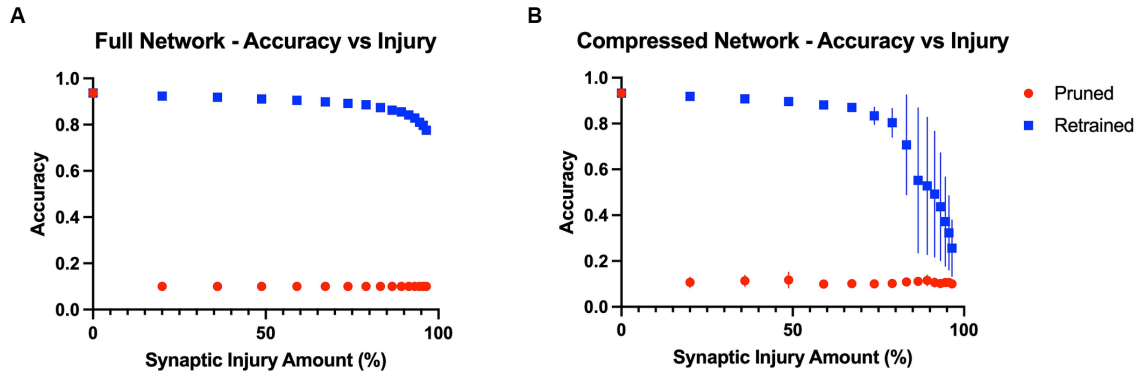


FIGURE 2
Model accuracies as a function of progressive injury and retraining. **(A)** Model accuracy and standard deviations as a function of progressive synaptic damage and retraining for the full VGG19 model. The standard deviations in accuracy of the full model are extremely low (~ $\pm$ *0.005%*). The model immediately has a substantial drop in accuracy after 20% of the synapses are removed but regains close to complete function after retraining on the training set. Substantial levels of accuracy are regained with the addition of retraining, even at extremely high levels of injury. **(B)** Compressed model accuracy as a function of synaptic damage. Retraining leads to large gains in accuracy until injury levels of 85% and higher, at which point the model shows a steeper decline in accuracy even with retraining.

the compressed model with a mere 0.560 million parameters. Furthermore, we assessed accuracy levels when the two models contained a similar number of parameters. This occurred when the full model had been injured by 96.5%, and thus had 0.705 million parameters, and when the compressed model had been injured by 59.0% and had 0.700 million parameters. The accuracy levels were significantly different, at 77.6% ± 0.010 and 88.1% ± 0.010%, respectively. These results indicate that model size and overparameterization are not the sole contributing factors to the impact adding plasticity has on the degenerative *in-silico* paradigm.

3.2 Representational dissimilarity matrices

In line with the results of the model accuracy evaluation, the internal representations of the models were able to regenerate and recover with retraining after injury. In the first iteration of injury and retraining (20% of synapses ablated) on the injured full model, the correlation to the healthy RDM revealed a Kendall's tau of 0.25 ± 0.05. After retraining for three epochs, the model was able to reconstruct activations more similar to those of the healthy model, resulting in a Kendall's tau value of 0.78 ± 0.01. Comparatively, the compressed model's internal activations also degraded after the first iteration of damage and showed a Kendall' tau value of 0.22 ± 0.04. Upon retraining, the internal activations displayed an increased correlation to the healthy activations that resulted in a Kendall's tau of 0.76 ± 0.02. Figures 3C,D show how retraining after each injury step led to regaining category-distinguishable activations and a smooth cognitive decline. A qualitative examination also reveals how the network activations were affected through injury and retraining. As seen in Figures 3A,B, the uninjured networks initially had clearly defined activations grouped according to object classes in the CIFAR10 dataset. Upon injury, the networks lost this categorical representation and the RDMs became noisy. After retraining, however, categorical structure between the classes was regained. This trend continued progressively as injury and retraining steps were applied to the full network, but at high levels of injury, there came a point where there was no longer a difference in Kendall's tau correlation between injured and retrained RDMs (e.g., at injury levels higher than 95%).

3.3 Brain-score assessment

The Brain-Score was computed for progressive steps of injury and retraining for both models (Table 1). Brain-Scores are reported as the mean Brain-Score of the four brain regions (V1, V2, V4, and inferior temporal (IT) cortices) that were used in the correlation analysis. The averages are reported with the standard deviations. Overall, it was found that Brain-Scores decreased in value as synaptic injury increased. Thus, the injured models tended to be less 'brain-like' than both of the healthy models (i.e., the uninjured full and compressed models) according to the scores. Following retraining, the models regained a level of Brain-Score comparable to healthy models. These finding indicate that adding retraining allows models to retain 'brain-like' features in terms of internal activations, while still exhibiting functional deficits (i.e., loss of object recognition abilities).

4 Discussion

4.1 Main findings

The proposed framework for *in-silico* modeling of visual impairments associated with neurological diseases and retraining to model neuroplasticity may lead to improved disease understanding. With further development, we may establish more biologically realistic computer models that can be injured in different ways, instead of having to collect data from hundreds of patients with different disease patterns to obtain similar information. Furthermore, the development of this branch of research may also enable us to investigate the benefit of potential interventions to re-learn specific brain functions, for example, cognitive rehabilitation therapies. This work specifically enhances the feasibility of these models by including neuroplasticity

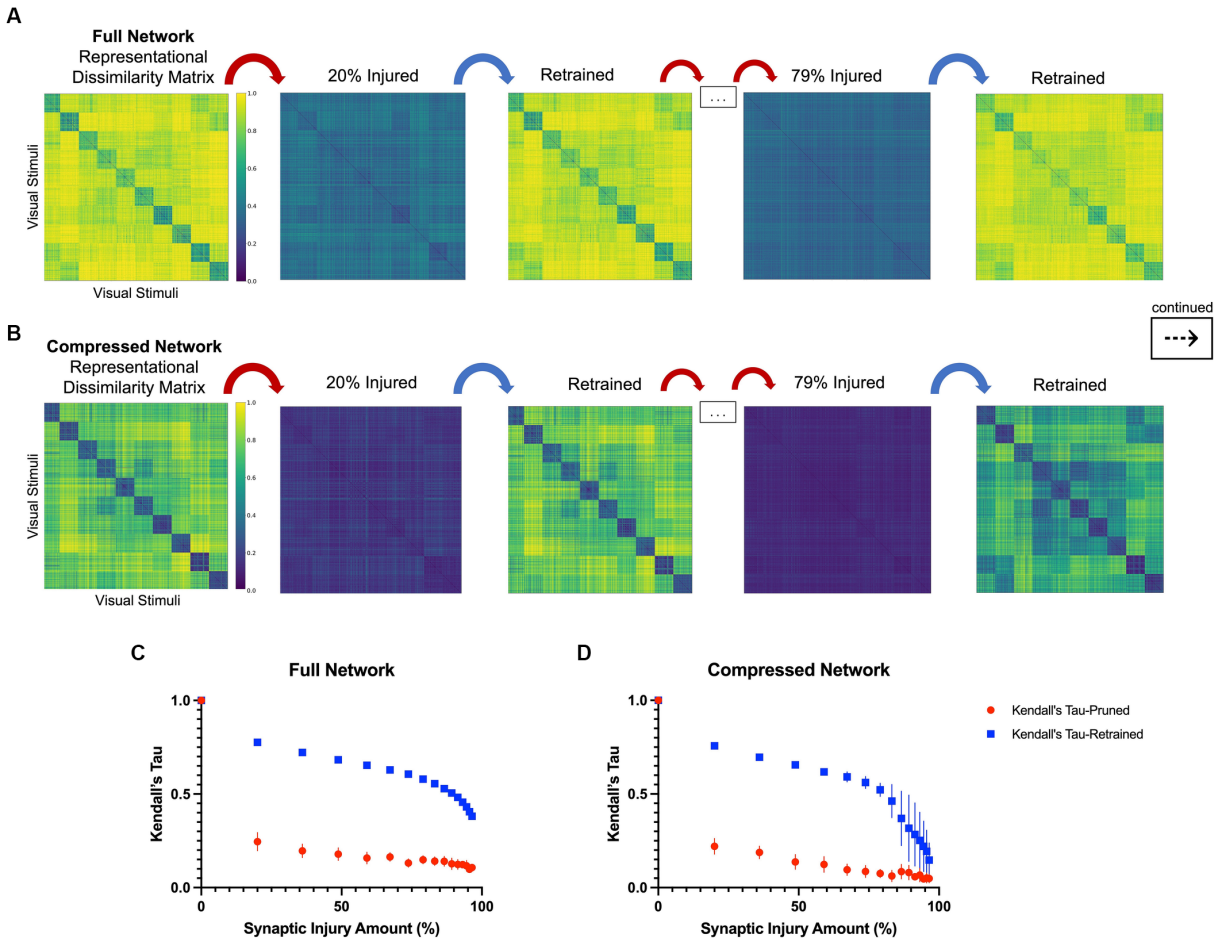


FIGURE 3
Representational dissimilarity matrices with injury and retraining, as described in Section 2.3. **(A)** A qualitative examination of representational dissimilarity matrices as an initial iteration and a later iteration of injury and retraining are completed in the full model and **(B)** in the compressed model. Before injury, both models have distinct activations for each of the 10 different classes in CIFAR10. **(C)** Kendall's tau correlation as a function of synaptic injury in the full model. As injury and retraining are imposed, the activations first lose their categorical nature, but are able to recover some of it with retraining. **(D)** Kendall's tau correlation as a function of synaptic injury for the compressed model. A similar trend is seen to that in the full model albeit there is larger variance in correlation values with higher levels of injury.

TABLE 1 **Brain-Scores are reported for the injury and retraining steps as injury level increases.**

Full network									
Injury amount	0% (healthy)	20%	48.8%	67.2%	79.0%	86.6%	91.4%	94.5%	96.5%
Brain score (injured)	0.342 (± 0.028)	0.335 (± 0.021)	0.329 (± 0.024)	0.318 (± 0.025)	0.309 (± 0.021)	0.312 (± 0.021)	0.320 (± 0.020)	0.244 (± 0.017)	0.278 (± 0.021)
Brain score (retrained)		0.341 (± 0.027)	0.345 (± 0.028)	0.348 (± 0.027)	0.341 (0.024)	0.336 (± 0.023)	0.349 (± 0.024)	0.338 (± 0.024)	0.342 (± 0.025)
Compressed network									
Brain score (injured)	0.343 (±0.028)	0.291 (± 0.025)	0.290 (± 0.024)	0.276 (± 0.018)	0.279 (± 0.018)	0.314 (± 0.020)	0.258 (± 0.020)	0.267 (± 0.020)	0.266 (± 0.017)
Brain score (retrained)		0.348 (± 0.026)	0.342 (± 0.023)	0.343 (± 0.022)	0.346 (0.026)	0.342 (± 0.025)	0.335 (± 0.025)	0.320 (± 0.024)	0.300 (± 0.020)

The Brain-Score values tend to decrease as injury is progressively applied, but become comparable to the healthy models' Brain-Scores with retraining.

in the simulated disease progression. It was found that this approach leads to a more biologically relevant pattern of cognitive decline with respect to the load of injury. The human brain has remarkable abilities to reorganize pathways, develop new connections, and arguably even create new neurons, typically referred to as neuroplasticity or as the neurocognitive reserve (Esiri and Chance, 2012). Simply damaging a

network all at once without allowing it to retrain in between or during injury ignores this important ability of the human brain. Previous works using CNNs to model neurodegenerative diseases used a static injury paradigm that led to extreme loss of object recognition abilities even with low levels (i.e., 15–20%) of synapses injured (Lusch et al., 2018; Tuladhar et al., 2021; Moore et al., 2022).

The main finding of the current study is that with the incorporation of retraining to simulate neuroplasticity after the progression of injury, the models' object recognition abilities progressively decline at a much smoother and slower rate than without retraining. This slow decline is more akin to the degradation of cognitive abilities seen in patients with AD and it's PCA variant (Hodges et al., 1995; Fox et al., 1999; Jefferson et al., 2006) than the decline patterns previously observed. Expanding upon this previous research simulating statically imposed injury, here we developed a framework that is able to simulate irreversible injury, while the unaffected filters and weights were subjected to 're-learning' processes to stimulate reorganization of the information flow that makes use of existing reserve capacities in the injured model. We found that the retrained models were able to compensate for the damaged pathways (synapses) and reconstruct the original activation patterns of the healthy models to a large extent when presented with images in the test set. Additionally, in this work we validate that this ability was not a direct function of initial model size. Generally, it is reasonable to expect that after being injured, an overparameterized model may exhibit large gains in task performance with retraining. However, here we show that a model that is much more compressed, and thus highly optimized in terms of number of parameters, displays remarkably similar abilities to re-gain task performance using increasingly minimal available pathways through the network, which is more similar to the human brain. Thus, we believe that the introduction of the biologically important concept of neuroplasticity, which equips our CNNs with a retraining mechanism, can be seen as an important step toward developing biologically more meaningful *in-silico* models of neurodegenerative diseases and other injuries of the human brain.

4.2 Limitations and future work

One important limitation of this work is related to the notable differences in information processing between CNNs and the biological visual system (Lonnqvist et al., 2021) (e.g., convolutional filters are global in a CNN while the human visual system also has filtering units that are responsible for certain parts of the receptive field). However, while this remains to be true, the object recognition performance of CNNs is comparable to that of humans, and CNNs have the ability to predict neural activation in the primate visual cortex better than any other computational model to date (Cadieu et al., 2014; Khaligh-Razavi and Kriegeskorte, 2014; Yamins et al., 2014). While a CNN works very differently at the neural level, the general organization is broadly representative of a visual network with a hierarchy of connections. We see this work as utilizing the similarities between CNNs and the visual cortex to further develop the feasibility of using deep learning models as an *in-silico* model for neurodegenerative diseases. The success of convolutional neural networks for predicting neural activity in the visual cortex makes them excellent models for modelling visual cognition. In theory, the setup presented in this work can be extended to other brain regions and cognitive or motor functions. For example, language models could be used to investigate how lesions in the auditory and frontal cortex affect language function. However, it should be noted that more research is probably needed first to investigate how similar other deep learning models for other tasks are to the human brain akin to the comparably extensive research investigating the biological feasibility of CNNs. Furthermore, while CNNs are well accepted models of the human visual system, there may be opportunities to increase the similarities to the human brain even more (Lake et al., 2015). Future work may be extended to simulate different neural damage, such as more localized lesions to model conditions like cerebral stroke or multiple sclerosis.

As previously mentioned, patients with AD often undergo cognitive assessments that probe visual object recognition abilities and recall (i.e., the Boston Naming Test). Such visual assessments together with longitudinal, high-resolution MRI data to assess atrophy could be used in future to optimize and validate the proposed in-silico model of AD but is outside the scope of this work.

Crucial future directions for this work will be to further investigate the details surrounding the iterative retraining process, as well as more realistically represent disease progression. Such investigation will allow for the exploration of rehabilitation strategies in terms of what methods of retraining enable in-silico models to regain the most function. Additionally, we can provide models with training data that are directly related to the types of errors the models begin to make with initial injury. This could be compared against re-training strategies that would simply re-use all initial training data. In addition, it may be important to evaluate the effects of other variables such as training the network on new data rather than previously seen data, or adjusting the number of epochs used in one iteration of retraining. Some studies have identified that specific task-oriented cognitive training strategies (i.e., face recognition practice) show higher memory related brain activity and task performance for patients with Alzheimer's disease (Cotelli et al., 2006; Choi and Twamley, 2013). Notably, it may be possible to model different pathological processes of AD by gradually decaying weight values to zero rather than fully removing synapses in a single iteration. This could, for example, be used to simulate the accumulation of hyperphosphorylated tau, which is often assumed to precede synaptic death.

By probing these types of differences in network plasticity and recovery, it may be possible to identify optimal intervention strategies and relate these findings to rehabilitation techniques used in patients with dementia. This study lays further groundwork toward using deep learning models to effectively simulate disease progression, with (bright) potential to develop cutting edge in-silico models.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

JM: Conceptualization, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft. MW:

Conceptualization, Methodology, Writing – review & editing. AG: Software, Writing – review & editing. ZI: Conceptualization, Resources, Writing – review & editing. KF: Methodology, Visualization, Writing – review & editing. FH: Conceptualization, Visualization, Writing – review & editing. CH: Conceptualization, Writing – review & editing. NF: Conceptualization, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the Canada Research Chairs program, the River Fund at Calgary Foundation, Natural Sciences and Engineering Research Council of Canada (NSERC), Alberta Innovates – Data Enabled Technologies, and from the NSERC – Hotchkiss Brain Institute Brain CREATE program. The funding agencies had no role in the study design, collection, analysis and interpretation of data, nor preparation, review or approval of the manuscript. KF was supported by Deutsche Forschungsgemeinschaft (DFG) grant SFB 936/A1, FH by DFG grant TRR 169/A1, and CH by DFG grants SFB 936/A1, SFB 936/Z3; TRR 169/A2 and SFB 1461/A4.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

References

Ayinde, B. O., Inanc, T., and Zurada, J. M. (2019). Redundant feature pruning for accelerated inference in deep neural networks. *Neural Netw.* 118, 148–158. doi: 10.1016/j.neunet.2019.04.021

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., et al. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Comput. Biol.* 10:e1003963. doi: 10.1371/journal.pcbi.1003963

Choi, J., and Twamley, E. W. (2013). Cognitive rehabilitation therapies for Alzheimer's disease: a review of methods to improve treatment engagement and self-efficacy. *Neuropsychol. Rev.* 23, 48–62. doi: 10.1007/s11065-013-9227-4

Choudhary, T., Mishra, V., Goswami, A., and Sarangapani, J. (2020). A comprehensive survey on model compression and acceleration. *Artif. Intell. Rev.* 53, 5113–5155. doi: 10.1007/s10462-020-09816-7

Cichy, R. M., and Kaiser, D. (2019). Deep neural networks as scientific models. *Trends Cogn. Sci.* 23, 305–317. doi: 10.1016/j.tics.2019.01.009

Cotelli, M., Calabria, M., and Zanetti, O. (2006). Cognitive rehabilitation in Alzheimer's disease. *Aging Clin. Exp. Res.* 18, 141–143. doi: 10.1007/BF03327429

Crutch, S. J., Lehmann, M., Schott, J. M., Rabinovici, G. D., Rossor, M. N., and Fox, N. C. (2012). Posterior cortical atrophy. *Lancet Neurol.* 11, 170–178. doi: 10.1016/S1474-4422(11)70289-7

Drachman, D. A. (2005). Do we have brain to spare? *Neurology* 64, 2004–2005. doi: 10.1212/01.WNL.0000166914.38327.BB

Esiri, M. M., and Chance, S. A. (2012). Cognitive reserve, cortical plasticity and resistance to Alzheimer's disease. *Alzheimers Res. Ther.* 4:7. doi: 10.1186/alzrt105

Fang, G., Ma, X., Song, M., Bi Mi, M., and Wang, X. (2023). Depgraph: towards any structural pruning'. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16091–16101.

Fox, N. C., Scahill, R. I., Crum, W. R., and Rossor, M. N. (1999). Correlation between rates of brain atrophy and cognitive decline in AD. *Neurology* 52, 1687–1689. doi: 10.1212/wnl.52.8.1687

Frankle, J., and Carbin, M. (2018). The lottery ticket hypothesis: Finding sparse, Trainable Neural Networks. *ar Xiv.* doi: 10.48550/arXiv.1803.03635

Freeman, J., Ziemba, C. M., Heeger, D. J., Simoncelli, E. P., and Movshon, J. A. (2013). A functional and perceptual signature of the second visual area in primates. *Nat. Neurosci.* 16, 974–981. doi: 10.1038/nn.3402

Güçlü, U., and van Gerven, M. A. J. (2014). Unsupervised feature learning improves prediction of human brain activity in response to natural images. *PLoS Comput. Biol.* 10:e1003724. doi: 10.1371/journal.pcbi.1003724

Hodges, J. R., Graham, N., and Patterson, K. (1995). Charting the progression in semantic dementia: implications for the organisation of semantic memory. *Memory* 3, 463–495. doi: 10.1080/09658219508253161

Hubel, D. H., and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* 160, 106–154. doi: 10.1113/jphysiol.1962.sp006837

Hubel, D. H., and Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *J. Physiol.* 195, 215–243. doi: 10.1113/jphysiol.1968.sp008455

James, C., Ranson, J. M., Everson, R., and Llewellyn, D. J. (2021). Performance of machine learning algorithms for predicting progression to dementia in memory clinic patients. *JAMA Netw. Open* 4:e2136553. doi: 10.1001/jamanetworkopen.2021.36553

Jefferson, A., et al. (2006). Object perception impairments predict instrumental activities of daily living dependence in Alzheimer's disease. *J. Clin. Exp. Neuropsychol.* 28, 884–897. doi: 10.1080/13803390591001034

John, A., and Reddy, P. H. (2021). Synaptic basis of Alzheimer's disease: focus on synaptic amyloid beta, P-tau and mitochondria. *Ageing Res. Rev.* 65:101208. doi: 10.1016/j.arr.2020.101208

Kaiser, L., Gomez, A. N., Shazeer, N., Vaswani, A., Parmar, N., Jones, L., et al. (2017). One model to learn them all. *ar Xiv.* doi: 10.48550/arXiv.1706.05137

Khaligh-Razavi, S.-M., and Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.* 10:e1003915. doi: 10.1371/journal.pcbi.1003915

Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* 2:4. doi: 10.3389/neuro.06.004.2008

Kubilius, J., Bracci, S., and Op de Beeck, H. P. (2016). Deep neural networks as a computational model for human shape sensitivity. *PLoS Comput. Biol.* 12:e1004896. doi: 10.1371/journal.pcbi.1004896

Lake, B. M., Salakhutdinov, R., and Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science* 350, 1332–1338. doi: 10.1126/science.aab3050

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2017). Building machines that learn and think like people. *Behav. Brain Sci.* 40:e253. doi: 10.1017/S0140525X16001837

Li, H., Kadav, A., Durdanovic, I., Samet, H., and Graf, H. P. (2016). Pruning filters for efficient conv nets. *arXiv*. doi: 10.48550/arXiv.1608.08710

Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2020). Explainable AI: a review of machine learning interpretability methods. *Entropy* 23:18. doi: 10.3390/e23010018

Lindsay, G. W. (2021). Convolutional neural networks as a model of the visual system: past, present, and future. *J. Cogn. Neurosci.* 33, 2017–2031. doi: 10.1162/jocn_a_01544

Lo Vercio, L., Amador, K., Bannister, J. J., Crites, S., Gutierrez, A., MacDonald, M. E., et al. (2020). Supervised machine learning tools: a tutorial for clinicians. *J. Neural Eng.* 17:062001. doi: 10.1088/1741-2552/abbff2

Lonnqvist, B., Bornet, A., Doerig, A., and Herzog, M. H. (2021). A comparative biology approach to DNN modeling of vision: a focus on differences, not similarities. *J. Vis.* 21:17. doi: 10.1167/jov.21.10.17

Lusch, B., Weholt, J., Maia, P. D., and Kutz, J. N. (2018). Modeling cognitive deficits following neurodegenerative diseases and traumatic brain injuries with deep convolutional neural networks. *Brain Cogn.* 123, 154–164. doi: 10.1016/j.bandc.2018.02.012

Maia da Silva, M. N., Millington, R. S., Bridge, H., James-Galton, M., and Plant, G. T. (2017). Visual dysfunction in posterior cortical atrophy. *Front. Neurol.* 8:389. doi: 10.3389/fneur.2017.00389

Majaj, N. J., Hong, H., Solomon, E. A., and DiCarlo, J. J. (2015). Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human Core object recognition performance. *J. Neurosci.* 35, 13402–13418. doi: 10.1523/JNEUROSCI.5181-14.2015

Mattsson, N., Andreasson, U., Zetterberg, H., and Blennow, K.for the Alzheimer's Disease Neuroimaging Initiative (2017). Association of Plasma Neurofilament Light with neurodegeneration in patients with Alzheimer disease. *JAMA Neurol.* 74, 557–566. doi: 10.1001/jamaneurol.2016.6117

Mehrer, J., Spoerer, C. J., Kriegeskorte, N., and Kietzmann, T. C. (2020). Individual differences among deep neural network models. *Nat. Commun.* 11:5725. doi: 10.1038/s41467-020-19632-w

Mizusaki, B. E. P., and O'Donnell, C. (2021). Neural circuit function redundancy in brain disorders. *Curr. Opin. Neurobiol.* 70, 74–80. doi: 10.1016/j.conb.2021.07.008

Moore, J. A., Tuladhar, A., Ismail, Z., Mouches, P., Wilms, M., and Forkert, N. D. (2022). Dementia in convolutional neural networks: using deep learning models to simulate neurodegeneration of the visual system. *Neuroinformatics* 21, 45–55. doi: 10.1007/s12021-022-09602-6

Perconti, P., and Plebe, A. (2020). Deep learning and cognitive science. *Cognition* 203:104365. doi: 10.1016/j.cognition.2020.104365

Pinto, M. F., Oliveira, H., Batista, S., Cruz, L., Pinto, M., Correia, I., et al. (2020). Prediction of disease progression and outcomes in multiple sclerosis with machine learning. *Sci. Rep.* 10:21038. doi: 10.1038/s41598-020-78212-6

Rajashekar, D., Wilms, M., MacDonald, M. E., Schimert, S., Hill, M. D., Demchuk, A., et al. (2022). Lesion-symptom mapping with NIHSS sub-scores in ischemic stroke patients. *Stroke Vasc. Neurol.* 7, 124–131. doi: 10.1136/svn-2021-001091

Rawat, W., and Wang, Z. (2017). Deep convolutional neural networks for image classification: a comprehensive review. *Neural Comput.* 29, 2352–2449. doi: 10.1162/NECO_a_00990

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., et al. (2019). A deep learning framework for neuroscience. *Nat. Neurosci.* 22, 1761–1770. doi: 10.1038/s41593-019-0520-2

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). Image net large scale visual recognition challenge. *Int. J. Comput. Vis.* 115, 211–252. doi: 10.1007/s11263-015-0816-y

Saxe, A., Nelli, S., and Summerfield, C. (2021). If deep learning is the answer, what is the question? *Nat. Rev. Neurosci.* 22, 55–67. doi: 10.1038/s41583-020-00395-8

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Iss, E. B., et al. (2018). Brain-score: which artificial neural network for object recognition is most brain-like? *bio Rxiv*. doi: 10.1101/407007

Schrimpf, M., Kubilius, J., Lee, M. J., Ratan Murty, N. A., Ajemian, R., and DiCarlo, J. J. (2020). Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron* 108, 413–423. doi: 10.1016/j.neuron.2020.07.040

Tuladhar, A., Moore, J. A., Ismail, Z., and Forkert, N. D. (2021). Modeling neurodegeneration in silico with deep learning. *Front. Neuroinform.* 15:748370. doi: 10.3389/fninf.2021.748370

Williams, B. W., Mack, W., and Henderson, V. W. (1989). Boston naming test in Alzheimer's disease. *Neuropsychologia* 27, 1073–1079. doi: 10.1016/0028-3932(89)90186-3

Winder, A., Wilms, M., Fiehler, J., and Forkert, N. D. (2021). Treatment efficacy analysis in acute ischemic stroke patients using in silico modeling based on machine learning: a proof-of-principle. *Biomedicine* 9:1357. doi: 10.3390/biomedicines9101357

Yamins, D. L. K., and DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* 19, 356–365. doi: 10.1038/nn.4244

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U. S. A.* 111, 8619–8624. doi: 10.1073/pnas.1403112111

frontiers | Frontiers in Neuroscience

TYPE Methods
PUBLISHED 18 January 2024
DOI 10.3389/fnins.2023.1268860

OPEN ACCESS

EDITED BY
Roberto Maffulli,
Italian Institute of Technology (IIT), Italy

REVIEWED BY
Fulvia Palesi,
University of Pavia, Italy
Noemi Montobbio,
University of Genoa, Italy

*CORRESPONDENCE
Dominique Sappey-Marinier
✉ dominique.sappey-marinier@univ-lyon1.fr

RECEIVED 28 July 2023
ACCEPTED 18 December 2023
PUBLISHED 18 January 2024

CITATION
Chen E, Barile B, Durand-Dubief F, Grenier T and Sappey-Marinier D (2024) Multiple sclerosis clinical forms classification with graph convolutional networks based on brain morphological connectivity.
Front. Neurosci. 17:1268860.
doi: 10.3389/fnins.2023.1268860

Multiple sclerosis clinical forms classification with graph convolutional networks based on brain morphological connectivity

Enyi Chen[1], Berardino Barile[1], Françoise Durand-Dubief[1,2], Thomas Grenier[1] and Dominique Sappey-Marinier[1,3]*

[1]CREATIS, CNRS UMR 5220, INSERM U1294, Université de Lyon, Université Claude Bernard-Lyon 1, INSA Lyon, Lyon, France, [2]Service de Sclérose en Plaques, des Pathologies de la Myéline et Neuro-Inflammation, Groupement Hospitalier Est, Hôpital Neurologique, Bron, France, [3]CERMEP - Imagerie du Vivant, Université de Lyon, Bron, France

Multiple Sclerosis (MS) is an autoimmune disease that combines chronic inflammatory and neurodegenerative processes underlying different clinical forms of evolution, such as relapsing-remitting, secondary progressive, or primary progressive MS. This identification is usually performed by clinical evaluation at the diagnosis or during the course of the disease for the secondary progressive phase. In parallel, magnetic resonance imaging (MRI) analysis is a mandatory diagnostic complement. Identifying the clinical form from MR images is therefore a helpful and challenging task. Here, we propose a new approach for the automatic classification of MS forms based on conventional MRI (i.e., T1-weighted images) that are commonly used in clinical context. For this purpose, we investigated the morphological connectome features using graph based convolutional neural network. Our results obtained from the longitudinal study of 91 MS patients highlight the performance (F1-score) of this approach that is better than state-of-the-art as 3D convolutional neural networks. These results open the way for clinical applications such as disability correlation only using T1-weighted images.

KEYWORDS

multiple sclerosis, graph convolutional network, CNN, classification, brain morphological connectivity, gray matter thickness

1 Introduction

Multiple sclerosis (MS) is a chronic autoimmune inflammatory and demyelinating disease of the central nervous system. While its etiology is still unknown (Polman et al., 2011), MS is the first cause of non-traumatic neurological disability in young adults, affecting about 2.8 million people worldwide (Goodin, 2014). Often starting with a preliminary clinical isolated syndrome (CIS) involving a large heterogeneity of clinical symptoms such as weak limbs, blurred vision, dizziness, fatigue, or tingling sensations, the disease may evolve along two main clinical courses. In 85% of patients, the disease starts as a relapsing-remitting course (RRMS, noted RR), with the occurrence of relapses. These RRMS patients can evolve over time into a non-systematic secondary-progressive course (SPMS, noted SP). In the 15% remaining patients, the disease evolves as primary-progressive MS (PPMS, noted PP) which corresponds to a steadily worsening of symptoms over time without any relapses (Lublin et al., 2014). The current McDonald diagnostic criteria for MS combine clinical assessment, imaging, and laboratory findings (Thompson et al., 2018). Despite such clinical classification, the status and the evolution of each patient could be very different from one to another, leading more and more to individual therapeutic approaches. Thus, to propose personalized

medical care and therapy, the neurologist needs to better predict the disease evolution based on early clinical, biological, and imaging markers available from disease onset.

Magnetic Resonance Imaging (MRI) is the most effective tool for the diagnosis of MS and for monitoring the disease modifying treatment. Conventional MRI provides T1-weighted (T1w), T2-weighted (T2w) and FLAIR images allowing the detection and follow-up of white matter (WM) lesions for clinical care (Mure et al., 2016). These conventional sequences allow the quantification of whole brain, WM or gray matter (GM) atrophy using dedicated software. More advanced MRI sequences such as diffusion-weighted imaging (DWI) and diffusion tensor imaging (DTI) have been developed to provide more sensitive markers of the inflammation processes occurring in WM and leading to T1- and T2-lesions. Several metrics of DTI such as the fractional anisotropy and the mean diffusion enable the detection of micro-architectural alterations in WM lesions as well as in normal-appearing WM (Jutten et al., 2019).

More recently, graph theory methods have been used to model brain network organization (Rubinov and Sporns, 2010; Guo et al., 2017). These graph models consist of nodes, based on the parcellation of brain GM regions, and edges, determined by the underlying links between the network nodes. In brain structural connectivity, these links are defined by the extraction of WM fibers using DTI tractography (Hagmann et al., 2007). Previously, Kocevar et al. (2016) have demonstrated an interest of such approaches for the classification of MS clinical profiles using Machine Learning (ML) methods, while Marzullo et al. (2019) improved the classification performance by a first approach using a Deep Learning (DL) model.

However, DTI data used for structural connectivity modeling require long acquisition time and complex processing techniques, which limits its applicability in clinical practice. Nevertheless, brain connectivity can also be obtained from conventional MRI by measuring morphological metrics of the GM on T1w images (Raamana and Strother, 2018). Indeed, several imaging investigations have shown that GM atrophy is present early in MS (Durand-Dubief et al., 2012; Eshaghi et al., 2018). Narayana et al. (2013) has found significant cortical thinning in RRMS patients compared to healthy subjects. Hence, the GM degeneration used in brain morphological connectivity models could provide a sensitive marker of the disease evolution. In such graphs, nodes represent GM areas obtained from the GM tissue parcellation, while edges represent a degree of (dis-)similarity between nodes features like GM thickness or curvature (MacDonald et al., 2000). Such approach has been recently used in Alzheimer's Disease (AD), showing that GM network measures predicted hippocampal atrophy rates in preclinical AD, in contrast to other AD biomarkers (Dicks et al., 2020). Also, Mahjoub et al. (2018) proposed to use morphological connectivity to discriminate late mild cognitive impairment from AD patients. Several studies of GM morphological network were used in Autism Spectrum Disorder (ASD) patients. Kong et al. (2019) proposed an auto-encoder-based deep neural network to identify ASD patients from typical controls, while Corps and Rekik (2019) used morphological networks to estimate the ASD patients' age and deduce the age-related cortical regions. In MS, Muthuraman et al. (2016) analyzed morphological GM thickness networks to classify CIS and RRMS patients using the Support Vector Machine model, obtaining a good level of accuracy. Meanwhile, several studies used graph metrics of GM networks to characterize MS patients. Hawkins et al. (2020) found reduced global efficiency and a more random network in RRMS subjects with cognitive impairment. Likewise, lower node degree and connectivity density were found by Rimkus et al. (2019) in MS patients with cognitive impairment. Rocca et al. (2021) combined functional connectivity and GM network to predict clinical worsening in MS, confirming that GM atrophy is an important predictor for the conversion from RRMS to SPMS. By using the source-based morphometry approach to decompose the cortical thickness map into different patterns, Steenwijk et al. (2016) have further shown that several anatomical patterns are strongly associated with clinical dysfunction in MS patients. Meanwhile, several studies also addressed the problem of age/gender and cortical thickness correlation, and removed their effects before further analysis. Eshaghi et al. (2016) fitted the linear regression between age and GM measurements and took only the residual part to classify MS cohort from neuromyelitis optical patients. Given the graph nature of brain connectivity, the use of graph neural network (GNN) to process such data is an evitable path. GNN allows us to deal with the heterogeneity of input data by capturing the message passing across nodes (Bronstein et al., 2021). More specifically, graph convolutional network (GCN), a reimplementation of convolution concept on GNN, is now ubiquitous in solving problems on non-euclidean data.

In the meantime, the application of convolutional neural network (CNN) has proven its strong ability in computer vision, especially in the biomedical image processing field. Leclerc et al. (2019) has successfully delineated cardiac structure on ultrasound images through an encoder-decoder-based model. 3D-CNN, a particular type of CNN, has been widely used in medical context since a huge amount of medical images were acquired and reconstructed in 3 dimensions. Various studies have focused on disease detection from anatomical neuroimaging (Wargnier-Dauchelle et al., 2023). Huang et al. (2019) have built a VGG-like CNN to adapt 3D image challenge for the purpose of Alzheimer's Disease (AD) classification using both T1w-MRI and FDG-PET modalities for a better outcome. Folego et al. (2020) have adapted LeNet, VGGNet, GoogLeNet, and ResNet in 3D domain to the aim of AD detection. Flaus et al. (2022) has proposed a 3D sequential ResNet to enhance PET images for better visualization of brain lesions. A transparent CNN framework proposed by Eitel et al. (2019) has revealed the decision process of CNN in the diagnosis of MS and pointed out more disease-relevant features in MR images. Optic nerve lesions, one of the first manifestations of MS, can be detected by the 3D-CNN model designed by Marti-Juan et al. (2022).

In this study, we proposed to use GCN for the classification of MS clinical forms based only on the measurement of GM morphological feature (thickness) obtained from T1w-MRI. The impacts of different methodological parameters such as the spatial resolution of the GM parcellation atlases and the level of different graph thresholds were compared. Finally, in order to demonstrate the interest of GCN for MS clinical forms classification, we compared the GCN with a classic 3D-CNN approach.

TABLE 1 MS cohort description of 660 scans including relapsing-remitting (RRMS), primary-progressive (PPMS), and secondary-progressive (SPMS) patients.

	RRMS	PPMS	SPMS
Number of patients (F/M)	42 (30/12)	21 (12/9)	28 (11/17)
Number of scans	299	143	218
Mean age at disease onset	28.5	35.0	27.6
Mean age at each scan (range)	35.4 (20.5–53.1)	43.0 (27.8–51.6)	42.9 (28.9–52.2)
Mean disease duration at first scan	4.9	5.6	13.4
Mean disease duration at each scan	7.3	7.5	15.1
EDSS median (range)	2 (0–5.5)	4 (2–7.5)	5.5 (3–8.5)

TABLE 2 Healthy controls cohort description of 335 T1-weighted MRI including 21 healthy controls (HC-AMSEP) acquired with the same protocol as MS cohort and 314 healthy controls (HC-IXI) obtained from the open-access IXI dataset.

HC	HC-AMSEP	HC-IXI
Number of subject (F/M)	21 (14/7)	314 (175/139)
Number of scans	21	314
Mean age at scan	42.9 (21.6–56.5)	50.8 (20.1–86.2)

2 Materials and methods

Our method was divided into three steps: (i) cortical feature extraction using FreeSurfer (Fischl, 2012); (ii) generation of brain morphological graphs using distance computation and threshold; and (iii) clinical forms classification using GCN.

2.1 MRI acquisition and data

The MS patient group (AMSEP) consists of 42 RR, 28 SP, and 21 PP participants included in a longitudinal MRI study. CIS patients (n = 12) were included in the RR patient group, in accordance with our clinical expert. Patients (n = 3) with change in clinical forms have been removed from the MS group. The patients underwent MR scans on a 1.5T Siemens Sonata system using an 8-channel head-coil at the Lyon CERMEP imaging platform, including a sagittal millimetric 3D-T1 MPRAGE (magnetization prepared rapid gradient echo-MPRAGE) sequence [(TR/TE/TI) = 1970/3.93/1100 ms, flip angle = 15°, field of view (FOV) = 256 × 256 mm, slice thickness = 1 mm, voxel size = 1 × 1 × 1 mm]. Table 1 provides information on the clinical data in further detail. During the first 3 years, MRI exams were performed every 6 months, and every year during the following years. These make up a MS patient dataset of 660 scans in total as detailed in Table 1. A healthy control (HC) group of 21 subjects following the AMSEP protocol was included in this study.

Another HC group of 314 scans from the IXI dataset (http://brain-development.org/ixi-dataset/) was introduced for the training process (noted IXI). These healthy subjects underwent MR scans on a 1.5T Philips Gyroscan Intera system using a T1w sequence (TR/TE = 9813/4603 ms, flip angle = 8°, 192 phase encoding steps, reconstruction diameter = 240 mm). These make up a HC dataset of 335 scans in total as detailed in Table 2.

2.2 Classification using graph-based convolutional network

As we explore the ability of cortical anatomical changes to identify MS forms, we extract features related to the shape of cortical regions. With such features, we then build a graph reflecting shape similarities between cortical regions and use the graph matrix to train the GCN. The full pipeline of the proposed network is shown in Figure 1.

2.2.1 Feature extraction

In order to obtain features of cortical regions, the brain GM was first segmented (Figure 1), the cortical surface was parcellated into N regions using a dedicated brain atlas. Morphological features of each region can thus be calculated and represented as a vector of values.

Automatic segmentation of GM and cortical surface reconstruction were performed on all T1w-MRI using FreeSurfer v6.0.0 image analysis suite (Fischl, 2012), a neuroimaging toolkit for human brain analysis. This includes 31 preprocessing steps such as motion correction, intensity normalization, skull stripping and non-linear registration. All FreeSurfer processing steps were done on the Virtual Imaging Platform (Glatard et al., 2013), the 1,001 images were processed simultaneously and it took 6 h per image on average. The input T1w-MRI brain was resampled onto an average brain (fsaverage) generated from 40 subjects using the Buckner dataset. The Buckner dataset is a subset of a large structural dataset created by the Buckner Lab, it was specifically selected for the intermediate processing step of FreeSurfer. The obtained cortical surface consists of a mesh with 163842 vertices. All outputs were smoothed at full-width/half-max (FWHM) value of 10 mm.

These smoothed outputs are then parcellated. In order to study the impact of the number of cortical regions N, three different atlases were used for brain parcellation and graph generation, namely the Desikan-Killiany (Desikan et al., 2006) with N = 68 regions, Destrieux (Destrieux et al., 2010) with N = 148 regions and Glasser (Glasser et al., 2016) with N = 360 regions. The cortex parcellation of the average template brain is demonstrated in Figure 2.

More specifically, a region number i (with $i = 1 \ldots N$) was assigned to each vertex according to the atlas chosen by registering the patient's brain mesh to the template brain. As mainly used in brain connectivity studies (reference), the cortical thickness was chosen as the morphological feature and calculated for each region.

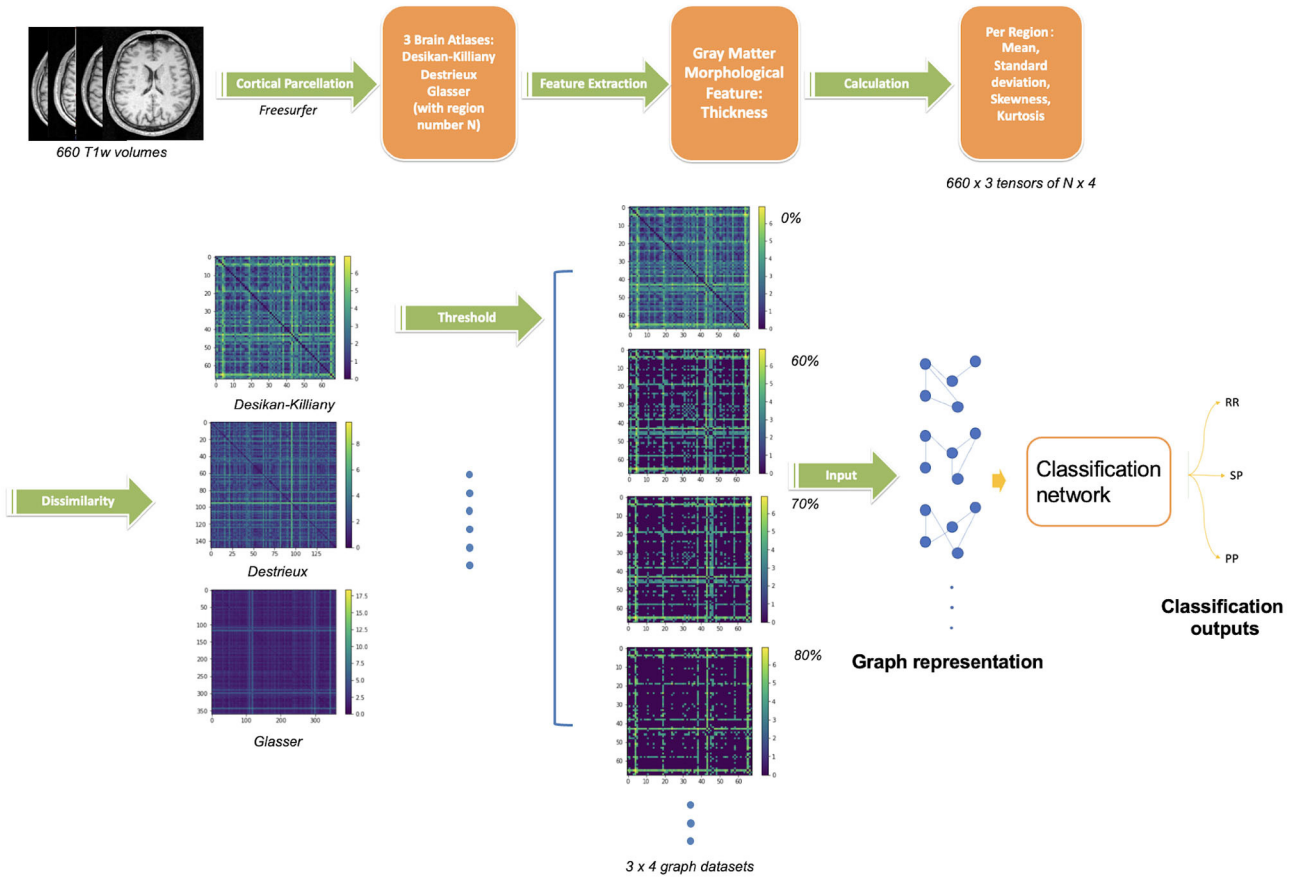


FIGURE 1
Proposed pipeline for GCN classification. The upper steps illustrate the cortical gray matter regions segmentation from T1w-MRI and parcellation using three atlases, the region feature extraction (thickness) and its vector values. The bottom steps describe the graph construction followed by the GCN classification network. Four threshold levels are applied on graphs (0, 60, 70, 80%), leading to four graphs per atlas. In summary, 12 networks are trained separately (3 atlases, 4 threshold levels) on 660 scans.

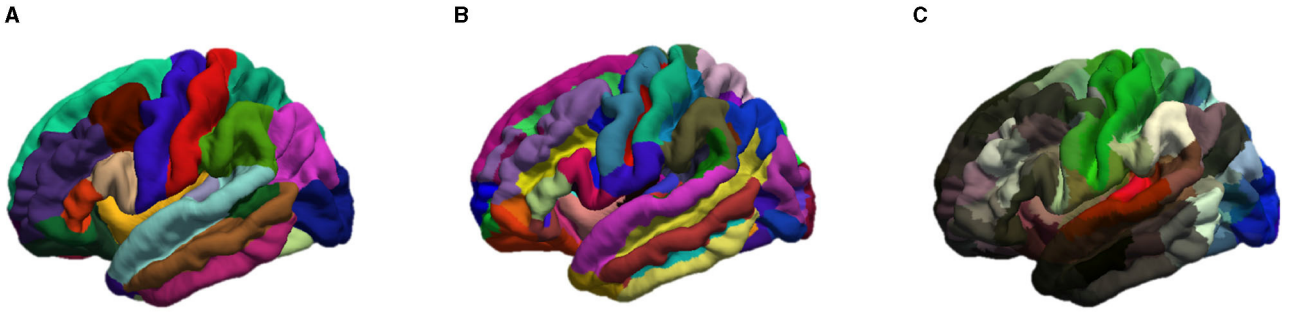


FIGURE 2
Representation of the cortical parcellation of the three atlases: **(A)** Desikan-Killiany; **(B)** Destrieux; **(C)** Glasser.

Since each region feature is a vector of thousands of elements on average, we summarize the distribution of the thickness values within one region i by a vector $\mathbf{x}_i \in \mathbb{R}^4$ containing the mean value, the standard deviation, the skewness, and the kurtosis: $\mathbf{x}_i = (\mu_i, \sigma_i, \gamma_i, k_i)$. We called the feature matrix $X \in \mathbb{R}^{N\times 4}$ the combination of the N vectors $\mathbf{x}_i$.

2.2.2 Age and gender normalization

Since women and men have different cortical atrophy manifestations with age (Narayana et al., 2013), we proposed two methods to normalize $\mathbf{x}_i$: a proportional normalization and a residual normalization. For the proportional normalization, we first calculated the average cortical thickness of the whole brain of all MS patients and healthy subjects from the IXI dataset. Then, we

performed a linear regression between age and cortical thickness as:

$$Cth = a * age + b$$

where Cth is the average cortical thickness of one person. Two different sets of coefficients (a_f, b_f) and (a_m, b_m) were calculated for healthy women and men respectively. If the slope represents the normal aging effect, we applied this slope to the MS patients group to correct the effect of age and sex. All MS patients' measurements were brought to the age of 20. Thus, the corrected thickness Cth_{20} of a patient can be expressed as:

$$Cth_{20} = a * 20 + b' = a * 20 + Cth - a * age$$

Therefore, the adjusted feature vector $\mathbf{x}_i'$ of each region with proportional correction with coefficient $\alpha = \frac{Cth_{20}}{Cth}$ can be represented as: $\mathbf{x}_i' = (\alpha\mu_i, \alpha\sigma_i, \gamma_i, k_i)$. The modified vectors were then used to calculate the new proportional normalized graphs following the same procedure as described above.

Inspired by the work of Eshaghi et al. (2016), we also proposed to adjust each cortical region for the effect of age and gender. For every brain region i of the healthy cohort, we fitted a linear regression where age was the regressor and the four attributes of the region were dependent variables. Therefore, for the four values of the feature vector, we have:

$$\mu_i = a_i^{(\mu)} * age + b_i^{(\mu)}$$

$$\sigma_i = a_i^{(\sigma)} * age + b_i^{(\sigma)}$$

$$\gamma^{(i)} = a_i^{(\gamma)} * age + b_i^{(\gamma)}$$

$$k_i = a_i^{(k)} * age + b_i^{(k)}$$

We then estimated the residual of each variable that was inexplicable by the healthy linear regression model: $r_i^{(\mu)} = \hat{\mu}_i - \mu_i = a_i^{(\mu)} * age + b_i^{(\mu)} - \mu_i$ for example in the case of average cortical thickness measure. The residual feature vector of one region became: $\mathbf{r}_i = (r_i^{(\mu)}, r_i^{(\sigma)}, r_i^{(\gamma)}, r_i^{(k)})$. The residual vectors were also used to calculate the residual graphs that were further used in the GCN classification. Notice that these regressions are performed for both males and females separately.

2.2.3 Graph generation

A graph G is a mathematical representation of a complex system and is defined by a collection of nodes V and edges E between pairs of nodes with the possibility to assign a weighted value w for each edge:

$$G = (V, E, w)$$

Therefore, a brain can be described as a graph, with each brain region being represented by a node $\mathbf{x}_i$, or $\mathbf{x}_i'$ and $\mathbf{r}_i$ in case of normalization. Here, we associate four attributes (mean value μ, standard deviation σ, skewness γ, and kurtosis k) to each node. The graph representation of brain morphological connectivity was defined as the dissimilarity across brain regions. We propose to compare two distances to calculate the region-wise connections. The first one is the Mahalanobis distance d_M:

$$d_M(\mathbf{x}_i, \mathbf{x}_j) = \left((\mathbf{x}_i - \mathbf{x}_j)^T S^{-1} (\mathbf{x}_i - \mathbf{x}_j)\right)^{1/2}$$

with S the covariance matrix of samples $\mathbf{x}_i$ and $\mathbf{x}_j$.

The second studied distance is the Taxicab (or Manhattan) distance d_T:

$$d_T(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^{4} |x_i^k - x_j^k|$$

where x_i^k is the kth dimension of the vector $\mathbf{x}_i$.

The adjacent matrix $A \in \mathbb{R}^{N \times N}$ is computed for all distances between $\mathbf{x}_i$ and $\mathbf{x}_j$: $A(i,j)_X = d(\mathbf{x}_i, \mathbf{x}_j)$.

Using both X and A, we generate weighted and undirected graphs. The edge weights are given by the adjacent matrix.

Thresholds were used to counteract the impact of the redundant information given by the brain adjacent matrix. A fixed rejection quantile τ is used as a threshold value to remove the lowest distances and thus maintains the same graph density for each subject.

For graph availability, the reader can refer to Section 5.

2.2.4 GCN classification

Graph convolutional networks were used as they exploit input data through graph structure. As a dimension reduction tool, graph representation can largely reduce input data size from 12 MB to 130 KB on average in our case. Intuitively speaking, brain network topology is an alternative method of image analysis. Sporns (2018) have confirmed the importance of graph theory for the understanding of brain structure. Based on our previous results using brain structural graph analysis (Marzullo et al., 2019), we explore a new approach using brain morphological graph.

For the graph $G = (V, E, w)$, the algorithm takes the adjacent matrix A and the associated node features matrix X as input. The layer-wise propagation rule is defined as follows (Kipf and Welling, 2017):

$$H^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H^{(l)} W^{(l)})$$

Where $\tilde{A}$ is the sum of A with the identity matrix I, $\tilde{D}$ is the corresponding diagonal degree matrix and the adjacent matrix is normalized by the step $\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}}$. W^l represents the trainable weight over each layer. The RELU activation function $\sigma(x) = max(0, x)$ is chosen for σ.

2.2.5 GCN architecture

The proposed GCN classification model was composed of 3 GCN layers followed by a global mean pool layer with a dropout rate of 0.3 to prevent overfitting. The proposed structure is shown in Figure 3. This led to 8835 trainable parameters.

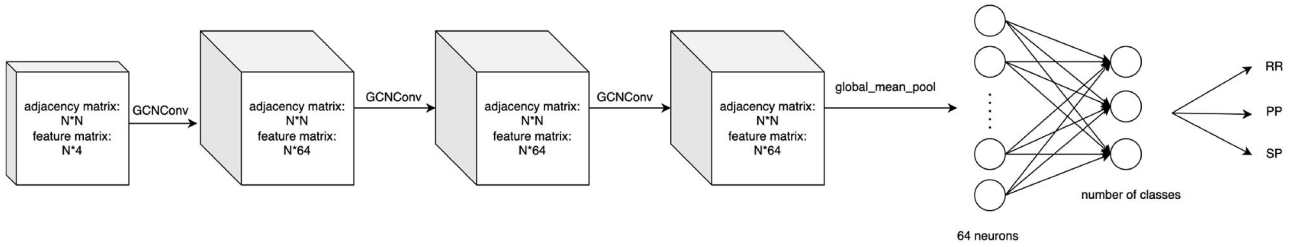


FIGURE 3
The overall structure of the proposed graph-based convolutional network. *N* is the number of regions according to the atlas chosen. Four represents the four elements of the feature vector per region. Input of the network consists of one adjacency matrix (N*N) and one feature matrix (N*4) per patient. The network starts with three graph convolutional layers of 64 filters each, then gathered into a vector using a global mean pooling. Two fully connected layers are used to obtain the classification into three classes (RR, PP, SP).

2.3 Classification using 3D convolutional neural network

To validate our GCN against classically used CNN architectures, we implemented a 3D-CNN architecture using a similar architecture by replacing graph convolutional layers with classical convolutional layers. The output of a filter of a 3D convolutional layer with kernel W of size ($f_h x f_w x f_d x f_c$) can be expressed as follows:

$$z_{i,j,k} = b + \sum_{p=0}^{f_h-1} \sum_{q=0}^{f_w-1} \sum_{r=0}^{f_d-1} \sum_{c=0}^{f_c-1} x_{i',j',k',c} . \mathbf{W}_{p,q,r,c}$$

with

$$i' = i + p - \lfloor f_h/2 \rfloor \; and \; j' = j + q - \lfloor f_w/2 \rfloor \; and \; k' = k + r - \lfloor f_d/2 \rfloor$$

Therefore, a 3D-CNN model was constituted of three 3D convolutional layer sets, including a 3D convolutional layer (kernel of 3 × 3 × 3), followed by a max pooling layer (subsampling spatial support by 2 × 2 × 2) and then a batch normalization layer. The tensor is then flattened and used as input of two consecutive fully connected layers of 128 and 2 neurons, respectively. These made up of 22,548,122 trainable parameters of the CNN network.

Before using a deep neural network to classify the 3D MRI, all scans were pre-processed using the brain extraction tool (BET) of FMRIB Software Library in order to eliminate non-brain structures. Then, the 3D-CNN image classification network predicts the class (RR, SP, or PP) of the T1w image of a patient's brain used as input. The architecture used is summarized in Figure 4. To prevent over-fitting, a dropout (Srivastava et al., 2014) rate of 0.3 is applied after the flattening layer.

As it is known that CNN classification needs numerous data to perform well, we compared its performance with the classification results using a graph-based neural network.

2.4 Experimental settings

According to our previous study using brain morphological connectivity (Barile et al., 2022), 4 threshold levels $\tau \in \{0, 0.6, 0.7, 0.8\}$ were applied to the adjacent matrix computed using the 3 atlases and the 2 distances. Thus, each GCN classification is carried out in 72 different ways, and one for CNN.

For both network architectures, the MS images were divided into two datasets: approximately 80% of scans used for training and 20% of the scans used only for testing, i.e., to evaluate the performance of networks. To avoid the impacts of repetition of the same patient, we carefully grouped all time points of one patient in the same train or test set using the stratified group k-fold technique. The exams of the same patient won't be in the train set and test set simultaneously.

The precision, recall, and the F1-score were used to assess both algorithms' effectiveness. To provide a more thorough assessment of the two models, cross-validation using five-folds was performed.

From hyperparameters manual optimization, we use the Adam optimizer with a learning rate of 0.001 for GCN and the Stochastic Gradient Descent optimizer with a learning rate of 0.001 for 3D-CNN.

GCN was trained on one GPU (NVIDIA GeForce RTX 3060), and CNN was trained on one NVIDIA RTX A5000. All experiments were done using PyTorch.

For code availability, the reader can refer to Section 5.

3 Results

In this section, we first present the GCN classification tasks and then the results without age and gender normalization to allow the comparison with 3D-CNN classification results. Second, the GCN classification results with age and sex normalization are presented.

3.1 Clinical forms classification tasks

Six classification tasks related to clinical needs were implemented: (1) RR vs. PP; (2) RR vs. SP; (3) PP vs. SP; (4) RR vs. PP+SP; (5) RR vs. PP vs. SP; (6) MS vs. HC. For this last task, the train set consists of 619 MS scans and 290 randomly selected scans from the IXI dataset. For the test set, 42 scans were selected from the MS group (24 RRMS, 10 PPMS, 8 SPMS) along with the 21 HC-AMSEP scans from the same study and 24 HC-IXI scans from the IXI dataset. For the other tasks, only the MS patients

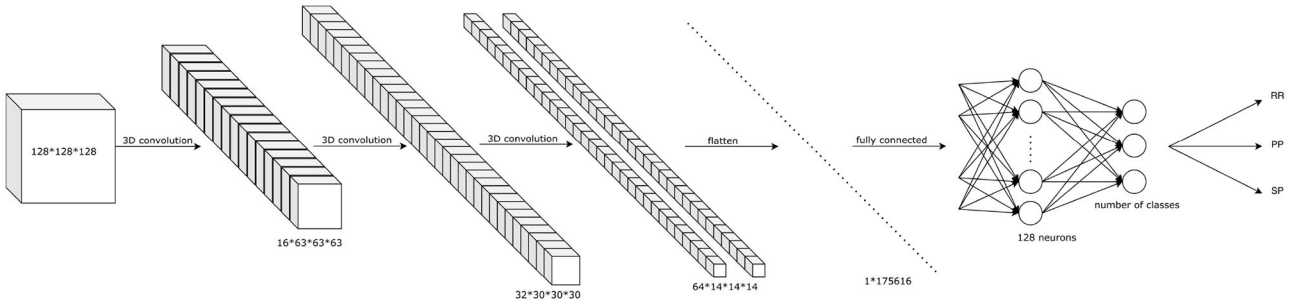


FIGURE 4
The overall structure of the proposed 3D-CNN network. It starts with three convolutional layers of 16, 32, and 64 filters respectively, each convolution layer followed by a max pooling layer. The tensor is then flattened and two fully connected layers are used to obtain the classification into three classes (RR, PP, SP).

dataset was used. A five-fold stratified cross-validation scheme was applied for all tasks.

3.2 GCN classification

3.2.1 Without normalization

F1-score of the three atlases (Desikan-Killiany, Destrieux, Glasser), four rejection rates and two distance calculation approaches were compared as shown in Tables 3, 4. Precision and Recall measures of corresponding experiments were included in Supplementary material.

Comparing classification results task by task, the best result was always found using Mahalanobis instead of Taxicab distance for the dissimilarity measurement. The classification of RR vs. PP gave the best result when an 80% rejection rate was applied to the Destrieux atlas with an F1-score of 72.5%. The separation between RR and SP patients provides an F1-score of 72.2% using an 80% rejection rate on the Glasser atlas. By grouping the PP and SP in a neurodegenerative group, the binary classification of RR vs. PP+SP reached an F1-score of 68.9%. The best three classes classification was obtained using an 80% rejection rate on the Glasser atlas with an F1-score of 64.2%. The optimal PP/SP splitting leading to an F1-score of 53.1% was obtained using the Glasser atlas and a rejection rate of 70%. Finally, all GCN classification networks can achieve a great result on MS vs. HC task (100% F1-score on the predefined unseen test dataset). Atlas-wise speaking, for Mahalanobis distance measurement, a 60% rejection rate gave the best result on the Desikan-Killiany atlas, while an 80% rejection rate yielded the best outcome on both Destrieux and Glasser atlases. For Taxicab distance measurement, a 70% rejection rate gave the best result on the Desikan-Killiany atlas, the graph without rejection generated the best on the Destrieux atlas, and a 60% rejection rate achieved the best performance on the Glasser atlas.

3.2.2 With normalization

In order to correct for age and gender, two normalization methods have been carried out. The results obtained using three atlases and two distance methods are shown in Tables 5–8. The best RR/PP separation can be found when the residual normalization was carried out to the Desikan-Killiany atlas with a threshold of 80%. The proportional normalization method applied to the Glasser atlas with an 80% rejection rate generated the best results of RR vs. SP, RR vs. PP+SP, and RR vs. PP vs. SP with F1-scores 71.1, 67.8, and 62.1% respectively. The best result of PP/SP classification can be found in residual normalization on the Desikan-Killiany atlas (rejection rate = 0) with an F1-score of 64.2%. For the proportional normalization method, the best overall result can be found using the Glasser atlas with 80% threshold. The best overall result for the residual normalization method was carried out by the same atlas with 60% threshold.

3.3 Comparing CNN and GCN

The results of the comparison between 3D-CNN classification and GCN without normalization are shown in Table 9. Comparing RR individually with PP and SP, 3D-CNN returned an F1-score of 72.1% and 69.7% respectively, which are slightly lower than GCN results. The separation between the RR and PP+SP groups on the F1-score was greater than that of the GCN technique at 70.7%. The 3D-CNN method generated a similar result on the multi-class classification task with an F1-score of 63.9%. Finally, 3D-CNN achieved a lower result than GCN for the PP vs. SP partition with a 49.5% F1-score. Overall, the best results were obtained using GCN over 3D-CNN while implementing an 80% rejection rate on the Glasser atlas and the Mahalanobis distance.

4 Discussion

Graph Convolutional Network is an innovative approach for the classification of clinical forms in multiple sclerosis. While functional and structural connectivities were previously used and provided good results (Ktena et al., 2018; Marzullo et al., 2019), they were constrained by the small size of the database available in clinical routine. To overcome this limitation, one approach is to develop a morphological connectivity method requiring only anatomical T1w MRI for brain studies. In order to test such a hypothesis, we developed a complete pipeline using morphological connectivity and graph convolutional networks. To

TABLE 3 F1-scores (mean value ± standard deviation) of clinical forms classification using GCN based on Mahalanobis graph for three parcellation atlases and four threshold levels τ.

Atlas	Tasks	$\tau = 0$	$\tau = 0.6$	$\tau = 0.7$	$\tau = 0.8$
Desikan-Killiany	RR vs. PP	0.701 ± 0.076	0.698 ± 0.068	**0.706 ± 0.056**	0.703 ± 0.052
	RR vs. SP	0.684 ± 0.064	**0.7 ± 0.077**	0.684 ± 0.061	0.674 ± 0.08
	RR vs. PP + SP	**0.654 ± 0.088**	0.648 ± 0.081	0.647 ± 0.081	0.638 ± 0.071
	RR vs. PP vs. SP	0.594 ± 0.047	0.593 ± 0.059	**0.603 ± 0.037**	0.567 ± 0.043
	PP vs. SP	0.438 ± 0.092	**0.475 ± 0.073**	0.466 ± 0.064	0.465 ± 0.101
	MS vs. HC	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
Destrieux	RR vs. PP	0.72 ± 0.103	0.721 ± 0.089	0.721 ± 0.088	**0.725 ± 0.085**
	RR vs. SP	0.684 ± 0.065	0.679 ± 0.066	0.666 ± 0.055	**0.686 ± 0.07**
	RR vs. PP + SP	0.649 ± 0.074	**0.657 ± 0.061**	0.656 ± 0.058	0.642 ± 0.071
	RR vs. PP vs. SP	0.569 ± 0.037	0.588 ± 0.059	0.587 ± 0.057	**0.596 ± 0.066**
	PP vs. SP	**0.485 ± 0.05**	0.45 ± 0.054	0.479 ± 0.058	0.466 ± 0.073
	MS vs. HC	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
Glasser	RR vs. PP	0.702 ± 0.096	**0.722 ± 0.102**	0.711 ± 0.099	0.714 ± 0.079
	RR vs. SP	0.711 ± 0.062	0.71 ± 0.059	0.694 ± 0.071	**0.722 ± 0.067**
	RR vs. PP + SP	0.627 ± 0.085	0.681 ± 0.085	0.687 ± 0.084	**0.689 ± 0.095**
	RR vs. PP vs. SP	0.609 ± 0.038	0.634 ± 0.055	0.62 ± 0.066	**0.642 ± 0.063**
	PP vs. SP	0.495 ± 0.076	0.479 ± 0.076	**0.531 ± 0.115**	0.471 ± 0.077
	MS vs. HC	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000

The best rejection rates within each atlas are in bold, the best overall results are in gray background.

TABLE 4 F1-scores (mean value ± standard deviation) of clinical forms classification using GCN based on Taxicab graph for three parcellation atlases and four threshold levels τ.

Atlas	Tasks	$\tau = 0$	$\tau = 0.6$	$\tau = 0.7$	$\tau = 0.8$
Desikan-Killiany	RR vs. PP	0.701 ± 0.075	**0.709 ± 0.065**	0.706 ± 0.056	0.693 ± 0.097
	RR vs. SP	0.682 ± 0.061	0.671 ± 0.063	**0.684 ± 0.061**	0.671 ± 0.052
	RR vs. PP + SP	0.654 ± 0.087	0.662 ± 0.08	**0.667 ± 0.073**	0.646 ± 0.078
	RR vs. PP vs. SP	0.596 ± 0.047	0.601 ± 0.04	**0.603 ± 0.037**	0.571 ± 0.033
	PP vs. SP	0.437 ± 0.092	0.458 ± 0.07	0.466 ± 0.064	**0.471 ± 0.07**
	MS vs. HC	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
Destrieux	RR vs. PP	0.721 ± 0.103	0.719 ± 0.097	**0.721 ± 0.088**	0.709 ± 0.075
	RR vs. SP	**0.683 ± 0.064**	0.674 ± 0.054	0.666 ± 0.055	0.649 ± 0.064
	RR vs. PP + SP	**0.65 ± 0.075**	0.647 ± 0.074	0.649 ± 0.066	0.648 ± 0.064
	RR vs. PP vs. SP	0.569 ± 0.037	**0.587 ± 0.055**	0.587 ± 0.057	0.58 ± 0.057
	PP vs. SP	0.481 ± 0.05	0.476 ± 0.057	0.479 ± 0.058	**0.493 ± 0.043**
	MS vs. HC	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
Glasser	RR vs. PP	0.701 ± 0.095	**0.722 ± 0.096**	0.711 ± 0.099	0.696 ± 0.099
	RR vs. SP	**0.711 ± 0.063**	0.708 ± 0.069	0.694 ± 0.071	0.672 ± 0.035
	RR vs. PP + SP	0.628 ± 0.086	**0.656 ± 0.09**	0.653 ± 0.096	0.63 ± 0.09
	RR vs. PP vs. SP	0.609 ± 0.039	**0.629 ± 0.068**	0.62 ± 0.066	0.593 ± 0.065
	PP vs. SP	0.494 ± 0.073	0.513 ± 0.089	**0.531 ± 0.115**	0.526 ± 0.09
	MS vs. HC	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000

The best rejection rates within each atlas are in bold, the best overall results are in gray background.

TABLE 5 F1-scores (mean value ± standard deviation) of clinical forms classification using GCN based on Mahalanobis age-gender proportional adjusted graph for three parcellation atlases and four threshold levels τ.

Atlas	Tasks	$\tau = 0$	$\tau = 0.6$	$\tau = 0.7$	$\tau = 0.8$
Desikan-Killiany	RR vs. PP	0.582 ± 0.091	0.581 ± 0.111	**0.616 ± 0.091**	0.611 ± 0.096
	RR vs. SP	**0.613 ± 0.08**	0.609 ± 0.07	0.6 ± 0.066	0.591 ± 0.065
	RR vs. PP + SP	0.615 ± 0.058	0.622 ± 0.044	**0.625 ± 0.047**	0.592 ± 0.048
	RR vs. PP vs. SP	**0.545 ± 0.049**	0.551 ± 0.069	0.535 ± 0.068	0.529 ± 0.049
	PP vs. SP	0.428 ± 0.044	**0.491 ± 0.056**	0.45 ± 0.043	0.463 ± 0.083
	MS vs. HC	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
Destrieux	RR vs. PP	0.629 ± 0.118	**0.635 ± 0.115**	0.625 ± 0.115	0.605 ± 0.127
	RR vs. SP	0.63 ± 0.076	0.634 ± 0.076	0.632 ± 0.102	**0.647 ± 0.105**
	RR vs. PP + SP	**0.608 ± 0.068**	0.601 ± 0.05	0.602 ± 0.069	0.589 ± 0.054
	RR vs. PP vs. SP	0.546 ± 0.043	0.548 ± 0.056	0.558 ± 0.061	**0.58 ± 0.073**
	PP vs. SP	0.476 ± 0.044	0.471 ± 0.055	**0.494 ± 0.058**	0.49 ± 0.066
	MS vs. HC	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
Glasser	RR vs. PP	0.635 ± 0.146	0.668 ± 0.124	0.669 ± 0.122	**0.671 ± 0.117**
	RR vs. SP	0.638 ± 0.092	0.679 ± 0.117	0.692 ± 0.114	**0.711 ± 0.107**
	RR vs. PP + SP	0.619 ± 0.063	0.643 ± 0.071	0.657 ± 0.075	**0.678 ± 0.063**
	RR vs. PP vs. SP	0.578 ± 0.077	0.582 ± 0.065	0.6 ± 0.044	**0.621 ± 0.032**
	PP vs. SP	**0.592 ± 0.086**	0.569 ± 0.097	0.525 ± 0.09	0.533 ± 0.116
	MS vs. HC	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000

The best rejection rates within each atlas are in bold, the best overall results are in gray background.

TABLE 6 F1-scores (mean value ± standard deviation) of clinical forms classification using GCN based on Taxicab age-gender proportional adjusted graph for three parcellation atlases and four threshold levels τ.

Atlas	Tasks	$\tau = 0$	$\tau = 0.6$	$\tau = 0.7$	$\tau = 0.8$
Desikan-Killiany	RR vs. PP	0.588 ± 0.089	0.581 ± 0.111	**0.615 ± 0.092**	0.611 ± 0.096
	RR vs. SP	0.607 ± 0.08	**0.609 ± 0.071**	0.6 ± 0.066	0.591 ± 0.063
	RR vs. PP + SP	0.615 ± 0.06	0.622 ± 0.045	**0.626 ± 0.047**	0.592 ± 0.047
	RR vs. PP vs. SP	**0.542 ± 0.049**	0.55 ± 0.069	0.535 ± 0.068	0.529 ± 0.046
	PP vs. SP	0.427 ± 0.044	**0.49 ± 0.053**	0.451 ± 0.042	0.462 ± 0.083
	MS vs. HC	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
Destrieux	RR vs. PP	0.632 ± 0.119	**0.636 ± 0.116**	0.631 ± 0.111	0.605 ± 0.128
	RR vs. SP	0.637 ± 0.075	0.633 ± 0.09	0.631 ± 0.101	**0.647 ± 0.105**
	RR vs. PP + SP	**0.609 ± 0.067**	0.601 ± 0.051	0.601 ± 0.07	0.588 ± 0.054
	RR vs. PP vs. SP	0.546 ± 0.042	0.549 ± 0.056	0.558 ± 0.061	**0.58 ± 0.074**
	PP vs. SP	0.48 ± 0.045	0.473 ± 0.057	**0.493 ± 0.057**	0.489 ± 0.067
	MS vs. HC	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
Glasser	RR vs. PP	0.618 ± 0.12	0.645 ± 0.098	0.63 ± 0.117	**0.655 ± 0.085**
	RR vs. SP	0.627 ± 0.092	0.669 ± 0.11	0.686 ± 0.106	**0.7 ± 0.096**
	RR vs. PP + SP	0.606 ± 0.055	0.0.632 ± 0.069	0.649 ± 0.069	**0.67 ± 0.059**
	RR vs. PP vs. SP	0.567 ± 0.068	0.572 ± 0.057	0.594 ± 0.039	**0.611 ± 0.029**
	PP vs. SP	**0.6 ± 0.094**	0.576 ± 0.097	0.538 ± 0.096	0.51 ± 0.101
	MS vs. HC	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000

The best rejection rates within each atlas are in bold, the best overall results are in gray background.

TABLE 7 F1-scores (mean value ± standard deviation) of clinical forms classification using GCN based on Mahalanobis age-gender residual adjusted graph for three parcellation atlases and four threshold levels τ.

Atlas	Tasks	$\tau = 0$	$\tau = 0.6$	$\tau = 0.7$	$\tau = 0.8$
Desikan-Killiany	RR vs. PP	0.7 ± 0.097	0.681 ± 0.097	0.679 ± 0.085	**0.715 ± 0.069**
	RR vs. SP	0.578 ± 0.105	0.577 ± 0.109	0.579 ± 0.114	**0.581 ± 0.126**
	RR vs. PP + SP	0.612 ± 0.055	**0.618 ± 0.064**	0.603 ± 0.069	0.61 ± 0.068
	RR vs. PP vs. SP	**0.525 ± 0.065**	0.484 ± 0.042	0.488 ± 0.066	0.503 ± 0.055
	PP vs. SP	**0.635 ± 0.079**	0.601 ± 0.09	0.595 ± 0.098	0.563 ± 0.118
	MS vs. HC	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
Destrieux	RR vs. PP	**0.709 ± 0.102**	0.693 ± 0.105	0.697 ± 0.107	0.696 ± 0.11
	RR vs. SP	0.58 ± 0.103	0.579 ± 0.11	0.599 ± 0.115	**0.603 ± 0.124**
	RR vs. PP + SP	**0.558 ± 0.035**	0.557 ± 0.015	0.547 ± 0.008	0.538 ± 0.025
	RR vs. PP vs. SP	0.483 ± 0.074	0.476 ± 0.092	0.481 ± 0.099	**0.49 ± 0.101**
	PP vs. SP	0.481 ± 0.105	0.498 ± 0.094	0.505 ± 0.083	**0.528 ± 0.077**
	MS vs. HC	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
Glasser	RR vs. PP	**0.711 ± 0.087**	0.707 ± 0.098	0.705 ± 0.096	0.644 ± 0.153
	RR vs. SP	0.595 ± 0.132	0.612 ± 0.131	0.619 ± 0.138	**0.637 ± 0.127**
	RR vs. PP + SP	0.588 ± 0.08	**0.617 ± 0.083**	0.607 ± 0.088	0.608 ± 0.094
	RR vs. PP vs. SP	0.51 ± 0.068	**0.54 ± 0.082**	0.537 ± 0.083	0.527 ± 0.066
	PP vs. SP	**0.566 ± 0.149**	0.509 ± 0.096	0.523 ± 0.093	0.561 ± 0.097
	MS vs. HC	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000

The best rejection rates within each atlas are in bold, the best overall results are in gray background.

TABLE 8 F1-scores (mean value ± standard deviation) of clinical forms classification using GCN based on Taxicab age-gender residual adjusted graph for three parcellation atlases and four threshold levels τ.

Atlas	Tasks	$\tau = 0$	$\tau = 0.6$	$\tau = 0.7$	$\tau = 0.8$
Desikan-Killiany	RR vs. PP	0.7 ± 0.097	0.681 ± 0.097	0.678 ± 0.085	**0.715 ± 0.072**
	RR vs. SP	0.579 ± 0.111	0.58 ± 0.106	**0.583 ± 0.113**	0.575 ± 0.12
	RR vs. PP + SP	0.611 ± 0.055	**0.617 ± 0.062**	0.607 ± 0.067	0.609 ± 0.067
	RR vs. PP vs. SP	**0.525 ± 0.065**	0.485 ± 0.042	0.482 ± 0.062	0.503 ± 0.056
	PP vs. SP	**0.642 ± 0.079**	0.604 ± 0.088	0.593 ± 0.101	0.567 ± 0.124
	MS vs. HC	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
Destrieux	RR vs. PP	**0.711 ± 0.101**	0.693 ± 0.105	0.694 ± 0.112	0.696 ± 0.112
	RR vs. SP	0.582 ± 0.105	0.579 ± 0.112	0.597 ± 0.122	**0.598 ± 0.123**
	RR vs. PP + SP	0.553 ± 0.036	**0.56 ± 0.015**	0.533 ± 0.025	0.531 ± 0.032
	RR vs. PP vs. SP	**0.491 ± 0.073**	0.476 ± 0.092	0.479 ± 0.098	0.48 ± 0.1
	PP vs. SP	0.48 ± 0.106	0.497 ± 0.091	0.526 ± 0.076	**0.527 ± 0.074**
	MS vs. HC	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
Glasser	RR vs. PP	**0.713 ± 0.088**	0.707 ± 0.098	0.705 ± 0.096	0.645 ± 0.155
	RR vs. SP	0.589 ± 0.126	0.611 ± 0.131	0.618 ± 0.135	**0.637 ± 0.128**
	RR vs. PP + SP	0.592 ± 0.086	**0.618 ± 0.084**	0.607 ± 0.088	0.608 ± 0.09
	RR vs. PP vs. SP	0.508 ± 0.067	**0.542 ± 0.083**	0.537 ± 0.081	0.523 ± 0.062
	PP vs. SP	**0.567 ± 0.126**	0.509 ± 0.095	0.529 ± 0.095	0.55 ± 0.088
	MS vs. HC	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000

The best rejection rates within each atlas are in bold, the best overall results are in gray background.

TABLE 9 Best F1-scores (mean value ± standard deviation) of clinical forms classification using 3D-CNN and GCN [three datasets: non-normalized (NN) graph, proportional normalized (PN) graph, and residual normalized (RN) graph].

Tasks	3D-CNN	NN GCN	PN GCN	RN GCN
RR vs. PP	0.697 ± 0.124	**0.725 ± 0.085**	0.671 ± 0.117	0.715 ± 0.069
RR vs. SP	0.721 ± 0.081	**0.722 ± 0.067**	0.711 ± 0.107	0.637 ± 0.128
RR vs. PP + SP	**0.707 ± 0.066**	0.689 ± 0.095	0.678 ± 0.063	0.618 ± 0.084
RR vs. PP vs. SP	0.639 ± 0.036	**0.642 ± 0.063**	0.621 ± 0.032	0.542 ± 0.083
PP vs. SP	0.495 ± 0.06	0.531 ± 0.115	0.6 ± 0.094	**0.642 ± 0.079**

The best F1-scores for each classification task are in bold.

our knowledge, this is the first attempt to use this approach for the classification of MS clinical forms. Brain graphs were established based on Desikan-Killiany, Destrieux, and Glasser atlases, for GM parcellation. Rejection rates of 60, 70, and 80% were applied to connectivity graphs to preserve solely main differences across brain regions. Morphological connectivity data were fed into GCN while 3D brain images were loaded in 3D-CNN to compare the two classification approaches.

First, non-normalized GCN was compared to 3D-CNN, which was unable to normalize age or gender based on image data. Generally speaking, GCN has outperformed 3D-CNN on 4 out of 5 predefined tasks when the threshold/atlas pair was carefully chosen. For the task RR vs. PP+SP, the F1-score generated by GCN was slightly weaker than the result of 3D-CNN with a 1.8 percentage point. However, it requires more computation resources to train a simple 3 convolutional layers network. In our case, GCN only took 5 h for network training while achieving a better result than 3D-CNN which took more than a week on the same computer. The proposed pipeline has gained in computation time thanks to its dimension-reduction ability. Instead of working on 256 × 256 × 256 volumetric images, the graph approach allowed us to use the adjacent matrix of size 360 × 360 in the most complex case.

The comparison of the two classification networks has also given us insights into the medical image processing field. In general, clinical image classification tasks can be easily affected by acquisition changes (manufacturers, centers, MR field, etc.). In particular, CNNs are sensitive to intensity changes with the use of convolution layers. To address this problem, CNN classification networks must be trained on a large number of images that represent both the variability of the acquisition process and the diversity of the patients. Since most medical datasets are composed of a small number of patients, CNN doesn't usually generate well due to its data-thirsty characteristic. In contrast, GCN can be trained on brain graph features that are less sensitive to image intensity changes. Indeed, cortical thinning is an important biomarker of the MS neurodegenerative process that is visible in T1w images (Narayana et al., 2013). With a brain graph generated from cortical thickness, these small changes in the brain were well-captured by the proposed GCN pipeline. Our pipeline returns a clearer relation between brain atrophy and clinical forms, compared to the 3D-CNN approach, which could be improved by using Grad-CAM (Selvaraju et al., 2020) or similar methods.

Second, normalized GCN was used to classify MS clinical forms. This is essential for clinical forms classification. Binary and multi-class classifications were performed between the three clinical forms (RR, PP, SP). The result of normalized GCN showed that GCN can return satisfactory results on binary classification between MS clinical courses. More specifically, the automatic separation of inflammatory forms from neurodegenerative forms, RR vs. SP and PP groups, has been carried out. The best F1-score was found when separating RR from PP patients, and a good result was also obtained in the RR/SP classification task. On one hand, RR patients present relapses corresponding to focal inflammatory processes. On the other hand, SP and PP patients share the experience of progressive clinical evolution, associated or not with inflammatory activity, resulting from degenerative phenomena of the gray matter. Thus, by grouping SP and PP patients, an adequate result was found when the finest atlas (Glasser) was applied.

The three-class classification is a difficult multi-class categorization task which is further worsened by the imbalanced data distribution. Nevertheless, a promising result was obtained using the Glasser parcellation atlas with a high rejection rate, indicating the advantage of dimension reduction when facing complex brain data such as our case.

Classification of SP and PP was the hardest binary classification task to be accomplished. this is partially due to the small amount of PP cases. Indeed, SP and PP are two neurodegenerative forms sharing similar pathological processes. Moreover, PP is a starting clinical form that can be divided into subclasses depending on the level of disability. With an EDSS score ranging from 2 to 7.5, our PP population is composed of both early and late stages of the disease. The latter ones are more relevant and probably more similar to SP patients as shown in the disease duration at scan. This large variability of disability scores reflects different progressions of the disease and thus different stages of brain alterations. Thus, the SP and some PP patients may share MRI phenotypes which makes the classification difficult, and perhaps even unnecessary.

Achieving good results, the binary classification of HC vs. MS patients was not our primary goal. In general, MS patients can be easily distinguished from healthy subjects in both clinical and imaging ways. In our experience, an F1-score of 100% was observed in all GCN outputs, meaning that all combinations of atlases and thresholds provided enough information for the classification task. Similar results were obtained in the previous work of Marzullo et al. (2019) on brain structural connectivity. Marzullo et al. (2019) has performed the test of HC vs. CIS+RR (24/253) and the test of HC vs. SP+PP (24/325) and achieved the best result (F-measure = 1), demonstrating an evident difference between HC and MS brain morphological and structural networks, respectively.

To further compare our work with other studies, we analyzed the results obtained from Marzullo et al. (2019) and Barile et al. (2022). Apart from the binary classification of HC vs. MS patients, Marzullo et al. (2019) have also tested the separation between early and progressive forms of MS (CIS+RR vs. SP+PP: 253/325) obtaining the highest F-measure at 0.99. Since CIS subjects are included in the RR group in our study, we can compare the previous result with our classification task of RR vs. SP+PP (299/361), leading to an F1-score of 0.678. This strong difference in performance demonstrates that white matter inflammation introduced significant information that facilitates the classification

of clinical forms in MS. In contrast, the work of Barile et al. (2022) was performed on GM morphological connectivity. Three similar tasks were reported: (1) CIS+RR vs. PP; (2) CIS+RR vs. SP; (3) CIS+RR vs. SP+PP. By employing the same pipeline of graph generation and atlas (Glasser) and an ensemble of machine learning methods, they have obtained an F1-score of 0.661 (0.12), 0.654 (0.12), 0.648 (0.11) for the three tasks, respectively. In our study, we obtained better F1 scores of 0.671 (0.117), 0.711 (0.107), 0.678 (0.063) for the same tasks. This gain in performance (higher F1-score and reduced standard deviation) demonstrated the interest of brain graph convolutional networks.

Taxicab distance is an L1-norm metric that is generally preferred over Euclidean distance for high-dimension data analysis (Aggarwal et al., 2001). However, since every dimension (mean, standard deviation, skewness, kurtosis) has the same attribution in the calculation of Taxicab distance, our feature vector of four dimensions could not have the same impact on the final value due to the difference in magnitude. In such cases, Mahalanobis distance can overcome the problem while removing redundant information from correlated variables. Since distance measurement was included as edge weight in the input data of GCN, the choice can surely affect the final result. Thus, it is not surprising to observe a better result with Mahalanobis distance supporting the graph generation.

Finally, this work presents several methodological limitations. First the classification results were biased by the class imbalance of the database and the insufficient number of patients. Since the current database consists of a series of multiple MR scans per patient, it does not cover enough variability of the disease, meaning a lack of global vision of the disease. Hence, even if we carefully stop the network training before overfitting, it is hard to extract sufficient features of each MS clinical course to classify an unseen patient by the proposed network, resulting in bad output in some cases. Nevertheless, our cohort study had no bias related to the protocol acquisition, which is unique, guaranteeing the homogeneity of the data. In contrast, a multi-center study is more variable and therefore requires a precise study and corrections of bias.

5 Conclusion

Although studies on MS mainly focus on white matter and lesion analysis, morphological change in gray matter is a non-negligible aspect of the disease. A full pipeline was proposed in this study for the classification of MS clinical forms. It starts from automatic GM segmentation and surface parcellation, followed by GM thickness analysis using three different granularity of atlases, two different distance measurements, and two different age-gender normalization methods. Thus, a brain resulted in a morphological connectivity graph accompanied by a feature matrix per graph. Four rejection rates corresponding to noise elimination were applied to the graph. A graph convolutional network was performed on these graphs to exploit the hidden information behind GM morphological features. In parallel, a classic 3D convolutional neural network was applied to the brain MRI directly for comparison. The best results were generated by proportional GCN that trained on Glasser parcellation-based graphs with Mahalanobis distance measurement and 80% rejection rate. In future studies, to fully exploit its capacity for clinical image analysis, our method can be implemented on a larger database to predict patients' disease evolution and obtain the correlation between images' information and patients' disability. However, to work with such a heterogeneous study will require developing more advanced graph networks (i.e., with attention) to limit biases such as gender, age and acquisition systems.

Data availability statement

The sources code and graph data supporting the conclusions of this article can be found at: https://gitlab.in2p3.fr/thomas.grenier/msgcn-classification.

Ethics statement

The studies involving humans were approved by Local Ethics Committee (CPP Sud-Est IV) and French National Agency for Medicine and Health Products Safety (ANSM). The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

EC: Conceptualization, Investigation, Software, Writing—original draft, Data curation, Formal analysis, Methodology, Validation, Visualization, Writing—review & editing. BB: Software, Writing—review & editing, Data curation. FD-D: Funding acquisition, Supervision, Writing—review & editing, Validation, Visualization. TG: Methodology, Supervision, Validation, Writing—review & editing, Conceptualization, Formal analysis, Investigation, Visualization. DS-M: Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Validation, Writing—review & editing, Resources.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. EC was founded by the LABEX PRIMES (ANR-11-LABX-0063) of Université de Lyon, within the program "Investments for the Future" operated by the French National Research Agency (ANR).

Acknowledgments

Part of the results presented in this work were achieved using the FreeSurfer application (Fischl, 2012) through the Virtual Imaging Platform (Glatard et al., 2013), which uses the resources provided by the biomed virtual organization of the EGI infrastructure. This work was done within the framework of Observatoire Français de la Sclérose en Plaques (OFSEP), a national cohort supported by a grant provided by the French State and handled by the French National Research Agency (ANR) within the framework of the "Investments for the Future" program, under the reference ANR-10-COHO-002.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2023.1268860/full#supplementary-material

References

Aggarwal, C. C., Hinneburg, A., and Keim, D. A. (2001). "On the surprising behavior of distance metrics in high dimensional space," in *Database Theory - ICDT 2001* (Berlin; Heidelberg). doi: 10.1007/3-540-44503-X_27

Barile, B., Ashtari, P., Stamile, C., Marzullo, A., Maes, F., Durand-Dubief, F., et al. (2022). Classification of multiple sclerosis clinical profiles using machine learning and grey matter connectome. *Front. Robot. AI* 9, 926255. doi: 10.3389/frobt.2022.926255

Bronstein, M. M., Bruna, J., Cohen, T., and Veličković, P. (2021). Geometric deep learning: grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*. doi: 10.48550/arXiv.2104.13478

Corps, J., and Rekik, I. (2019). Morphological brain age prediction using multi-view brain networks derived from cortical morphology in healthy and disordered participants. *Sci. Rep*. 9, 9676. doi: 10.1038/s41598-019-46145-4

Desikan, R. S., Segonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage* 31, 968–980. doi: 10.1016/j.neuroimage.2006.01.021

Destrieux, C., Fischl, B., Dale, A., and Halgren, E. (2010). Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *Neuroimage* 53, 1–15. doi: 10.1016/j.neuroimage.2010.06.010

Dicks, E., van der Flier, W. M., Scheltens, P., Barkhof, F., and Tijms, B. M. (2020). Single-subject gray matter networks predict future cortical atrophy in preclinical Alzheimer's disease. *Neurobiol. Aging* 94, 71–80. doi: 10.1016/j.neurobiolaging.2020.05.008

Durand-Dubief, F., Belaroussi, B., Armspach, J. P., Dufour, M., Roggerone, S., Vukusic, S., et al. (2012). Reliability of longitudinal brain volume loss measurements between 2 sites in patients with multiple sclerosis: comparison of 7 quantification techniques. *Am. J. Neuroradiol.* 33, 1918–1924. doi: 10.3174/ajnr.A3107

Eitel, F., Soehler, E., Bellmann-Strobl, J., Brandt, A. U., Ruprecht, K., Giess, R. M., et al. (2019). Uncovering convolutional neural network decisions for diagnosing multiple sclerosis on conventional MRI using layer-wise relevance propagation. *Neuroimage Clin*. 24, 102003. doi: 10.1016/j.nicl.2019.102003

Eshaghi, A., Marinescu, R. V., Young, A. L., Firth, N. C., Prados, F., Cardoso, M. J., et al. (2018). Progression of regional grey matter atrophy in multiple sclerosis. *Brain* 141, 1665–1677. doi: 10.1093/brain/awy088

Eshaghi, A., Wottschel, V., Cortese, R., Calabrese, M., Sahraian, M. A., Thompson, A. J., et al. (2016). Gray matter MRI differentiates neuromyelitis optica from multiple sclerosis using random forest. *Neurology* 87, 2463–2470. doi: 10.1212/WNL.0000000000003395

Fischl, B. (2012). Freesurfer. *Neuroimage* 62, 774–781. doi: 10.1016/j.neuroimage.2012.01.021

Flaus, A., Deddah, T., Reilhac, A., Leiris, N. D., Janier, M., Merida, I., et al. (2022). Pet image enhancement using artificial intelligence for better characterization of epilepsy lesions. *Front. Med*. 9, 1042706. doi: 10.3389/fmed.2022.1042706

Folego, G., Weiler, M., Casseb, R. F., Pires, R., and Rocha, A. (2020). Alzheimer's disease detection through whole-brain 3D-CNN MRI. *Front. Bioeng. Biotechnol*. 8, 534592. doi: 10.3389/fbioe.2020.534592

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., et al. (2016). A multi-modal parcellation of human cerebral cortex. *Nature* 536, 171–178. doi: 10.1038/nature18933

Glatard, T., Lartizien, C., Gibaud, B., Silva, R. F. D., Forestier, G., Cervenansky, F., et al. (2013). A virtual imaging platform for multi-modality medical image simulation. *IEEE Trans. Med. Imaging* 32, 110–118. doi: 10.1109/TMI.2012.2220154

Goodin, D. S. (2014). "Chapter 11: The epidemiology of multiple sclerosis: insights to disease pathogenesis," in *Multiple Sclerosis and Related Disorders, volume 122 of Handbook of Clinical Neurology*, ed D. S. Goodin (Elsevier), 231–266. doi: 10.1016/B978-0-444-52001-2.00010-8

Guo, Y., Nejati, H., and Cheung, N. M. (2017). "Deep neural networks on graph signals for brain imaging analysis," in *2017 IEEE International Conference on Image Processing (ICIP)* (Beijing). doi: 10.1109/ICIP.2017.8296892

Hagmann, P., Kurant, M., Gigandet, X., Thiran, P., Wedeen, V. J., Meuli, R., et al. (2007). Mapping human whole-brain structural networks with diffusion MRI. *PLoS ONE* 2, e597. doi: 10.1371/journal.pone.0000597

Hawkins, R., Shatil, A. S., Lee, L., Sengupta, A., Zhang, L., Morrow, S., et al. (2020). Reduced global efficiency and random network features in patients with relapsing-remitting multiple sclerosis with cognitive impairment. *Am. J. Neuroradiol.* 41, 449–455. doi: 10.3174/ajnr.A6435

Huang, Y., Xu, J., Zhou, Y., Tong, T., Zhuang, X., and the Alzheimer's Disease Neuroimaging Initiative (ADNI) (2019). Diagnosis of Alzheimer's disease via multi-modality 3D convolutional neural network. *Front. Neurosci.* 13, 509. doi: 10.3389/fnins.2019.00509

Jutten, K., Mainz, V., Gauggel, S., Patel, H. J., Binkofski, F., Wiesmann, M., et al. (2019). Diffusion tensor imaging reveals microstructural heterogeneity of normal-appearing white matter and related cognitive dysfunction in glioma patients. *Front. Oncol*. 9, 536. doi: 10.3389/fonc.2019.00536

Kipf, T. N., and Welling, M. (2017). "Semi-supervised classification with graph convolutional networks," in *5th International Conference on Learning Representations, ICLR 2017* (Toulon).

Kocevar, G., Stamile, C., Hannoun, S., Cotton, F., Vukusic, S., Durand-Dubief, F., et al. (2016). Graph theory-based brain connectivity for automatic classification of multiple sclerosis clinical courses. *Front. Neurosci.* 10, 478. doi: 10.3389/fnins.2016.00478

Kong, Y., Gao, J., Xu, Y., Pan, Y., Wang, J., and Liu, J. (2019). Classification of autism spectrum disorder by combining brain connectivity and deep neural network classifier. *Neurocomputing* 324, 63–68. doi: 10.1016/j.neucom.2018.04.080

Ktena, S. I., Parisot, S., Ferrante, E., Rajchl, M., Lee, M., Glocker, B., et al. (2018). Metric learning with spectral graph convolutions on brain connectivity networks. *Neuroimage* 169, 431–442. doi: 10.1016/j.neuroimage.2017.12.052

Leclerc, S., Smistad, E., Pedrosa, J., Ostvik, A., Cervenansky, F., Espinosa, F., et al. (2019). Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. *IEEE Trans. Med. Imaging* 38, 2198–2210. doi: 10.1109/TMI.2019.2900516

Lublin, F. D., Reingold, S. C., Cohen, J. A., Cutter, G. R., Sorensen, P. S., Thompson, A. J., et al. (2014). Defining the clinical course of multiple sclerosis: the 2013 revisions. *Neurology* 83, 278–86. doi: 10.1212/WNL.0000000000000560

MacDonald, D., Kabani, N., Avis, D., and Evans, A. C. (2000). Automated 3-d extraction of inner and outer surfaces of cerebral cortex from mri. *Neuroimage* 12, 340–356. doi: 10.1006/nimg.1999.0534

Mahjoub, I., Mahjoub, M. A., Rekik, I., Weiner, M., Aisen, P., Petersen, R., et al. (2018). Brain multiplexes reveal morphological connectional biomarkers fingerprinting late brain dementia states. *Sci. Rep*. 8, 1–14. doi: 10.1038/s41598-018-21568-7

Marti-Juan, G., Frias, M., Garcia-Vidal, A., Vidal-Jordana, A., Alberich, M., Calderon, W., et al. (2022). Detection of lesions in the optic nerve with magnetic resonance imaging using a 3d convolutional neural network. *Neuroimage Clin*. 36, 103187. doi: 10.1016/j.nicl.2022.103187

Marzullo, A., Kocevar, G., Stamile, C., Durand-Dubief, F., Terracina, G., Calimeri, F., et al. (2019). Classification of multiple sclerosis clinical profiles via graph convolutional neural networks. *Front. Neurosci.* 13, 594. doi: 10.3389/fnins.2019.00594

Mure, S., Grenier, T., Guttmann, C. R. G., Cotton, F., and Benoit-Cattin, H. (2016). "Classification of multiple sclerosis lesion evolution patterns a study based on unsupervised clustering of asynchronous time-series," in *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)* (Prague), 1315–1319. doi: 10.1109/ISBI.2016.7493509

Muthuraman, M., Fleischer, V., Kolber, P., Luessi, F., Zipp, F., and Groppa, S. (2016). Structural brain network characteristics can differentiate cis from early rrms. *Front. Neurosci.* 10, 14. doi: 10.3389/fnins.2016.00014

Narayana, P. A., Govindarajan, K. A., Goel, P., Datta, S., Lincoln, J. A., Cofield, S. S., et al. (2013). Regional cortical thickness in relapsing remitting multiple sclerosis: a multi-center study. *Neuroimage Clin.* 2, 120–131. doi: 10.1016/j.nicl.2012.11.009

Polman, C. H., Reingold, S. C., Banwell, B., Clanet, M., Cohen, J. A., Filippi, M., et al. (2011). Diagnostic criteria for multiple sclerosis: 2010 revisions to the Mcdonald criteria. *Ann. Neurol.* 69, 292–302. doi: 10.1002/ana.22366

Raamana, P. R., and Strother, S. C. (2018). graynet: single-subject morphometric networks for neuroscience connectivity applications. *J. Open Source Softw.* 3, 924. doi: 10.21105/joss.00924

Rimkus, C. M., Schoonheim, M. M., Steenwijk, M. D., Vrenken, H., Eijlers, A. J., Killestein, J., et al. (2019). Gray matter networks and cognitive impairment in multiple sclerosis. *Multiple Scler. J.* 25, 382–391. doi: 10.1177/1352458517751650

Rocca, M. A., Valsasina, P., Meani, A., Pagani, E., Cordani, C., Cervellin, C., et al. (2021). Network damage predicts clinical worsening in multiple sclerosis: a 6.4-year study. *Neurol. Neuroimmunol. NeuroInflam.* 8, e1006. doi: 10.1212/NXI.0000000000001006

Rubinov, M., and Sporns, O. (2010). Complex network measures of brain connectivity: uses and interpretations. *Neuroimage* 52, 1059–1069. doi: 10.1016/j.neuroimage.2009.10.003

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2020). Grad-CAM: visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.* 128, 336–359. doi: 10.1007/s11263-019-01228-7

Sporns, O. (2018). Graph theory methods: applications in brain networks. *Dialog. Clin. Neurosci.* 20, 111–121. doi: 10.31887/DCNS.2018.20.2/osporns

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15, 1929–1958. Available online at: https://dl.acm.org/doi/10.5555/2627435.2670313

Steenwijk, M. D., Geurts, J. J., Daams, M., Tijms, B. M., Wink, A. M., Balk, L. J., et al. (2016). Cortical atrophy patterns in multiple sclerosis are non-random and clinically relevant. *Brain* 139(Pt 1), 115–126. doi: 10.1093/brain/awv337

Thompson, A. J., Banwell, B. L., Barkhof, F., Carroll, W. M., Coetzee, T., Comi, G., et al. (2018). Diagnosis of multiple sclerosis: 2017 revisions of the mcdonald criteria. *Lancet Neurol.* 17, 162–173. doi: 10.1016/S1474-4422(17)30470-2

Wargnier-Dauchelle, V., Grenier, T., Durand-Dubief, F., Cotton, F., and Sdika, M. (2023). A weakly supervised gradient attribution constraint for interpretable classification and anomaly detection. *IEEE Trans. Med. Imaging* 42, 3336–3347. doi: 10.1109/TMI.2023.3282789

frontiers | Frontiers in Computational Neuroscience

TYPE Original Research
PUBLISHED 07 February 2024
DOI 10.3389/fncom.2024.1328699

OPEN ACCESS

EDITED BY
Pablo Martinez-Cañada,
University of Granada, Spain

REVIEWED BY
Angeliki Zarkali,
University College London, United Kingdom
Vignayanandam Ravindernath Muddapu,
Ecole polytechnique fédérale de Lausanne (EPFL), Switzerland

*CORRESPONDENCE
Marina C. Ruppert-Junck
✉ marina.ruppert@uni-marburg.de

RECEIVED 27 October 2023
ACCEPTED 22 January 2024
PUBLISHED 07 February 2024

CITATION
Ruppert-Junck MC, Kräling G, Greuel A, Tittgemeyer M, Timmermann L, Drzezga A, Eggers C and Pedrosa D (2024) Random forest analysis of midbrain hypometabolism using [18F]-FDG PET identifies Parkinson's disease at the subject-level.
Front. Comput. Neurosci. 18:1328699.
doi: 10.3389/fncom.2024.1328699

Random forest analysis of midbrain hypometabolism using [18F]-FDG PET identifies Parkinson's disease at the subject-level

Marina C. Ruppert-Junck[1,2,3]*, Gunter Kräling[2], Andrea Greuel[4], Marc Tittgemeyer[5,6], Lars Timmermann[1,2,3], Alexander Drzezga[6,7,8], Carsten Eggers[1,9] and David Pedrosa[1,2,3]

[1]Department of Neurology, Philipps-University of Marburg, Marburg, Germany, [2]Clinic for Neurology, University Hospital Gießen and Marburg GmbH, Marburg, Germany, [3]Center for Mind, Brain and Behavior, Philipps-University of Marburg and Justus-Liebig University Gießen, Marburg, Germany, [4]Department of Psychiatry, Psychotherapy and Psychosomatics, Vivantes Hospital Neukölln, Berlin, Germany, [5]Max Planck Institute for Metabolism Research, Cologne, Germany, [6]Cluster of Excellence in Cellular Stress and Aging Associated Disease (CECAD), Cologne, Germany, [7]Cognitive Neuroscience, Institute of Neuroscience and Medicine (INM-2), Research Center Jülich, Jülich, Germany, [8]Multimodal Neuroimaging Group, Department of Nuclear Medicine, Medical Faculty, University Hospital Cologne, Cologne, Germany, [9]Department of Neurology, Knappschaftskrankenhaus Bottrop, Bottrop, Germany

Parkinson's disease (PD) is currently diagnosed largely on the basis of expert judgement with neuroimaging serving only as a supportive tool. In a recent study, we identified a hypometabolic midbrain cluster, which includes parts of the substantia nigra, as the best differentiating metabolic feature for PD-patients based on group comparison of [18F]-fluorodeoxyglucose ([18F]-FDG) PET scans. Longitudinal analyses confirmed progressive metabolic changes in this region and, an independent study showed great potential of nigral metabolism for diagnostic workup of parkinsonian syndromes. In this study, we applied a machine learning approach to evaluate midbrain metabolism measured by [18F]-FDG PET as a diagnostic marker for PD. In total, 51 mid-stage PD-patients and 16 healthy control subjects underwent high-resolution [18F]-FDG PET. Normalized tracer update values of the midbrain cluster identified by between-group comparison were extracted voxel-wise from individuals' scans. Extracted uptake values were subjected to a random forest feature classification algorithm. An adapted leave-one-out cross validation approach was applied for testing robustness of the model for differentiating between patients and controls. Performance of the model across all runs was evaluated by calculating sensitivity, specificity and model accuracy for the validation data set and the percentage of correctly categorized subjects for test data sets. The random forest feature classification of voxel-based uptake values from the midbrain cluster identified patients in the validation data set with an average sensitivity of 0.91 (Min: 0.82, Max: 0.94). For all 67 runs, in which each of the individuals was treated once as test data set, the test data set was correctly categorized by our model. The applied feature importance extraction consistently identified a subset of voxels within the midbrain cluster with highest importance across all runs which spatially converged with the left substantia nigra. Our data suggest midbrain metabolism measured by [18F]-FDG PET as a promising diagnostic

imaging tool for PD. Given its close relationship to PD pathophysiology and very high discriminatory accuracy, this approach could help to objectify PD diagnosis and enable more accurate classification in relation to clinical trials, which could also be applicable to patients with prodromal disease.

KEYWORDS

Parkinson's disease, imaging biomarker, machine learning, random forest, metabolic imaging

1 Introduction

Parkinson's disease (PD) is the second most common neurodegenerative disorder (Lau and de Breteler, 2006) and characterized by a spread of α-synuclein containing Lewy bodies and the loss of neuromelanin pigmented neurons in the substantia nigra. The consequential depletion of dopaminergic transmission to lateral nigral projection areas (Kish et al., 1988), and primarily the posterior putamen, results in aberrant striato-thalamo-cortical information processing causing motor symptoms like bradykinesia or rigidity (Albin et al., 1989; DeLong, 1990). Diagnosing the condition can yet be a challenge for physicians, as no reliable biomarker is currently available and only clinical criteria can be used (Postuma et al., 2015). Especially at early stages, when symptoms were present for <5 years, a diagnostic accuracy of only 53% in PD patients has been reported (Adler, 2014). Not only does this limit disease management, but it also underlies the dilemma that neuroprotective therapies are likely to fail if used too late. Therefore, one of the main goals of PD research is to find biomarkers that can be applied easily and early and are as objective as possible (Adler, 2014). Future-oriented concepts claim a biological staging system for PD continuum, whereby degeneration of midbrain dopaminergic neurons represents a crucial, universal feature of the disease.

Currently, there is no causative therapy for PD, but significant efforts have been directed at neuroprotective therapies targeting molecular pathways before disease onset. Nigral neurons are highly energy consuming neural populations relying on effective mitochondria which makes them vulnerable to exhaustion possibly contributing to neurodegeneration (Braak et al., 2006b; Seibyl et al., 2012). When patients experience motor symptoms, typically up to 70% of nigral neurons have already been depleted. Due to lack of applicable α-synuclein tracers, no possibility exists to date for *in vivo* examination of α-synuclein load (Fearnley and Lees, 1991). However, there are indirect measures of nigral dopaminergic cell loss, particularly in the field of molecular imaging. As a surrogate marker for presynaptic dopaminergic activity, semiquantitative analysis of [^{123}I]-FP-CIT-SPECT regularly serves as supportive diagnostic tool. To enable diagnosis from a pathophysiological rather than clinical perspective and demonstrate prospects for reducing disease progression through interventions, indicators of biological processes that are immediately applicable and show a strong correlation with established neuropathological markers are urgently needed, especially at early disease stages (Höglinger et al., 2023).

Molecular imaging has been proposed to trace ongoing disease-related processes and subclinical changes. In a recent study applying [^{18}F]-fluorodeoxyglucose PET([^{18}F]-FDG PET), which uses a labeled glucose analogon, we identified a hypometabolic midbrain cluster as the best differentiating metabolic feature for PD-patients compared to healthy controls (Ruppert et al., 2020). The level of individual hypometabolism was found to match contralateral motor symptoms. Subsequent examinations of a subset of these patients over the course of the disease confirmed progressive metabolic changes in this region which were accompanied by worsened motor symptoms (Steidel et al., 2022). An independent study reported nigral metabolism in PD based on non-high-resolution [^{18}F]-FDG PET and demonstrated the great potential of nigral metabolism for differential diagnostics of parkinsonian syndromes (Schröter et al., 2022). Nigral hypometabolism was worse in entities associated with most severe nigrostriatal pathology (Schröter et al., 2022).

Hence, there is a growing body of evidence for the midbrain as an important region to differentiate PD patients from healthy controls based on metabolic group comparisons. Nevertheless, for applications as diagnostic marker, the informative value of the measure for the individual needs to be verified. In this context, machine learning approaches are increasingly applied to evaluate the discriminative accuracy of measures under consideration. Several studies have conducted region-of-interest wise machine learning analysis of [^{18}F]-Desmethoxyfallypride PET data extracted from either striatal structures or whole-brain and revealed an accuracy of 59.7% or of about 70% for differentiating between PD patients and atypical parkinsonism (Segovia et al., 2015, 2017a). Studies focusing on [^{18}F]-FDG PET as diagnostic marker have rarely been carried out and focused on whole brain scans, or an atlas-based parcellation but did not include the midbrain region despite its crucial role in neuropathology (Wu et al., 2019). In this study, we evaluated midbrain metabolism derived by high-resolution [^{18}F]-FDG PET as a diagnostic marker for PD using random forest analysis.

Abbreviations: [18F]-FDG PET, [18F]-fluorodeoxyglucose positron emission tomography; DD, disease duration; FWE, family-wise error; FWHM, full-width at half-maxium; LEDD, levodopa equivalent daily dose; PD, Parkinson's disease; MMSE, Mini-Mental state examination; MNI, Montreal Neurological Institute; SNpc, substantia nigra pars compacts; SNpr, substantia nigra pars reticulata; UPDRS, unified Parkinson's disease rating scale; VTA, ventral tegmental area.

2 Materials and methods

2.1 Participants

All participants provided informed consent to their data analyses in conformation with the Declaration of Helsinki. The study was confirmed by the local ethics committee (EK12-265) and the Federal Bureau for Radiation Protection. In total, 25 healthy control subjects and 60 patients with clinically established PD were enrolled. Patient recruitment was carried out at the University Hospital of Cologne and affiliated neurology practices, whereas healthy control participants were recruited via advertising. Exclusion criteria were age < 40 years, suspected atypical parkinsonian syndromes, advanced parkinsonism, i.e., Hoehn and Yahr stages >3 (Hoehn and Yahr, 1967), dementia, neurological diseases other than PD, and any safety concerns for MRI scanning. In order to exclude patients with dementia, criteria published by the Movement Disorder Society including a neuropsychological test battery and an assessment of the patient's ability to manage daily life (Emre et al., 2007) were applied. The Mini-Mental State Examination (MMSE) was used as cognitive screening tool (Folstein et al., 1975). Clinical examination and functional imaging were conducted at the Max Planck institute for Metabolism Research Cologne and the University Hospital Cologne, Department of Neurology. Patients were examined in the OFF state, defined as a 12-h period without dopaminergic medication (Langston et al., 1992) (72 h in cases of dopamine agonists). Levodopa-equivalent daily dose (LEDD) was calculated for total antiparkinsonian medication based on standard conventions (Tomlinson et al., 2010). Disease severity was quantified by the Unified Parkinson's Disease Rating Scale (UPDRS) part III (Fahn et al., 1987).

Statistical analysis of demographical, clinical and behavioral data was performed in R (R-project for statistical computing, Vienna, Austria). Depending on the assumptions met, parametric or non-parametric tests were performed. Results were considered significant if $p < 0.05$.

2.2 [^{18}F]-FDG PET acquisition and preprocessing

All PET scans were acquired on an ECAT HRRT-PET-Scanner (CTI) at the Max-Planck-Institute for Metabolism Research in Cologne after overnight fasting and OFF dopaminergic medication. Under standardized conditions (dimmed light, closed eyes, quiet room) subjects were positioned along the kantho-meatal line. Following a transmission scan, 185 MBq of the radioligand was injected intravenously and tomographic images were acquired in dynamic PET scans (60 min). Using camera-specific filters, PET data were corrected for attenuation and scattered radiation, and reconstructed to 207 slices with a 256 × 256 matrix and 1.22 mm voxel size, creating one frame per 10 min. Frames were realigned for motion correction by rigid-body transformation and frames numbered three to six were averaged into one static image for further analysis. The data set used in the presented analysis has been analyzed in previous publications from different research perspectives (Greuel et al., 2020; Ruppert et al., 2020, 2021; Steidel et al., 2022) and once in context of machine learning but with a whole brain approach and in specific combination with metabolomic data (Glaab et al., 2019).

Static PET scans were spatially normalized into Montreal Neurological Institute (MNI) space in SPM12 (www.fil.ion.ucl.ac.uk/spm/software/spm12, Wellcome Trust Center for Human Imaging, London) using an [^{18}F]-FDG PET template for elderly subjects (Della Rosa et al., 2014) and smoothed with a Gaussian kernel of 6 mm full-width at half-maximum (FWHM). The midbrain cluster, reflecting hypometabolic regions in our PD cohort and defining our regions of interest in the current analysis, was derived by a voxel-wise group comparison as specified in our previous work (Ruppert et al., 2020), with the number of included subjects referring to all subjects with [^{18}F]-FDG-PET scans here (PD = 51, healthy controls = 16). PET data were proportionally scaled with reference to the global mean as implemented in SPM. Group comparisons were carried out via a general linear model in SPM12. Results were considered significant when $p < 0.05$ after family-wise error (FWE) rate correction at cluster level (Figure 1).

Voxel-wise normalized (proportional scaling) uptake values were extracted from the obtained midbrain cluster (Figure 1) for all subjects with the region of interest toolbox Marsbar (Brett et al., 2002). A class label column was added for supervised machine learning with 0 for healthy control and 1 for PD class. To check whether the approach also performs with a not data-driven region, which would enable easier transferability to independent data sets, we repeated the machine learning analysis with uptake measures from an atlas-based midbrain region (Talairach-Daemon atlas, WFU PickAtlas, RRID:SCR_007378) and with a whole brain gray matter mask [ICBM 2009c non-linear symmetric, FSL (Collins et al., 1999)].

2.3 Machine learning analysis

Extracted uptake values were subjected to machine learning analyses, applying different feature classification algorithms. An adapted leave-one-out cross validation approach with reassignment of the training and validation test set (70:30) at every step was applied for testing robustness of the model for differentiating between patients and controls. First, we compared performance of the most commonly applied machine learning classification algorithms in our data set using PyCaret tool (https://pycaret.org/) in python. Specifically, the following algorithms were tested: Extra Tree Classifier, Naive Bayes, K Neighbors Classifier, Random Forest Classifier, Logistic Regression, Ada Boost Classifier, Light Gradient Boosting Machine, Dummy Classifier, Decision Tree Classifier, Ridge Classifier, Linear Discriminant Analysis, Gradient Boosting Classifier, Support-Vector-Machine—Linear Kernel, and Quadratic Discriminant Analysis. The random forest ensemble algorithm is one of the most widely applied machine learning techniques for classification problems. It is an ensemble learning method, which used a combination of decision trees to make predictions. Each decision tree is trained based on a subset of the data generated by Bootstrap-sampling. A prediction is

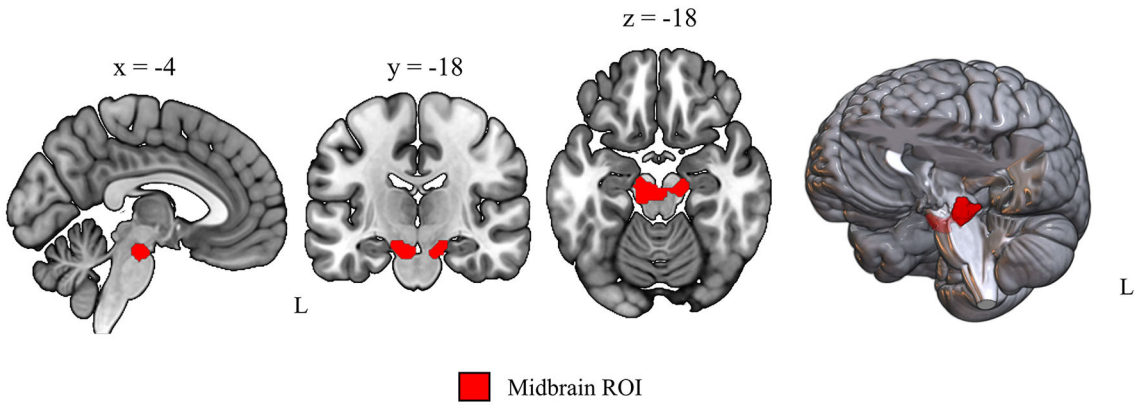


FIGURE 1
Data-driven region of interest used for voxel-wise uptake extraction. Sagittal, coronal and axial view of the midbrain region of interest obtained by voxel-wise group comparison of [^{18}F]-FDG PET scans from 16 healthy controls and 51 PD patients ($p < 0.01$ after FWE cluster level correction, $t = 6.46$, cluster size = 376 voxels).

offered by every decision tree and the final prediction of the model is driven by the majority of votes on the predictions (cf. Figure 2). For hyperparameter tuning, the default option in PyCaret was used which applies a random grid search. Robustness and discriminatory performance of the model across all runs was evaluated in four ways: (1) for model evaluation we averaged performance measures across 67 runs with one of the 67 subjects left out and dividing the remaining 66 in respective training and validation data sets (70:30), (2) in each of these runs a 10-fold nested cross-validation was performed on the training data set with 1-fold serving as validation and 9-fold serving as training data set per one of the 10 validations, (3) in each of the 67 runs, an independent validation data set was used to evaluate model performance by calculating sensitivity and specificity for the resulting confusion matrix, and (4) the percentage of correctly categorized subjects for test data sets (the one not considered in training and validation per run) with reference to movement disorder expert opinion. Included subjects per training and validation data set can be found on our GitHub repository (https://github.com/ruppertm/Midbrain-FDG-PD.git). An evaluation of potential between-group differences in clinical and demographic variables is reported in the Supplementary material.

Feature importance reflects the relevance an individual feature has for correct classification. Feature importance for individual voxels was calculated according to the default settings implemented in PyCaret which refers to the method in the scikit-learn library (mean decrease impurity). Each voxel's coordinates (in all three axes) derived by Marsbar were transformed into MNI-space coordinates using the provided transformation matrix. 3D displays were created in MRIcroGL using the Marsbar coordinates and feature importance values derived via the feature importance analysis above. Spatial colocalization with dopaminergic midbrain nuclei was verified using the Automated Anatomical Labeling version 3 (AALv3) atlas. The code generated to analyze all data is freely available on GitHub (https://github.com/ruppertm/Midbrain-FDG-PD.git).

3 Results

3.1 Cohort characteristics

[^{18}F]-FDG PET scans were available for 51 patients with MRI (66.45 ± 8.53 years, 18 female) and 16 control subjects (64.63 ± 8.33, 9 female) with no significant differences in terms of age, sex and general cognitive performance (cf. Table 1). The included patients were moderately affected with an average UPDRS-III of 25.10 ± 9.54 points and 453.88 ± 244.72 mg LEDD. Detailed information on included participants (mean ± standard deviation) can be found in Table 1. Across all runs, there were no between-group differences in terms of age or motor severity between training and validation data set (Supplementary material).

3.2 Random forest analysis

3.2.1 Classification based on midbrain [^{18}F]-FDG uptake

Across all runs, the random forest algorithm performed best in most cases. Therefore, random forest classifier analysis was applied to evaluate the diagnostic potential of midbrain metabolism in our study. The random forest feature classification of voxel-based uptake values from the 376 voxels spanning midbrain cluster distinguished between the groups with an average sensitivity of 0.91 (Min: 0.82, Max: 0.94) in the validation data set (Table 2). For all 67 runs, in which each of the individuals was treated once as test data set, the test data set was correctly categorized by our model. The separately performed analyses with uptake values from the midbrain atlas region showed slightly lower sensitivity measures, and lower specificity and accuracy (Table 2). Whole-brain analysis revealed a slightly better sensitivity, but worse specificity and accuracy (Table 2).

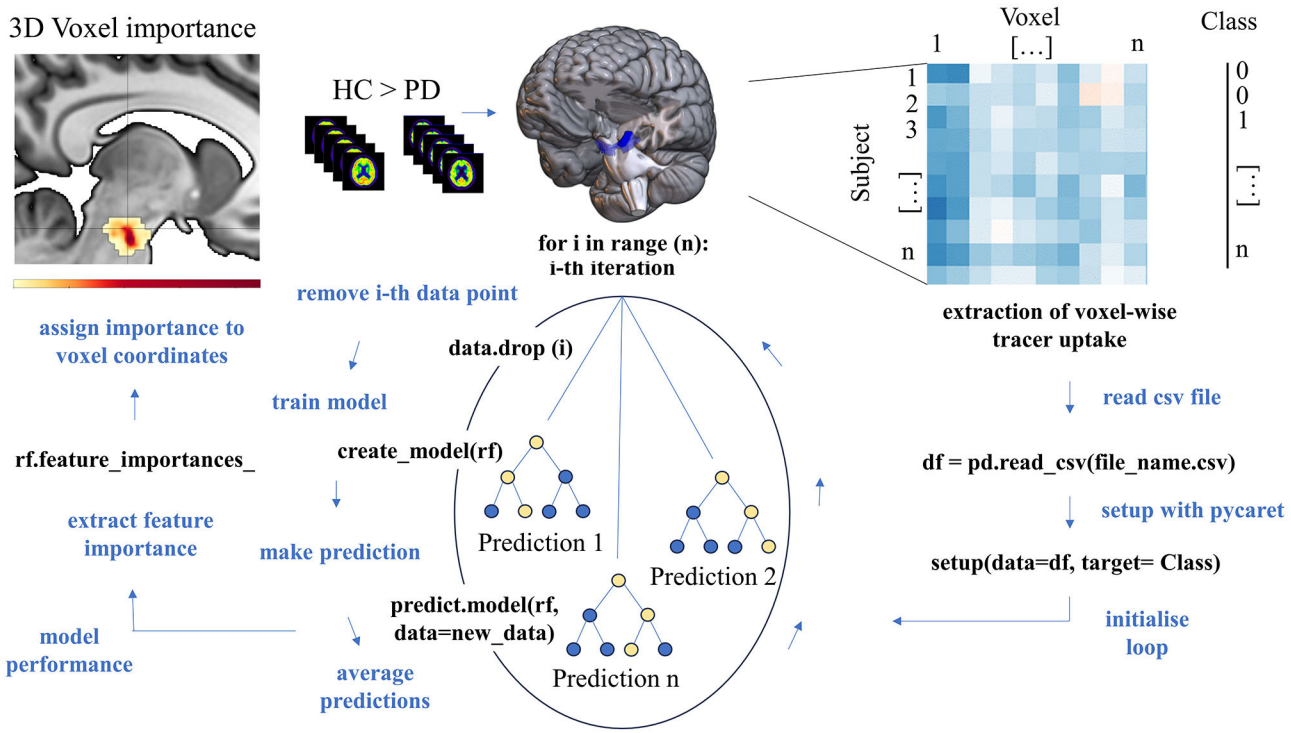


FIGURE 2
Schematic representation of the data extraction process and random forest analysis. The subcortical region of interest was defined by a between-group comparison of [18F]-FDG PET scans **(middle top)**. Voxel-wise uptake values were extracted from that region for every subject from normalized [18F]-FDG PET scans **(top right)** and subjected to random forest analysis using PyCaret. After importing the data, random forest model was conducted on the data from all but one subject (each subject was left out once) with 10-fold cross-validation. The data set was splitted into a training and validation data set (70:30) to derivate accuracy measures and the class of the removed subject was predicted based on that model. The model's accuracy measures, and each voxel's feature importance were averaged across all runs and reassigned to voxel coordinates to enable 3D representation of voxels with highest importance for class decision.

TABLE 1 **Demographic, clinical and behavioral characteristics of the [18F]-FDG PET cohort including all PD patients and healthy controls.**

Groups	HC ($n = 16$)	PD ($n = 51$)	Statistics	p-value
Age (in years)	64.63 ± 8.33	66.45 ± 8.53	$t = 0.75$	0.455
Female (%)	9 (56.25%)	18 (35.29%)	$X^2 = 1.44$	0.231
DD (in years)	-	4.56 ± 3.29	-	-
UPDRS III total	-	25.10 ± 9.54	-	-
LEDD (in mg)	-	453.88 ± 244.72	-	-
MMSE	28.94 ± 1.00	28.37 ± 1.82	$W = 351.5$	0.392

Between-group comparison of numeric variables was performed via t-tests or Mann-Whitney U tests. Dichotomous variables were compared via chi-square test. DD, disease duration; HC, healthy control subjects; Levodopa equivalent daily dose; PD, Parkinson's disease; MMSE, Mini-Mental Status Examination.

TABLE 2 **Accuracy measures of the random forest classifier model based on [18F]-FDG PET uptake for the data-driven region of interest, midbrain atlas region, and whole brain gray matter.**

Model performance (mean ± SD)	Accuracy	Sensitivity	Specificity
Midbrain (data-driven)	0.83 ± 0.06	0.91 ± 0.03	0.67 ± 0.14
Midbrain (atlas)	0.82 ± 0.04	0.88 ± 0.05	0.63 ± 0.15
Whole brain gray matter mask	0.76 ± 0.04	0.98 ± 0.03	0.10 ± 0.11

SD, standard deviation.

3.2.2 Feature importance

Since our region of interest is closely related to PD pathophysiology and we included individual voxels as features in our model, the spatial location of features with greatest importance for the class decision was of great interest. The applied feature importance extraction consistently identified a subset of voxels within the midbrain cluster with highest importance across all runs. Among the top voxels with highest importance across all runs were V70 (0.029, MNI: x = −8, y = −20, z = −22) and V148 (0.021, MNI: x = 14, y = −18, z = −20, see Supplementary Table 1 for all values). The two voxels with highest importance were localized in the left ventrolateral tier of the midbrain cluster and next to the atlas region substantia nigra pars compacts (SNpc) from AALv3 atlas (cf. Figure 3 top, Supplementary Figure S1). As indicated by overlay plots in Supplementary Figure S1, there is a spatial overlap between midbrain voxels with high importance and dopaminergic midbrain nuclei. Among the nuclei with a spatial convergence were: left SNpc, left substantia nigra pars reticulata (SNpr), left ventral tegmental areas (VTA), right SNpc, right SNpr, and right VTA.

The left SNpc was the atlas region that had the greatest spatial overlap with left-sided voxels of highest importance (dark orange-to-red color) (cf. Figure 3 top, Supplementary Figure S1). Voxels with a feature importance above 0.008 overlapped exclusively with the left SNpc. Right-sided voxels with highest importance were localized more laterally. The separate analysis performed with an atlas-based midbrain region revealed nearly identical coordinates for voxels with highest feature importance (V332 MNI: $x = -6$, $y = -22$, $z = -20$, see Supplementary Figure S2). Including whole brain gray matter [^{18}F]-FDG uptake per voxel in a separate analysis, also indicated that our defined region is the most important region for classification (cf. Supplementary Figure S3).

4 Discussion

In this study, we demonstrate the diagnostic potential of midbrain [^{18}F]-FDG uptake for PD. In a cohort of well-characterized mild-to-moderately affected patients, we showed that it may differentiate between patients and controls with high precision. The presented analyses were motivated by the previous description of the cohort, highlighting the hypometabolic midbrain cluster as the region that exhibited the highest deficit in PD that correlated with contralateral clinical severity (Ruppert et al., 2020) and showed disease-related decline over time (Steidel et al., 2022). In order to evaluate the informative value of [^{18}F]-FDG uptake within that region for the individual's classification, a random forest feature classification algorithm was applied with an adapted leave-one-out cross validation approach. Across all runs, the individual test data set was correctly categorized by our model. The applied feature importance extraction consistently identified a subset of voxels within the midbrain cluster with highest importance for class decision across all runs, which spatially overlapped with the left substantia nigra pars compacta. Our results confirm that [^{18}F]-FDG uptake in the midbrain is a promising neuroimaging feature with spatial convergence to known pathophysiology that is feasible in the individual patient and can be similarly applied to independent cohorts using midbrain atlas regions.

The loss of dopaminergic cells in the midbrain is a histopathological hallmark of PD and serves as neurobiological correlate of its progression (Damier et al., 1999). However, the significant denervation in the lateral substantia nigra prior to the onset of symptoms in those affected has not been clinically utilized due to a lack of suitable *in-vivo* examination techniques. Notably, particular voxels within our region of interest hold significant importance from a neurobiological viewpoint. There is a spatial overlap of voxels with a feature importance above 0.008 located in the left SNpc, substantiating the hypothesis that the observed hypometabolism might indicate a relationship to degenerating nigral cells or lowered metabolic activity in these naturally energy-demanding cells (Braak et al., 2006a; Seibyl et al., 2012). Corresponding to our earlier analyses, a higher count of voxels with increasing significance for class decision were located in the left midbrain. Our results complement the previous studies in the sense that exactly this region is suitable for the classification of an individual with high precision.

Machine learning techniques are used to identify elusive patterns that are difficult to detect using conventional statistical methods and to test their predictive power at individual level (Peng et al., 2020). Notably, despite certain efforts to apply machine learning to [^{18}F]-FDG-uptake for identifying-PD patients (Shen et al., 2019; Wu et al., 2019), none have targeted the midbrain region specifically. Another study has reported the identification of critical diagnostic features in the midbrain based on deep-learning, and claimed that this region, despite its crucial involvement in PD pathophysiology, has not been considered in conventional [^{18}F]-FDG PET studies (Zhao et al., 2019). Yet, several parallels might be drawn to previous attempts of applying machine learning to PET data of PD cohorts. Wu et al. (2019) extracted radiomic features from PET images using atlas regions excluding the midbrain (Wu et al., 2019). Shen et al. (2019) followed an approach with Group Lasso Sparse Deep Belief Network (GLS-DBN) for identifying PD based on [^{18}F]-FDG PET scans. Both studies report a diagnostic accuracy comparable to our results (Shen et al., 2019; Wu et al., 2019), but do not elaborate the importance of a specific subcortical region that has a close association with the known pathology as our results do. Another study has conducted a machine learning analysis with the here presented [^{18}F]-FDG PET data set but focused on whole brain uptake for PD diagnosis. Our region of interest-based approach revealed higher accuracy for the PET modality (Glaab et al., 2019). In line with our findings, Segovia and colleagues also reported a higher diagnostic accuracy with a focus on specific disease-related regions of interest rather than whole brain analysis in a dopaminergic PET study (Segovia et al., 2015, 2017b). A combination of multiple imaging modalities, supported by a specific focus on disease-related regions as in the presented approach, could increase model performance and could be crucial for tracking disease progression. Particularly, our results may be of relevance for efforts of establishing objective markers for a purely biological-based staging system for the disease spectrum, as recently proposed and already established for other neurodegenerative disorders (Chahine et al., 2023; Höglinger et al., 2023). In the latter conceptual framework, degeneration of dopaminergic neurons in the midbrain is a crucial feature evident universally in PD syndromes (Chahine et al., 2023) and present in both presumed retro- and anterograde spreading subtypes. This fact and the recognized significance of FDG-PET patterns in PD (Höglinger et al., 2023) lends our target an important status with potential applicability within the framework.

Based on Schröter et al. (2022)'s findings, our approach may additionally serve to distinguish between atypical Parkinson's syndromes and PD. The fact that the latter study reported similar evidence for midbrain hypometabolism based on not high-resolution PET data suggests that the presented approach is likely to be replicated with standard clinical PET data and therefore easily integrable into clinical practice.

4.1 Limitations

One limitation of this study is the small sample size, especially in the healthy control group, which especially contributes to very unbalanced validation data sets. The limited number of controls was a deliberate decision in line with the specifications of the Federal Office for Radiation Protection to include as few healthy

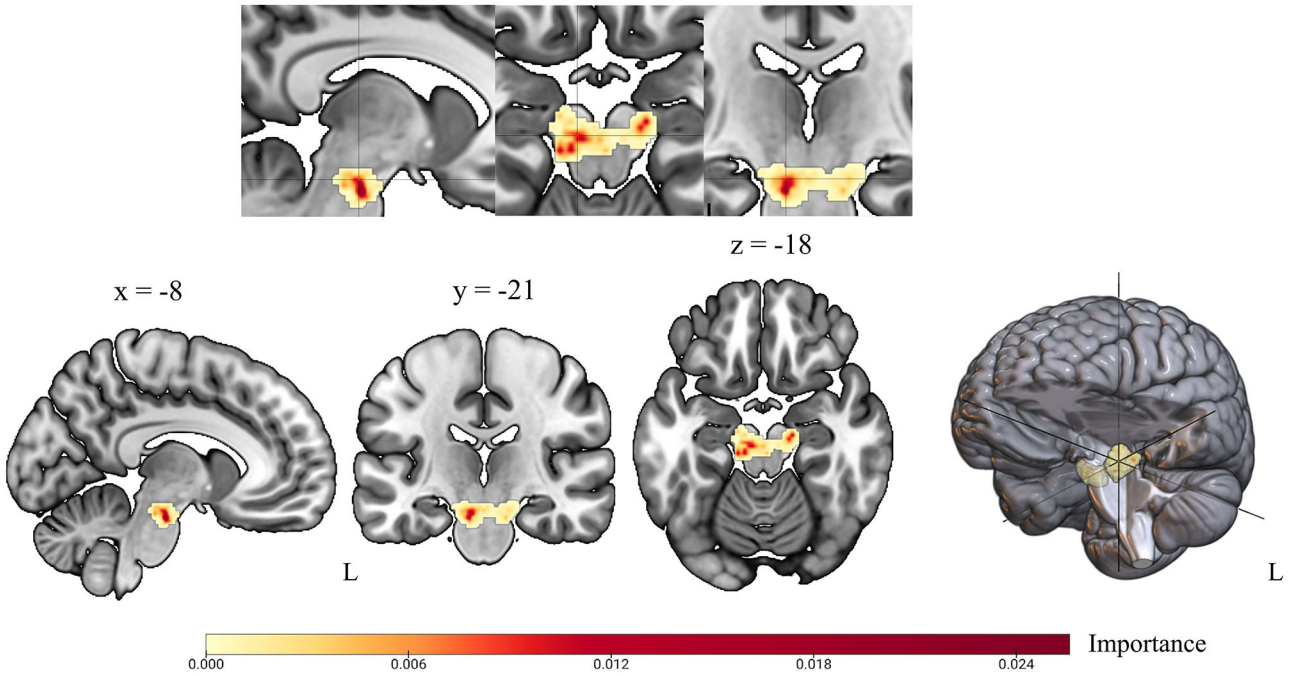


FIGURE 3
Voxel-wise feature importance for the midbrain region of interest. Feature importance revealed by random forest classifier is shown in color scaled 3D representation for every voxel in the midbrain region of interest. Yellow color indicates minor feature importance; red color indicates high voxel importance for class decision based on random forest classifier.

subjects as possible. The presence of unbalanced data and a rather small sample size warrants some caution on generalizable conclusions. In particular, unbalanced data may aid in more precise identification of PD patients compared to healthy control subjects' categorization since the individual model was likely trained on a higher number of patients compared to controls. Subsequent studies should therefore include larger sample sizes and equally sized groups. However, the present study reveals initial implications for the approach by applying an appropriate model for unbalanced data. In addition, appropriate techniques like conducting an ensemble of random forest analyses and evaluating the variability of model performance across runs, internal 10-fold cross validation and model evaluation based on respective validation data sets and prediction for one independent subject were taken into account. As a logical consequence of the preliminary work, however, the study provides initial indications, and an interesting proximity to neuropathology with accessibility on subject level, which could also be applicable to patients with prodromal disease in future studies.

A major limitation of the present study is the absence of testing the model on an external cohort, which would supplement the generalizability of the results. We conducted the analysis with high-resolution HRRT PET data to enable tracing back effects on smallest midbrain structures in terms of pathophysiological relevance. We have not tested our approach in an independent sample, as there is no large public dataset of high-resolution PET data. However, a more widespread availability of higher resolution scanners in the future and a multicenter initiative for collecting data may foster possibilities for an independent data set. In addition, future projects could focus on the comparability with lower resolution data as recent studies suggest that our approach might be feasible in non-high-resolution data that are more widely available. Furthermore, our implementation of supervised learning relied on the subjective evaluations of two independent clinical experts in movement disorders, which may not always reflect the ground truth, and should be supported by more objective diagnostic criteria as proposed by biological PD models, including molecular CSF markers, evidence of rapid eye movement sleep behavior disorder (RBD) and dopaminergic imaging, especially in prodromal stages.

4.2 Future perspectives

Similar to other studies using machine learning techniques, there is a question about scalability or applicability of this relatively simple measures in independent cohorts. Future studies could validate the approach presented here in early or prodromal stages of the disease, such as patients with RBD, as differences could be expected according to the longitudinally observed midbrain hypometabolism (Steidel et al., 2022). As recent studies highlight a pivotal role for evidence of nigrostriatal degeneration also in the pre-motor phase of the disease, our results may have direct implications for the emerging field of early diagnostics and identifying at-risk persons. The application of such kind of *in-vivo* accessible, objective biomarkers is of greatest interest in context of new therapeutic treatment strategies and paralleled by the development of disease-modifying agents. As longitudinal midbrain changes were demonstrated in mid-stage patients, future studies should verify if midbrain hypometabolism can be identified in prodromal stages like RBD-patients with high-resolution PET. Identifying prodromal biomarkers may be helpful for identifying

early disease stages, a crucial element for clinical trials of potential neuroprotective drugs, antibody studies or cell-based therapies.

5 Conclusion

Midbrain metabolism measured by [^{18}F]-FDG PET is a promising imaging tool for detecting PD-related midbrain degeneration on subject-level. Given its close relationship to PD pathophysiology and very high sensitivity, this approach can index midbrain degeneration and help to establish neurobiological staging systems, addressing the nigrostriatal system.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by local Ethics Committee University of Cologne, Faculty of Medicine. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

MR-J: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. GK: Conceptualization, Formal analysis, Investigation, Methodology, Software, Validation, Writing – review & editing. AG: Data curation, Investigation, Writing – review & editing. MT: Writing – review & editing. LT: Resources, Supervision, Writing – review & editing. AD: Resources, Writing – review & editing, Funding acquisition. CE: Conceptualization, Project administration, Resources, Validation, Writing – review & editing, Funding acquisition. DP: Resources, Writing – review & editing, Conceptualization, Project administration, Validation.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This study received funding by the German Research Association (DFG) in context of the Clinical Research Group 219 (KFO 219, EG350/1–1). Open access funding was provided by the Open Access Publishing Fund of Philipps-Universität Marburg.

Acknowledgments

We would especially like to thank our participants who made this study possible and our colleagues who helped with data acquisition. Our sincere thanks go to our colleague Marcus Belke for his help in the implementation of the analysis.

Conflict of interest

DP has received honoraria as a speaker at symposia sponsored by Boston Scientific Corp, Medtronic, AbbVie Inc., Zambon and Esteve Pharmaceuticals GmbH. He received payments as a consultant for Boston Scientific Corp and Bayer, and he received a scientific grant from Boston Scientific Corp for a project entitled: "Sensor-based optimisation of Deep Brain Stimulation settings in Parkinson's disease" (COMPARE-DBS). Finally, DP was reimbursed by Esteve Pharmaceuticals GmbH and Boston Scientific Corp for travel expenses to attend congresses. Between September 2021 and September 2023 LT received occasional payments as a consultant for Boston Scientific, he received honoraria as a speaker on symposia sponsored by Boston Scientific, AbbVIE, Novartis, Neuraxpharm, Teva, the Movement Disorders Society und DIAPLAN. The institution of LT, not he personally received funding by Boston Scientific, the German Research Foundation, the German Ministry of Education and Research, the Otto-Loewi-Foundation and the Deutsche Parkinson's Vereinigung. Neither LT nor any member of his family holds stocks, stock options, patents, or financial interests in any of the above mentioned companies or their competitors. LT serves as the president of the German Neurological Society without any payment or any income. MT is supported by funding from the German Center for DiabetesResearch as well as by the Deutsche Forschungsgemeinschaft DFG, German ResearchFoundation under Germany's Excellence Strategy. AD: Research support: Siemens Healthineers, Life Molecular Imaging, GE Healthcare, AVID Radiopharmaceuticals, SOFIE, Eisai, and Speaker Honorary/Advisory Boards: Siemens Healthineers, Sanofi, GE Healthcare, Biogen, Novo Nordisk, Invicro, and Novartis/AAA, Stock: Siemens Healthineers and Lantheus Holding, Patents: Patent pending for 18F-PSMA7 (PSMA PET imaging tracer). CE has received honoraria as speaker or consultant from AbbVie Inc., Stada Pharma Inc., Everpharma Inc., and Philyra.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fncom.2024.1328699/full#supplementary-material

References

Adler, C. H. (2014). Low clinical diagnostic accuracy of early vs. advanced Parkinson disease. *Neurology* 83, 406–412. doi: 10.1212/WNL.0000000000000641

Albin, R. L., Young, A. B., and Penney, J. B. (1989). The functional anatomy of basal ganglia disorders. *Trends Neurosci.* 12, 366–375. doi: 10.1016/0166-2236(89)90074-X

Braak, H., Bohl, J. R., Müller, C. M., Rüb, U., Vos, R. A. I., and Del Tredici, K. (2006a). Stanley Fahn Lecture 2005: the staging procedure for the inclusion body pathology associated with sporadic Parkinson's disease reconsidered. *Mov. Disord.* 21, 2042–2051. doi: 10.1002/mds.21065

Braak, H., Rüb, U., Schultz, C., and Del Tredici, K. (2006b). Vulnerability of cortical neurons to Alzheimer's and Parkinson's diseases. *J. Alzheimers Dis.* 9, 35–44. doi: 10.3233/JAD-2006-9S305

Brett, M., Anton, J. L., Valabregue, R., and Poline, J.-B. (2002). "Region of interest analysis using an SPM toolbox," in *Presented at the 8th International Conferance on Functional*.

Chahine, L. M., Merchant, K., Siderowf, A., Sherer, T., Tanner, C., Marek, K., et al. (2023). Proposal for a biologic staging system of Parkinson's disease. *J. Parkinsons Dis.* 13, 297–309. doi: 10.3233/JPD-225111

Collins, D. L., Zijdenbos, A. P., Baaré, W. F. C., and Evans, A. C. (1999). "ANIMAL+INSECT: improved cortical structure segmentation," in *Information Processing in Medical Imaging*, eds. G. Goos, J. Hartmanis, J. van Leeuwen, A. Kuba, M. Šáamal, and A. Todd-Pokropek (Berlin, Heidelberg: Springer Berlin Heidelberg), 210–223. doi: 10.1007/3-540-48714-X_16

Damier, P., Hirsch, E. C., Agid, Y., and Graybiel, A. M. (1999). The substantia nigra of the human brain. II. Patterns of loss of dopamine-containing neurons in Parkinson's disease. *Brain* 122, 1437–1448. doi: 10.1093/brain/122.8.1437

Della Rosa, P. A., Cerami, C., Gallivanone, F., Prestia, A., Caroli, A., Castiglioni, I., et al. (2014). A standardized 18F-FDG-PET template for spatial normalization in statistical parametric mapping of dementia. *Neuroinformatics* 12, 575–593. doi: 10.1007/s12021-014-9235-4

DeLong, M. R. (1990). Primate models of movement disorders of basal ganglia origin. *Trends Neurosci.* 13, 281–285. doi: 10.1016/0166-2236(90)90110-V

Emre, M., Aarsland, D., Brown, R., Burn, D. J., Duyckaerts, C., Mizuno, Y., et al. (2007). Clinical diagnostic criteria for dementia associated with Parkinson's disease. *Mov. Disord.* 22, 1689–707. doi: 10.1002/mds.21507

Fahn, S., Marsden, C. D., Goldstein, M., Calne, D. B. (1987). Recent developments in Parkinsons disease. *Movement Disor.* 2, 153–163.

Fearnley, J. M., and Lees, A. J. (1991). Ageing and Parkinson's disease: substantia nigra regional selectivity. *Brain* 114, 2283–2301. doi: 10.1093/brain/114.5.2283

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). Mini-mental state. *J. Psychiatr. Res.* 12, 189–198. doi: 10.1016/0022-3956(75)90026-6

Glaab, E., Trezzi, J.-P., Greuel, A., Jäger, C., Hodak, Z., Drzezga, A., et al. (2019). Integrative analysis of blood metabolomics and PET brain neuroimaging data for Parkinson's disease. *Neurobiol. Dis.* 124, 555–562. doi: 10.1016/j.nbd.2019.01.003

Greuel, A., Trezzi, J.-P., Glaab, E., Ruppert, M. C., Maier, F., Jäger, C., et al. (2020). GBA variants in Parkinson's disease: clinical, metabolomic, and multimodal neuroimaging phenotypes. *Mov. Disord.* 35, 2201–2210. doi: 10.1002/mds.28225

Hoehn, M. M., and Yahr, M. D. (1967). Parkinsonism: onset, progression and mortality. *Neurology* 17, 427–442. doi: 10.1212/WNL.17.5.427

Höglinger, G. U., Adler, C. H., Berg, D., Klein, C., Outeiro, T. F., Poewe, W., et al. (2023). Towards a biological definition of Parkinson's disease. *Preprints*. doi: 10.20944/preprints202304.0108.v1

Kish, S. J., Shannak, K., and Hornykiewicz, O. (1988). Uneven pattern of dopamine loss in the striatum of patients with idiopathic Parkinson's disease. *New England J. Med.* 318, 876–880. doi: 10.1056/NEJM198804073181402

Langston, J. W., Widner, H., Goetz, C. G., Brooks, D., Fahn, S., Freeman, T., et al. (1992). Core assessment program for intracerebral transplantations (CAPIT). *Mov. Disord.* 7, 2–13. doi: 10.1002/mds.870070103

Lau, L. M. L., and de Breteler, M. M. B. (2006). Epidemiology of Parkinson's disease. *Lancet Neurol.* 5, 525–535. doi: 10.1016/S1474-4422(06)70471-9

Peng, S., Spetsieris, P. G., Eidelberg, D., and Ma, Y. (2020). Radiomics and supervised machine learning in the diagnosis of parkinsonism with FDG PET: promises and challenges. *Ann. Transl. Med.* 8:808. doi: 10.21037/atm.2020.04.33

Postuma, R. B., Berg, D., Stern, M., Poewe, W., Olanow, C. W., Oertel, W., et al. (2015). MDS clinical diagnostic criteria for Parkinson's disease. *Mov. Disord.* 30, 1591–1601. doi: 10.1002/mds.26424

Ruppert, M. C., Greuel, A., Freigang, J., Tahmasian, M., Maier, F., Hammes, J., et al. (2021). The default mode network and cognition in Parkinson's disease: a multimodal resting-state network approach. *Hum. Brain Mapp.* 42, 2623–2641. doi: 10.1002/hbm.25393

Ruppert, M. C., Greuel, A., Tahmasian, M., Schwartz, F., Stürmer, S., Maier, F., et al. (2020). Network degeneration in Parkinson's disease: multimodal imaging of nigro-striato-cortical dysfunction. *Brain.* 143, 944–959. doi: 10.1093/brain/awaa019

Schröter, N., Blazhenets, G., Frings, L., Jost, W. H., Weiller, C., Rijntjes, M., et al. (2022). Nigral glucose metabolism as a diagnostic marker of neurodegenerative parkinsonian syndromes. *NPJ Parkinsons Dis.* 8:123. doi: 10.1038/s41531-022-00392-x

Segovia, F., Górriz, J. M., Ramírez, J., Martínez-Murcia, F. J., Levin, J., Schuberth, M., et al. (2017a). Multivariate analysis of 18F-DMFP PET data to assist the diagnosis of parkinsonism. *Front. Neuroinform.* 11:23. doi: 10.3389/fninf.2017.00023

Segovia, F., Górriz, J. M., Ramírez, J., Martínez-Murcia, F. J., and Salas-Gonzalez, D. (2017b). Preprocessing of 18F-DMFP-PET data based on hidden markov random fields and the gaussian distribution. *Front. Aging Neurosci.* 9:326. doi: 10.3389/fnagi.2017.00326

Segovia, F., Illán, I. A., Górriz, J. M., Ramírez, J., Rominger, A., and Levin, J. (2015). Distinguishing Parkinson's disease from atypical parkinsonian syndromes using PET data and a computer system based on support vector machines and Bayesian networks. *Front. Comput. Neurosci.* 9:137. doi: 10.3389/fncom.2015.00137

Seibyl, J., Russell, D., Jennings, D., and Marek, K. (2012). Neuroimaging over the course of Parkinson's disease: from early detection of the at-risk patient to improving pharmacotherapy of later-stage disease. *Semin. Nucl. Med.* 42, 406–414. doi: 10.1053/j.semnuclmed.2012.06.003

Shen, T., Jiang, J., Lin, W., Ge, J., Wu, P., Zhou, Y., et al. (2019). Use of overlapping group LASSO sparse deep belief network to discriminate parkinson's disease and normal control. *Front. Neurosci.* 13:396. doi: 10.3389/fnins.2019.00396

Steidel, K., Ruppert, M. C., Greuel, A., Tahmasian, M., Maier, F., Hammes, J., et al. (2022). Longitudinal trimodal imaging of midbrain-associated network degeneration in Parkinson disease. *NPJ Parkinsons Dis.* 8:79. doi: 10.1038/s41531-022-00341-8

Tomlinson, C. L., Stowe, R., Patel, S., Rick, C., Gray, R., and Clarke, C. E. (2010). Systematic review of levodopa dose equivalency reporting in Parkinson's disease. *Mov. Disord.* 25, 2649–2653. doi: 10.1002/mds.23429

Wu, Y., Jiang, J.-H., Chen, L., Lu, J.-Y., Ge, J.-J., Liu, F.-T., et al. (2019). Use of radiomic features and support vector machine to distinguish Parkinson's disease cases from normal controls. *Ann. Transl. Med.* 7:773. doi: 10.21037/atm.2019.11.26

Zhao, Y., Cumming, P., Rominger, A., Zuo, C., Shi, K., Wu, P., et al. (2019). "A 3D deep residual convolutional neural network for differential diagnosis of parkinsonian syndromes on 18F-FDG PET images," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* 3531–3534. doi: 10.1109/EMBC.2019.8856747

frontiers | Frontiers in Computational Neuroscience

TYPE Technology and Code
PUBLISHED 22 March 2024
DOI 10.3389/fncom.2024.1357607

OPEN ACCESS

EDITED BY
Roberto Maffulli,
Italian Institute of Technology (IIT), Italy

REVIEWED BY
Heidi Nag,
Frambu, Norway
Sadiq Hussain,
Dibrugarh University, India

*CORRESPONDENCE
Daniel Palacios-Alonso
✉ daniel.palacios@urjc.es

RECEIVED 18 December 2023
ACCEPTED 23 February 2024
PUBLISHED 22 March 2024

CITATION
Fernández-Ruiz R, Núñez-Vidal E, Hidalgo-delaguía I, Garayzábal-Heinze E, Álvarez-Marquina A, Martínez-Olalla R and Palacios-Alonso D (2024) Identification of Smith–Magenis syndrome cases through an experimental evaluation of machine learning methods.
Front. Comput. Neurosci. 18:1357607.
doi: 10.3389/fncom.2024.1357607

Identification of Smith–Magenis syndrome cases through an experimental evaluation of machine learning methods

Raúl Fernández-Ruiz[1], Esther Núñez-Vidal[1], Irene Hidalgo-delaguía[2], Elena Garayzábal-Heinze[3], Agustín Álvarez-Marquina[4], Rafael Martínez-Olalla[4] and Daniel Palacios-Alonso[1,4]*

[1]Escuela Técnica Superior de Ingeniería Informática, Universidad Rey Juan Carlos, Madrid, Spain, [2]Departament of Spanish Language and Theory of Literature, Universidad Complutense de Madrid, Madrid, Spain, [3]Departament of Linguistics, Universidad Autónoma de Madrid, Madrid, Spain, [4]Center for Biomedical Technology, Universidad Politécnica de Madrid, Madrid, Spain

This research work introduces a novel, nonintrusive method for the automatic identification of Smith–Magenis syndrome, traditionally studied through genetic markers. The method utilizes cepstral peak prominence and various machine learning techniques, relying on a single metric computed by the research group. The performance of these techniques is evaluated across two case studies, each employing a unique data preprocessing approach. A proprietary data "windowing" technique is also developed to derive a more representative dataset. To address class imbalance in the dataset, the synthetic minority oversampling technique (SMOTE) is applied for data augmentation. The application of these preprocessing techniques has yielded promising results from a limited initial dataset. The study concludes that the k-nearest neighbors and linear discriminant analysis perform best, and that cepstral peak prominence is a promising measure for identifying Smith–Magenis syndrome.

KEYWORDS

Smith–Magenis syndrome, machine learning, cepstral peak prominence, acoustics, children

1 Introduction

Over time, artificial intelligence (AI) has experienced substantial growth in a variety of scientific areas and disciplines (Górriz et al., 2020, 2023). In the medical field, AI has been used for disease diagnosis and treatment (Rother et al., 2015; Shen et al., 2017; Jia et al., 2018; Li et al., 2019; Zhang et al., 2019; Spiga et al., 2020), as well as for new drug research, since, in scientific research, AI accelerates data analysis and complex phenomena monitoring (Cifci and Hussain, 2018; Firouzi et al., 2018). The versatility and transformative potential of AI offers new possibilities in disease diagnosis. The origins of AI date back to the 1950s, with the development of the first neural network (machine learning), although its roots can be traced even further back in time, considering previous approaches such as Bayesian statistics or Markov chains, which share similar concepts. In the case of Parkinson's disease, the authors of Ali et al. (2019) worked on phonation in combination with ML. The results were applicable to

other diseases that, due to their low incidence in the population, are understudied and, consequently, underdiagnosed.

Patients face considerable challenges with dealing with underdiagnosed pathologies. The lack of early detection and limited information deprives them of timely, pathology-specific care, which is especially important for young patients. The use of AI techniques for early disease detection is an ongoing challenge. In this study, the focus is on determining the discriminatory as well as pathological characteristics of young patients' voices. Acoustic phonation studies provide relevant speaker information that can be used to detect diseases such as Alzheimer's dementia, Parkinson's, and amyotrophic lateral sclerosis, among others, based on the biomechanical uniqueness of each individual. Such uniqueness is evident in the EWA-DB dataset, which focuses on Slovak speakers with Alzheimer's and Parkinson's diseases (Rusko et al., 2023), and a dataset that focuses on Spanish native speakers with Parkinson's disease (Orozco-Arroyave et al., 2014), as well as recent acoustic studies on Alzheimer's (Cai et al., 2023; Zolnoori et al., 2023) and Parkinson's (Warule et al., 2023) diseases. In the 2021 study by Lee (2021), two types of neural network models were developed for dysphonia detection: a Feedforward Neural Network (FNN) and a Convolutional Neural Network (CNN). These models were designed to utilize Mel Frequency Cepstral Coefficients (MFCCs) for the detection process.

The determined laryngeal biomechanics, elastin deficiency in Williams syndrome (WS) or excess laryngeal tension in the case of Smith–Magenis syndrome (SMS) (Watts et al., 2008; Moore and Thibeault, 2012; Hidalgo-De la Guía et al., 2021b) discriminate these syndromes from others caused by neurological pathologies based on genetics (Antonell et al., 2006; Albertini et al., 2010; Hidalgo et al., 2018; Jeffery et al., 2018; Hidalgo-De la Guía et al., 2021a). Specifically, the voice profile of an SMS patient is determined by excess laryngeal and acute tension f0. These patients may also have a certain degree of dysphonia, which is observed in both children and adults. Likewise, there are studies that suggest that certain syndromes present characteristic alterations in the voice that give rise to specific vocal phenotypes (Edelman et al., 2007; Brendal et al., 2017; Linders et al., 2023).

SMS is a genetic disease that affects neurological development from the embryonic stage, specifically due to the alteration of the RAI1 gene, which is considered responsible for most of the clinical abnormalities observed in SMS individuals (Slager et al., 2003; Vlangos et al., 2003). Given its prevalence, i.e., 1:15,000–25,000 births (Greenberg et al., 1996; Elsea and Girirajan, 2008; Girirajan et al., 2009), SMS is considered a rare disease and, therefore, is underdetected.

It is more common to approach the problem of rare disease detection from areas other than genetics, where the fundamental focus has been on characterization. ML techniques have recently been implemented in rare disease research, including SMS. Bozhilova et al. (2023) identified different profiles of autism characteristics in genetic syndromes associated with some intellectual disability. SMS was among the 13 syndromes studied. The *Social Communication Questionnaire* was used to train a support vector machine (SVM) that achieved an overall precision of 55%. The main limitations of this work were that only social communication skill metrics were used and imbalanced sample sizes across groups. One of the main results seems to indicate that autistic individuals with genetic syndromes have different characteristics than those without any genetic syndrome. In Frassineti et al. (2021) different ML models were proposed to allow the automatic identification of four different diseases, including SMS. They made recordings of subjects and extracted 34 acoustic characteristics with Praat and 24 with BioVoice. The *cepstral peak prominence* (CPP) was not among the extracted characteristics. After the results achieved by BioVoice for SMS (true positive rate of 55.6% and false-negative rate of 44.4%), the authors suggested that the vowel /a/ is not sufficient for the definition of phenotypes. In an extension of their previous work, the same authors (Calà et al., 2023) incorporated the vowels /a/, /I/, and /u/, and introduced a new control group of normative individuals. Utilizing BioVoice, they extracted 77 acoustic features, excluding CPP, and organized the subjects into three distinct groups: pediatric subjects (age < 12), adult females, and adult males. Each group was treated independently, with a unique Machine Learning model generated for each. The results, obtained through a 10-fold cross validation, are presented as mean accuracy along with the standard deviation. The pediatric group achieved an accuracy of 87 ± 9%, adult women achieved 77 ± 19%, and men achieved 84 ± 17%. However, the outcomes appear inconclusive due to the high variability in measures such as precision, recall, and f-score.

This work compares different Machine Learning techniques for the detection of SMS in young people using audio samples, from which only the CPP is computed and extracted. In addition, a novel windowing method is proposed to improve the performance of the models. In addition, the SMOTE technique is used, aiming outcomes in precision rates above 85%. This approach proposes a non-invasive, low-cost, and rapid detection method with only one acoustic parameter, which contrasts with methods based on genetic techniques.

Unfortunately, it is difficult to compare medical research works, which used genetic techniques, with non-invasive SMS detection. Likewise, mathematical and computational approaches to this syndrome use acoustical features such as formants, shimmer, and jitter, among others. However, this study case aims to open the exploration of new ways to identify SMS individuals. The fact to use only one feature (CPP) allows faster models with lower computational performance. Therefore, the ultimate goal is to detect the syndrome early using this single feature.

This article is organized as follows. In the following section, the methods and materials are explained, the dataset structure and the "window" method are highlighted, and the ML methods used are briefly explained from a theoretical perspective. In Section 3, the results are included, and the model training and validation, as well as the approach and results of the case studies, are detailed. Next, in Section 4, the obtained results are discussed, and finally, the conclusions and future lines of work are proposed.

Abbreviations: AI, artificial intelligence; CPP, cepstral peak prominence; FISH, fluorescent *in situ* hybridization; FFT, fast Fourier transform; GMM, Gaussian mixture model; IFFT, inverse fast Fourier transform; KNN, k-nearest neighbors; LDA, linear discriminant analysis; LOO, leave one out; MFCC, mel frequency cepstral coefficients; ML, machine learning; RF, random forest; SMOTE, synthetic minority oversampling technique; SMS, Smith–Magenis syndrome; SVM, support vector machine; WS, Williams syndrome.

2 Materials and methods

2.1 Cepstral peak prominence

This research work is based on the use of the CPP as a discriminant measure for the identification of SMS (nonnormotypic) individuals compared to a control group of normotypic individuals. The CPP is an acoustic parameter that allows determining the degree of periodicity of a voice, showing the prominence of a cepstral peak that varies according to the periodicity of phonation. The more pronounced the peak is, the more harmonic a voice (Hidalgo-De la Guía et al., 2021b).

In the past decade, it has been found that the CPP presents a strong correlation with the degree of voice dysphonia (Peterson et al., 2013; Brinca et al., 2014). In fact, higher correlations were found between the CPP, and dysphonic voices compared to those of typical distortion parameters (Moers et al., 2012). Currently, the CPP is considered one of the best acoustic parameters for estimating the degree of vocal pathology. In addition, it has been found that the CPP in SMS individuals is low, which could be related to a possible relationship between the syndrome and laryngeal biomechanics (Hidalgo-De la Guía et al., 2021b).

In SMS, a dysphonic voice is one of the characteristics with the highest rate of appearance (Linders et al., 2023), and to achieve dysphonic voice detection in this study, the CPP is used. The CPP is calculated as follows.

1. The signal is segmented into overlapping fragments (1,024 samples 87.5% overlap). Each fragment is multiplied by a Hamming window function, and the fast Fourier transform (FFT) is calculated. Based on this calculated signal, the absolute value is found, and its logarithm is calculated. Finally, the inverse fast Fourier transform (IFFT) is performed on the previous result, and the real part is obtained. Thus, a set of frames is created in the cepstral domain.

$$c = real\left(IFFT\left(\log\left(abs\left(FFT\left(x.\times w\right)\right)\right)\right)\right)$$

where c is the cepstrum vector, x the input signal vector, w is a vector with a Hamming window function and the operation × represents the sample-to-sample product of both vectors.

2. A smoothing filter (smoothing in the cepstral direction) is applied to each of the frames obtained in the cepstral domain. This filter is applied to eliminate spurious signal values while preserving the true cepstral peaks, thus avoiding cepstral peak detection errors.

$$c_f[n] = \sum_{i=0}^{l-1} a_i c[n-i]$$

where *cf.* is the value of the smoothed cepstrum, a_i are the coefficients of the filter, and $l = 7$ is the length of the filter in samples.

3. The cepstrum is then limited between the quefrency values corresponding to the minimum (22 samples) and maximum (400 samples) fundamental periods expected for the range of vocal frequencies of the study population.

4. The maximum value of the previous signal (cepstral peak) is calculated, and the CPP is obtained as the difference between this maximum and the average of the rest of the signal.

$$CPP[i] = \max_{T_{\min}}^{T_{\max}}\left(c_{fi}\right) - \frac{\sum_{j=T_{\min}}^{T_{\max}} c_{fi}[j] - \max_{T_{\min}}^{T_{\max}}\left(c_{fi}\right)}{T_{\max} - T_{\min}}$$

5. A vector is formed with the CPP values thus obtained (CPP[n]), which is smoothed by a filter with a 56 ms window (smoothing in the temporal direction). This smoothing operation reduces the noise of the signal obtained while preserving large variations in the CPP value, which can be present in dysphonic voices.

$$CPP_f[n] = \sum_{i=0}^{m-1} a_i CPP[n-i]$$

with a filter length $m = 7$ for a displacement of 128 samples and 16,000 Hz of sampling frequency, and where ai are the coefficients of the filter (following a hamming window function), and CPPf the smoothed CPP.

2.2 Dataset

Most rare disease databases, such as those for SMS, are private, and accessing these databases is difficult. In the specific case of databases in Spanish, the Orphanet website (Orphanet, 2023) offers genetic biobank searches. Such searches were carried out, and three results were obtained: Basque Biobank, CIBERER Biobank, and the National Biobank for Rare Diseases (BioNer). However, two of the three results do not have information about SMS, and the one that does contain genetic information.

The difficulty of obtaining this type of data is well known. Given that the number of subjects suffering from these syndromes is small and heterogeneous, the datasets are strongly unbalanced. Consequently, this situation requires synthetic data augmentation methods to be applied. These techniques have been widely used in the field of image processing since the appearance of convolutional neural networks (CNNs) in 2012 (Shorten and Khoshgoftaar, 2019). Likewise, to process data such as those mentioned above, oversampling techniques such as the *synthetic minority oversampling technique* (SMOTE) and its variants are used. As described by Alabi et al. (2020), these techniques can be used to increase of amount of data in early tongue cancer detection. In Joloudari et al. (2023), the effectiveness of different solutions to data imbalance in Deep Neural Networks and CNNs is verified. The best result is obtained by combining SMOTE with a CNN plus a normalization process between both stages, achieving an accuracy of 99.08% across 24 imbalanced datasets.

In this study, the dataset contains voice quality information from normotypic and nonnormotypic individuals for comparison. To create this dataset, we worked with a total of 22 individuals between the ages

of 5 and 33 who belong to the Smith–Magenis Spain Association (ASME), comprising 20% of the Spanish population diagnosed with this syndrome. The diagnosis of all the individuals with SMS was obtained by means of the fluorescent *in situ* hybridization (FISH) technique. Samples were collected from subjects through recordings in which they had to hold the vowel /a/ for a few seconds (minimum 500 ms of phonation). The recording quality was guaranteed by ruling out comorbidity of associated vocal pathology, such as vocal fold nodules or any other additional vocal problem. Likewise, the recording context was addressed as follows: the rooms were completely silent (some soundproofed), only of the researcher and the diagnosed person were in the room, and a cardioid lapel microphone was used. From all the audio, the CPP information, an acoustic voice quality measure and one of the best dysphonia metrics (vocal timbre alteration), as described by Heman-Ackah et al. (2003), was extracted.

In this study, a subset of these data was used, consisting of 12 individuals SMS, all of whom were between the ages of 5 and 12 years. These individuals were used because we wanted to verify the possibility of developing a system that allows early disease identification, since a late diagnosis leads to a worse quality of life. The group of 12 individuals with SMS is made up of two subgroups: a group of young children aged 5 to 7 years and another group of older children aged 8 to 12 years. Both subgroups had 3 boys and 3 girls.

To complete the dataset, 12 recordings of participants with typical development were added. Sample collection from normotypic individuals was the same as that used for SMS individuals, and the same age distribution as that of the SMS individuals was followed.

The dataset in the study contains 2,685 CPP values extracted from audio from the 24 participants (12 normotypic and 12 nonnormotypic participants). The number of CPP values per participant varied in relation to the number of voice samples obtained and their duration. Each entry in the dataset has the following fields defined: subject identifier, sex, age, CPP value, as well as whether the participant suffers from SMS and whether they belong to the "younger" or "older" group.

A descriptive analysis of the CPP stored in the database was prepared as presented in Figure 1, where the *X*-axis represents the CPP values, and the *Y*-axis represents the data divided by sex. The orange boxplots represent the SMS group, and the blue boxplots represent the normative group. It is observed that the SMS group has much lower CPP values than those of the normative group. Likewise, it can be observed that the range of values for normative boys and girls is very similar. However, the range of values for SMS boys is slightly more dispersed than that of SMS girls. Finally, in Figure 1, it is observed that the *boxplot* of SMS girls is slightly larger, and the whiskers are somewhat longer than those of normotypic girls.

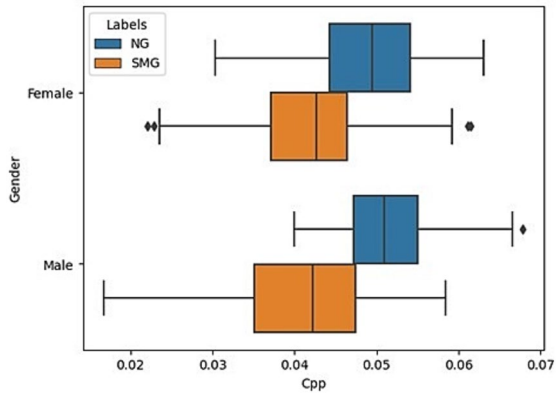


FIGURE 1
Representation of CPP values by sex, comparing normative vs. nonnormative groups.

Given the importance of age and sex and to improve the explainability of the results, the aforementioned information was segmented by "young children" (5–7 years) and "older children" (8–12 years). The results are reflected in Figure 2. From the generated histograms, it is observed that in the group of girls between 8 and 12 years old and that of boys between 5 and 7, there is a greater differentiation in the CPP values between normotypic and SMS individuals. However, in the other two groups (girls between 5 and 7 years old and boys between 8 and 12 years old), there is a greater overlap between the data of both groups. Specifically, the overlap is greater in girls between 5 and 7 years old than in the group of boys between 8 and 12 years old.

It is important to point out some of the potential research gaps in this research work. A larger number of individuals with SMS could be enriching and it could avoid lead to biases by gender, age, or other characteristics. The second issue is the lack of exploration of different alternatives to SMOTE. There are different variants of this technique and other oversampling methods that could be implemented and could lead to better solutions. Finally, other ML methods could also be searched. All four methods used in this research work have a multitude of variants that may improve the performance of the baseline method. Regarding the problem of the number of individuals, as previously mentioned, it has been decided to use a subset of the data as a first approach due to the number of patients who suffer from this syndrome.

2.3 Preprocessing and data augmentation

When working with machine learning models, the data must have adequate structure that guarantees correct training. It should be noted that group the information by speaker does not require that all individuals have the same number of samples (the number of voice recordings). It is also unlikely that the recordings will have the same duration. However, to directly apply one or more of the extracted features, the problem of comparing patterns of different sizes must be solved. Therefore, a proprietary "window" algorithm was developed, and to explain its operation, Figure 3 is used as a reference.

Although there are several subgroups that belong to the same person, they should not be treated independently within the dataset. Consequently, they should be assigned exclusively to either the validation set or the training set, but never simultaneously. Though CPP is not an efficient acoustic measure for speaker identification, compared to others such as Mel Frequency Cepstral Coefficients (MFCC) (Ayvaz et al., 2022), it is preferred to avoid mixing subgroups of the same person in the validation and training sets to pre-vent possible data leakage. Table 1 illustrates the result of the windowing process by means of a dataframe, where each row represents a sample in the dataset. With this process, a usable data structure was achieved to train the different ML models, as detailed in the following section.

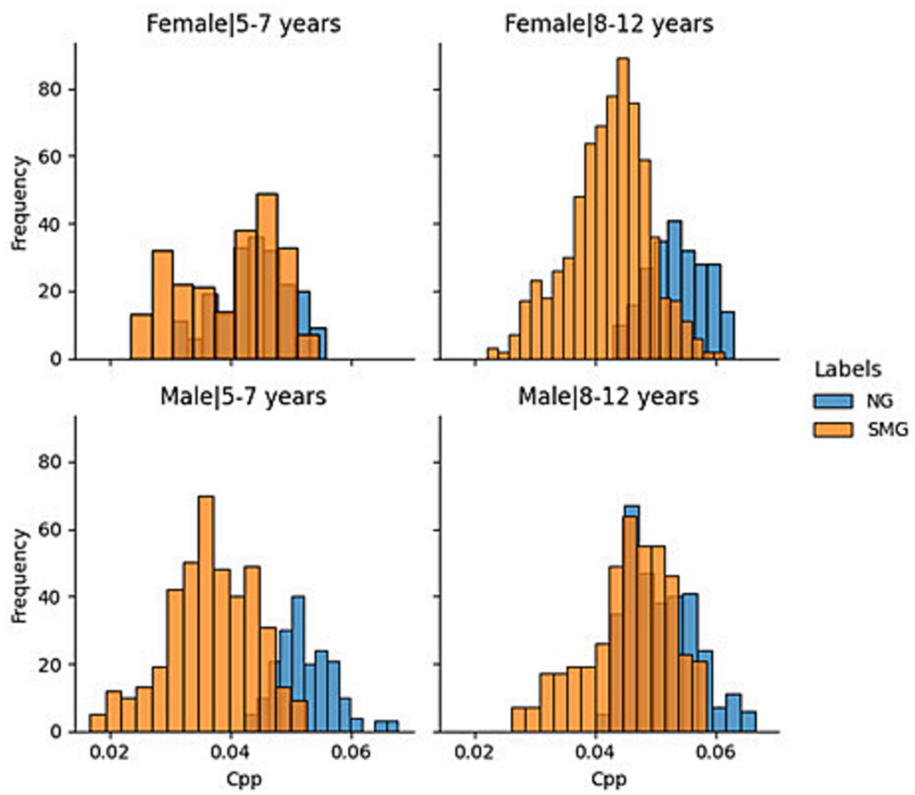


FIGURE 2
Normotypic vs. nonnormotypic CPP decomposition by sex and group.

In addition to the problem indicated above, there is a second problem, i.e., the imbalance between the classes to be predicted (246 entries from SMS individuals and 100 entries from normotypic individuals). This fact directly affects the performance of models that tend to overfit. To solve this problem, various solutions have been explored, e.g., assigning a higher weight to the minority class during the training or eliminating majority class samples. Finally, it was decided to use the SMOTE technique (Chawla et al., 2002), an oversampling technique based on the creation of synthetic examples of the minority class. With SMOTE, new samples are introduced along the segments that join the k nearest neighbors of the minority class. The number of k neighbors selected depends on the number of samples generated samples required. As the number of samples increases, the number of neighbors employed decreases. The great advantage of this technique is that it allows the generation of synthetic samples instead of resorting to oversampling, where samples of the minority class are reintroduced into the dataset, which tends to lead to overfitting.

2.4 ML techniques

In this work, both supervised and unsupervised methods were considered to compare the different techniques and create combined models. Among unsupervised methods, the Gaussian mixture model (GMM) (Rasmussen, 1999) and K-means clustering (Sinaga and Yang, 2020) were used. In addition, the following supervised methods were used: SVM, random forest (RF), linear discriminant analysis (LDA) and k-nearest neighbors (KNN).

Unsupervised methods were not included in this work as they do not offer results that contribute any new research knowledge. These techniques generated clusters based on the sex and age of the individuals, ignoring the CPP. Therefore, the experiment was repeated after eliminating these two variables. However, the clusters did not provide any new information.

Because supervised techniques are well known, only a brief description of the methods is given. The SVM (Jakkula, 2006) builds hyperplanes that allow an optimal separation of the data, and the power of this method resides in the kernel trick, allowing data transfer to spaces of greater dimensionality in an optimal manner. Depending on the kernel used, the shape of the decision boundary varies; in Figure 4, the influence of the different types of kernels is observed.

The RF (Pachange et al., 2015) is an assembly method, where multiple decision trees are combined to generate predictions. This method is based on building decision trees, where data are divided using the problem variables, applying some criterion that evaluates and maximizes the gain of information. LDA searches for a linear combination of the characteristics that generates the greatest variance between classes and minimizes it within each class (Izenman, 2008). KNN allows for the prediction of a class of data based on its k closest neighbors (Uddin et al., 2022). The way in which the influence of each neighbor is determined in the final prediction can vary according to the technique used. For example, if the weight of each neighbor in the final decision is "uniform," all neighbors have an equal influence on

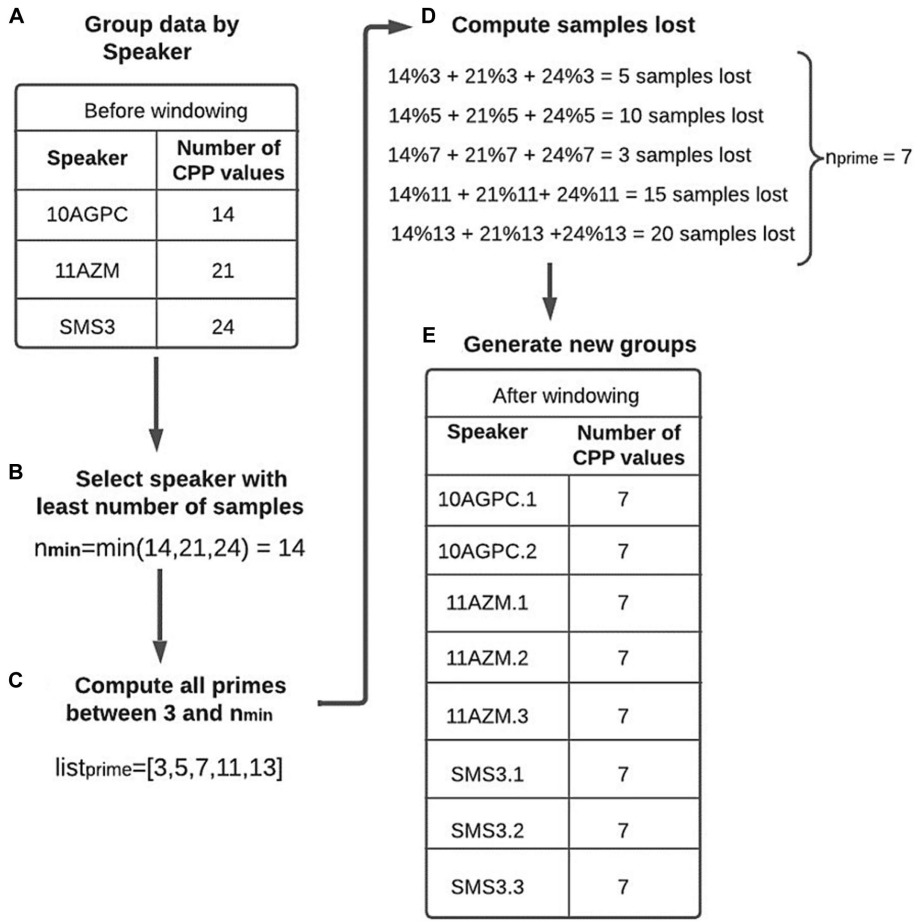


FIGURE 3
Windowing example. **(A)** All the CPP values stored in the database are grouped for each speaker. In the illustrated example, the first speaker has 14 CPP values, the second has 21 and the third has 24. **(B)** The speaker with the lowest number of samples (14 in this case) is identified ($n_{\min}$). **(C)** All prime numbers between 3 and $n_{\min}$ (14) are stored in a list ($list_{prime}$). **(D)** For each value stored in $list_{prime}$, the number of samples that would be lost when dividing the sampling into groups of that size is calculated. This calculation is equivalent to determining the modulus of the group size between that value. Suppose that in the example described above that a value of three is used. Since the first speaker has 14 samples, it is possible to generate four new groups of size three and lose two samples; for the second speaker, no samples would be lost, and for the third speaker, three samples would be lost. Therefore, if a size of three is used to generate the new groups, a total of five samples would be lost. For this reason, an n_{prime} value is sought that minimizes the number of lost samples. **(E)** The samples grouped by speaker are divided into groups of size n_{prime}. Each subgroup generated from the same individual has a number added to the end of the identifier to distinguish them. In the case of the above example, the number of samples of speaker SMS3 is 24, and n_{prime} is equal to 7. Therefore, three new groups of size 7 are obtained (SMS3.1, SMS3.2, and SMS3.3), and the remaining three samples are lost.

TABLE 1 Dataframe generated after windowing when $n_{prime} = 7$.

Name	CPP1	CPP2	CPP3	CPP4	CPP5	CPP6	CPP7	Target	Sex	Group
10APG.1	0.0488	0.0502	0.5050	0.0501	0.0494	0.0481	0.0476	N	Female	Older
10APG.2	0.0490	0.0508	0.0467	0.0483	0.0457	0.0458	0.0466	N	Female	Older
...	...	...	...	...	...	...	...	...	...	...
SMS3.3	0.0477	0.0475	0.480	0.0511	0.055	0.058	0.058	SMS	Male	Young

the vote; on the other hand, if each neighbor is "weighted," the closest neighbors will have a greater influence on the final decision.

2.5 Wilcoxon rank sum test

The Wilcoxon rank-sum test, also called the Mann–Whitney U test, is a powerful tool for comparing two sets of data without relying on specific assumptions about their distribution (unlike some other tests). It works by ranking the observations in each set instead of using their raw values. This makes it especially useful when the data might be skewed or non-normally distributed.

The goal of the Wilcoxon rank-sum test is to assess whether the medians of two populations differ significantly. This is particularly helpful when the precise shape of the data distribution is unknown.

To calculate the test statistic, the formula is shown as follows:

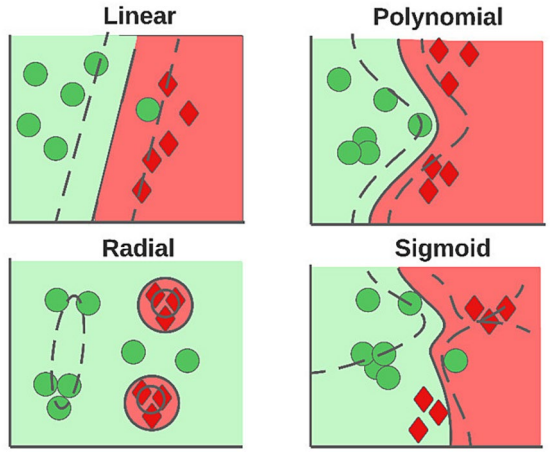


FIGURE 4
Hyperplanes generated according to the kernel used.

$$U = n_1 m_2 + \sum R_1 - (n_1)(n_2 + 1)/2$$

Where:

- U: The test statistic
- n_1: Size of the first sample
- n_2: Size of the second sample
- ΣR_1: Sum of the ranks in the first sample
- m_2: Median of the second sample

3 Results

3.1 Training and validation

The consistency of this study lies in its data, as well as the techniques and methods used. Therefore, it was decided to apply the methodical procedure described in Figure 5 to the data. This procedure is summarized in four fundamental phases: windowing, Leave One Out, SMOTE, ML methods.

1. Windowing: Each sample is composed of seven CPP values, sex, and group. Therefore, $n_{prime} = 7$.

2. Leave One Out (LOO): It is used to implement a training and validation model that ensured that different subgroups of the same person do not end up in different datasets. To do this, all subgroups of the same person are extracted to be used as a validation set, while the rest of the samples are used in the training phase. This process is repeated for each of the 24 people in the study.

3. SMOTE: It is used to generate new synthetic samples of the minority class (normotypic). The objective is to avoid creating biased models that tend to over-identify the dominant class (SMS). Although the number of SMS and normative individuals in the training set is always 11 versus 12, depending on which group is used for validation, the number of SMS subgroups (248) is higher than that of normative subgroups (131). It should be noted that this technique is only applied to the training set. The SMOTE technique is not suitable for the validation set. In such a way that the two groups are separated and do not mix and therefore data leakage is avoided.

4. ML methods: Once the training and validation sets are obtained, the different ML models are trained. Previously, exhaustive tests were carried out with different hyperparameters to identify the most effective combinations. It should be noted that, for each validation set, not only one but ten iterations are carried out. An augmented training set is generated in each iteration by using the SMOTE technique. Then, the performance of the used model is evaluated on the validation set. This process is repeated ten times, generating new training sets with SMOTE and training a new model in each iteration. The aim is to obtain a robust and accurate estimate of the model's performance over iterations. This process consists of a Leave One Out Cross Validation.

To statistically compare the performance of the different models on each individual, the following process will be followed: the 10 values obtained in the LOO for each subject in each method will be recorded. Then, all the results of each method for the same individual will be compared one by one using the Wilcoxon Rank Sum Test (Boslaugh, 2012), in order to obtain the p-values of and thus determine the statistical significance of the methods. The results are reflected in Tables 2, 3.

3.2 Results

Two different case studies were established in order to evaluate the behavior and quality of the predictions in the models.

1. The first case study (CE1) applies the windowing process but does not use SMOTE, resulting in an unbalanced training set in favor of the SMS class. Each training/validation sample contains seven CPP values used to predict whether it belongs to the SMS or normative class.

2. The second case study (CE2) involves the data passing through the windowing process and subsequently applying SMOTE to the training set. The data maintains the same structure as in the previous case.

Each case relates to the four ML techniques proposed in Section 2.4. Each figure (Figures 6–13) groups individuals by their age, sex, and study case, corresponding to the subgroups identified in Section 2.2. Each figure is divided into tables which share the same column structure: the first identifies the speaker, the second shows the number of samples per person obtained after the windowing process. The next ten columns represent the values obtained using leave-one-out (LOO) cross-validation, with the samples treated as the validation group, these ten values reflect the repetitions of the process. The last column is the average value of the ten iterations plus the standard deviation. Every table displays three normative (blue) and the non-normative (orange) individuals. In each iteration of the Leave-One-Out (LOO) cross-validation, all samples belonging to a single individual are consistently used as the validation set. This means we exclude all samples from a particular subject and test the model on them in each iteration.

Importantly, the tables associated with CE2 (Figures 10–13) exhibit higher standard deviations and different results on the score columns compared to those of CE1 (Figures 6–9). This issue occurs

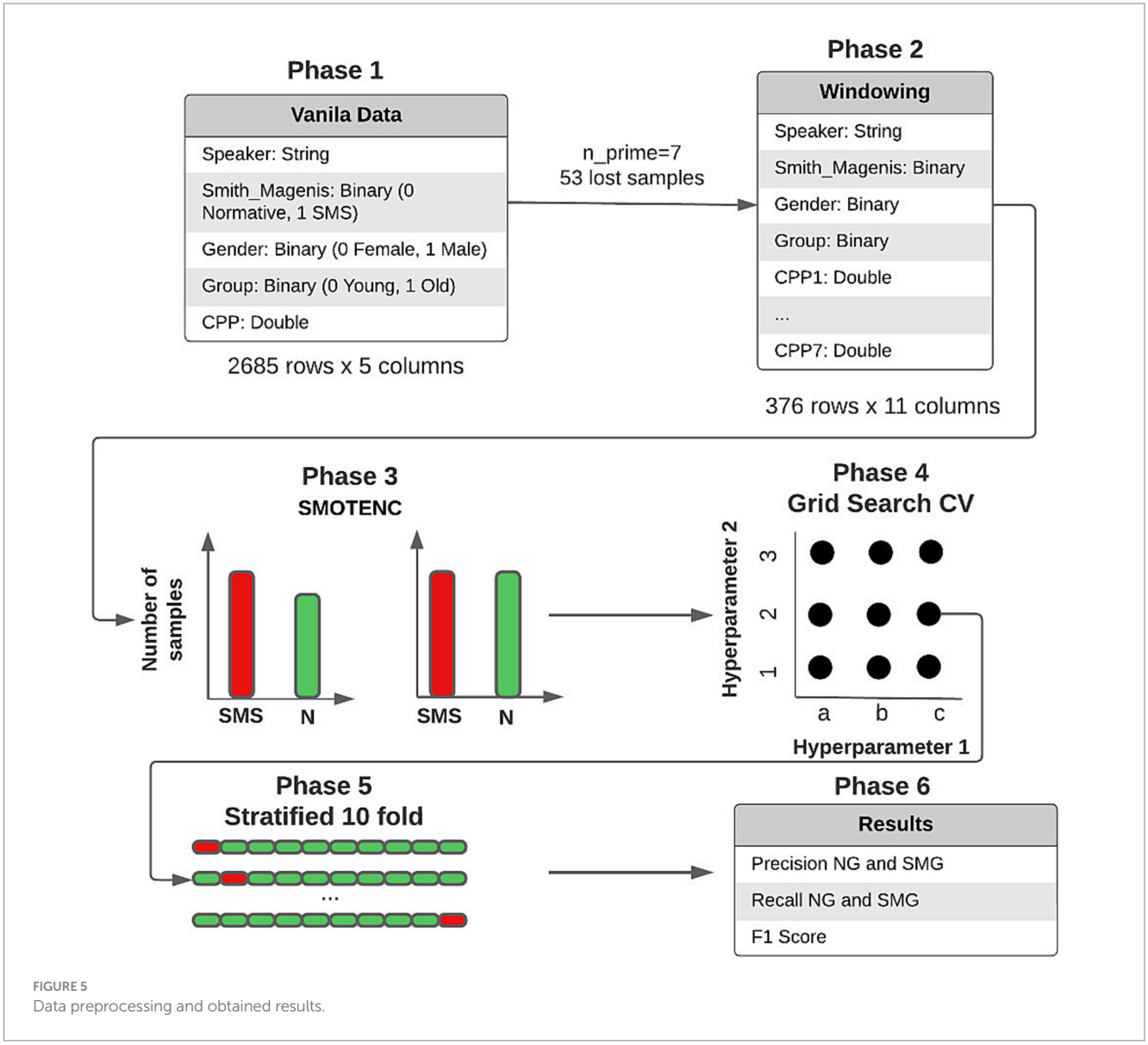


FIGURE 5
Data preprocessing and obtained results.

because, in CE2, each iteration augments the training set with SMOTE, generating new synthetic data, making each training set different from the others. Furthermore, significant variation between iterations for the same subject is possible due to the limited size of the individual validation sets (i.e., 15 samples). If the algorithm fails or hits two samples of the available data during a specific iteration, the resulting value for that iteration can fluctuate significantly across different runs.

3.2.1 Case study 1

The results of CE1 are elaborated in Figures 6–9. It is noteworthy that the subgroups of Female Old and Male Young (Figures 6A–9A, 6C–9C) do not exhibit exceptionally low detection rates. However, a stark contrast is observed in the Female Young subgroup (Figures 6B–9B), where the three normative individuals display significantly lower results compared to the SMS group. In the final subgroup, Male Old (Figures 6D–9D), both normative and SMS individuals demonstrate low detection rates.

When individuals are evaluated independently, it is observed that several normative subjects, such as 10AGPC, 11OADS, 517A, 612A, 637A, and 842O, exhibit low precision rates across various methods. Some of these subjects achieve low rates on the order of 0.1%. Within the SMS group, only SMS7 and SMS9 display significantly low detection rates. SMS11 also has a low rate, albeit higher than the previous two speakers. These results align with the tendencies of a biased model, which tends to over-identify the majority groups. In this scenario, the dominant class (SMS) demonstrates better detection than the minority class (normative).

3.2.2 Case study 2

Figures 10–13 depict the outputs of CE2. In Figures 10A–13A, there is a noticeable enhancement in the detection of 10AGPC compared to the previous case, notwithstanding with a minor decline for SMS11. In the Female Young subgroup (Figures 10B–13B), detection rates for subjects 517A and 637A have increased, but performance for patient SMS06 has decreased. In the Male Old

TABLE 2 Summary and comparison of the four ML methods, providing average and pairwise precision rates using the Wilcoxon Rank Sum Test for CE1.

Average accuracy					Comparison Wilcoxon Test (p-value)					
RF	KNN	SVM	LDA	Speaker	RF_vs_SVM	RF_vs_KNN	RF_vs_LDA	SVM_vs_KNN	SVM_vs_LDA	KNN_vs_LDA
59.23%	53.85%	46.15%	61.54%	10AGPC	0.002	0.011	0.149	0.002	0.002	0.002
99.23%	92.31%	100.00%	100.00%	11AAZM	1	0.003	1	0.002	NA	0.002
23.08%	15.38%	0.00%	0.00%	11OADS	0.002	0.005	0.002	0.002	NA	0.002
88.33%	91.67%	75.00%	75.00%	511O	0.002	0.072	0.002	0.002	NA	0.002
16.67%	16.67%	16.67%	16.67%	517A	NA	NA	NA	NA	NA	NA
5.83%	0.00%	0.00%	0.00%	612A	0.011	0.011	0.011	NA	NA	NA
87.14%	71.43%	71.43%	71.43%	618O	0.002	0.002	0.002	NA	NA	NA
49.00%	40.00%	40.00%	40.00%	637A	0.008	0.008	0.008	NA	NA	NA
85.71%	85.71%	85.71%	85.71%	743O	NA	NA	NA	NA	NA	NA
87.06%	94.12%	82.35%	94.12%	819O	0.018	0.012	0.012	0.002	0.002	NA
67.33%	66.67%	46.67%	66.67%	842O	0.002	0.783	0.783	0.002	0.002	NA
95.00%	100.00%	100.00%	100.00%	12109A	0.149	0.149	0.149	NA	NA	NA
99.64%	100.00%	100.00%	100.00%	SMS1	1.000	1.000	1.000	NA	NA	NA
100.00%	100.00%	100.00%	100.00%	SMS2	NA	NA	NA	NA	NA	NA
86.15%	92.31%	92.31%	92.31%	SMS3	0.006	0.006	0.006	NA	NA	NA
100.00%	100.00%	100.00%	100.00%	SMS4	NA	NA	NA	NA	NA	NA
77.50%	87.50%	100.00%	100.00%	SMS5	0.002	0.006	0.002	0.002	NA	0.002
68.46%	76.92%	84.62%	84.62%	SMS6	0.002	0.010	0.002	0.002	NA	0.002
29.23%	46.15%	38.46%	38.46%	SMS7	0.007	0.002	0.007	0.002	NA	0.002
90.61%	87.88%	93.94%	93.94%	SMS8	0.002	0.003	0.002	0.002	NA	0.002
36.15%	15.38%	7.69%	0.00%	SMS9	0.002	0.002	0.002	0.002	0.002	0.002
100.00%	100.00%	100.00%	100.00%	SMS10	NA	NA	NA	NA	NA	NA
62.88%	65.38%	65.38%	61.54%	SMS11	0.026	0.026	0.104	NA	0.002	0.002
91.76%	94.12%	100.00%	100.00%	SMS12	0.002	0.006	0.002	0.002	NA	0.002
71.1%	70.6%	68.6%	70.1%							

subgroup (Figures 10D–13D), all normative subjects exhibit improvements in their detection rates, despite a minor decrease for subjects SMS07 and SMS08. Lastly, the Male Young subgroup (Figures 10C–13C) mirrors the Male Old, with improved detection for all normative individuals and a slight decrease for SMS.

Highlighting some individual cases, it is significant to note that subjects 10AGPC and 842O from the normative set have seen substantial improvements in their detection compared to the previous case. The individual 11OADS depicts a considerable increase in SVM detection from 0 to 0.815 (Figures 8D vs. 12D) and an increase from 0 to 0.577 in LDA (Figures 9D vs. 13D). For 637A (Female Young), there is a global enhancement in detection across methods, with both SVM and LDA (Figures 12B, 13B) yielding favorable results. However, no significant improvement is observed for subjects 517A and 612A (Female Young). Conversely, the SMS group results indicate a marked decrease in performance, especially for individuals SMS6 (Female Young) and SMS11 (Female Old), which achieved identification rates below 0.5. SMS7 (Male Old) and SMS9 (Male Old) present identification rates comparable to the previous case. Lastly, the SMOTE technique boosts the precision rates of the minority class, albeit at a slight detriment to the majority class.

4 Discussion

In this work, we propose the development of ML models that allow for the identification of SMS versus normotypic individuals. One clinical feature of the SMS pathology is voice hoarseness (Elsea and Girirajan, 2008), as described in previous studies (Hidalgo-De la Guía et al., 2021b), it has been demonstrated that by utilizing the CPP values of SMS and normotypic individuals, it is possible to create divisions into highly differentiated subgroups. This differentiation is primarily due to the hoarseness present in individuals with this genetic pathology. These types of studies are necessary to improve early disease detection. Currently, the average SMS diagnosis age is approximately seven years (Hidalgo-De la Guía et al., 2021b), leading to problems for these patients. Problem arises because SMS requires specific therapies that, when implemented late, cause different kinds of delays. As presented in this research work, the voice is a versatile, inexpensive, and minimally invasive medium that helps to discriminate possible pathologies (Jeffery et al., 2018; Lee, 2021; Calà et al., 2023).

The initial data were not suitable for ML model training. The main problem was sample imbalance between groups. Two techniques were

TABLE 3 Summary and comparison of the four ML methods, providing average and pairwise precision rates using the Wilcoxon Rank Sum Test for CE2.

Average accuracy					Comparison Wilcoxon test (*p*-value)					
RF	KNN	SVM	LDA	Speaker	RF_vs_SVM	RF_vs_KNN	RF_vs_LDA	SVM_vs_KNN	SVM_vs_LDA	KNN_vs_LDA
72.30%	90.00%	99.23%	89.23%	10AGPC	0.002	0.002	0.002	0.008	0.008	0.679
100.00%	100.00%	100.00%	100.00%	11AAZM	NA	NA	NA	NA	NA	NA
33.80%	33.85%	81.54%	57.69%	11OADS	0.002	1.000	0.002	0.002	0.002	0.002
93.30%	100.00%	100.00%	95.83%	511O	0.006	0.006	0.299	NA	0.037	0.037
21.70%	55.00%	50.00%	35.00%	517A	0.002	0.002	0.015	0.149	0.003	0.002
9.20%	10.00%	8.33%	0.00%	612A	0.414	0.679	0.006	0.186	0.002	0.007
92.90%	100.00%	100.00%	100.00%	618O	0.037	0.037	0.037	NA	NA	NA
56.00%	65.00%	70.00%	88.00%	637A	0.002	0.058	0.002	0.240	0.002	0.002
85.71%	91.43%	100.00%	100.00%	743O	0.002	0.072	0.002	0.020	NA	0.020
84.12%	100.00%	100.00%	100.00%	819O	0.002	0.002	0.002	NA	NA	NA
73.33%	74.00%	98.00%	96.00%	842O	0.002	1.000	0.002	0.002	0.149	0.002
100.00%	100.00%	100.00%	100.00%	12109A	NA	NA	NA	NA	NA	NA
99.64%	98.21%	92.86%	98.93%	SMS1	0.002	0.129	0.424	0.002	0.002	0.484
92.94%	91.18%	94.12%	94.12%	SMS2	0.186	0.322	0.186	0.037	NA	0.037
86.15%	85.38%	76.92%	73.08%	SMS3	0.002	0.408	0.002	0.012	0.037	0.008
100.00%	100.00%	100.00%	100.00%	SMS4	NA	NA	NA	NA	NA	NA
75.00%	70.00%	75.00%	97.50%	SMS5	1.000	0.129	0.002	0.072	0.002	0.002
43.85%	37.69%	27.69%	30.00%	SMS6	0.002	0.098	0.002	0.034	0.149	0.033
14.62%	21.54%	7.69%	23.08%	SMS7	0.048	0.090	0.026	0.002	0.002	0.186
84.24%	77.58%	67.27%	86.67%	SMS8	0.002	0.006	0.229	0.009	0.002	0.002
30.77%	13.08%	0.00%	0.00%	SMS9	0.002	0.002	0.002	0.002	NA	0.002
100.00%	100.00%	100.00%	100.00%	SMS10	NA	NA	NA	NA	NA	NA
56.15%	46.15%	34.62%	44.23%	SMS11	0.002	0.009	0.009	0.002	0.002	0.322
88.53%	88.82%	80.59%	98.24%	SMS12	0.002	1.000	0.002	0.002	0.002	0.002
70.6%	72.9%	73.5%	75.3%							

proposed to solve this problem. The first technique is CPP sample "windowing," a novel approach. In Section 2.3, it was explained that "windowing" consists of grouping the samples by speaker and making new subgroups of the same size to solve the sample imbalance problem. The second technique is the application of SMOTE, with which new synthetic samples of the minority class are generated until a balance between the two classes is achieved. The authors maintain that, with the combination of the "windowing" and SMOTE methods, the dataset is improved. To demonstrate how the yields of the models vary according to the applied techniques, two different case studies were proposed.

The LOO technique was implemented to prevent the inclusion of subgroups of the same person in the validation and training sets, avoiding the risk of data leakage. This technique is especially beneficial in small datasets because it allows the use of all n-1 available data for training. It should be noted that training involves the 23 individuals present in the dataset, while the remaining person is reserved for validation. This validation and training process is iterated ten times for each speaker. This iterative approach contributes to obtaining robust results, reducing the possibility of achieving biased or circumstance-influenced performances. The different models tend to over-identify the dominant group (SMS) in CE1. In contrast, in CE2, the SMOTE technique was implemented in the training dataset to address the class imbalance. It should be highlighted that the application of SMOTE was limited to the training set to prevent possible data leakage.

This approach increased the identification of the normative group and led to an overall improved performance but reduced slightly the identification of the SMS speakers. To evaluate the ML techniques against each other, it has been decided to give the arithmetic median obtained in the SMS and normative classes, as it is not affected by outsider high or low performances in certain individuals. Firstly, SVM offered the worst results, especially in CE1, since it was necessary to use models with a hyperparameter configuration that tends to overfit the model due to its inability to detect the normative class. This led to labeling all results as SMS, obtaining an average of 0.59 and 0.97 for the normative and SMS classes. However, in CE2, a model that does not depend on hyperparameters is obtained, with a median of 0.99 for normative and 0.75 for SMS. In this second case study, its high detection rate in the normative group stands out. Individual 11OADS is far superior to the rest of the methods. Nonetheless, it is not able to achieve such good generalization in the SMS group.

Random Forest

A CE1 + Female Old

Name	N Sample	Score1	Score2	Score3	Score4	Score5	Score6	Score7	Score8	Score9	Score10	avg.+std
10AGPC	13	0,615	0,615	0,615	0,615	0,538	0,538	0,615	0,538	0,615	0,615	0,592 +/- 0,037
11AAZM	13	1,000	1,000	1,000	1,000	0,923	1,000	1,000	1,000	1,000	1,000	0,992 +/- 0,024
12109A	6	1,000	1,000	1,000	1,000	1,000	0,833	0,833	1,000	0,833	1,000	0,95 +/- 0,081
SMS10	13	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1 +/- 0
SMS11	52	0,635	0,654	0,654	0,615	0,615	0,615	0,615	0,596	0,673	0,615	0,629 +/- 0,024
SMS12	34	0,912	0,912	0,912	0,912	0,912	0,912	0,912	0,941	0,941	0,912	0,918 +/- 0,012

B CE1 + Female Young

Name	N Sample	Score1	Score2	Score3	Score4	Score5	Score6	Score7	Score8	Score9	Score10	avg.+std
517A	6	0,167	0,167	0,167	0,167	0,167	0,167	0,167	0,167	0,167	0,167	0,167 +/- 0
612A	12	0,083	0,083	0,083	0,083	0,083	0,083	0,000	0,083	0,000	0,000	0,058 +/- 0,04
637A	10	0,500	0,400	0,500	0,500	0,400	0,600	0,500	0,500	0,500	0,500	0,49 +/- 0,057
SMS4	11	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1 +/- 0
SMS5	8	0,875	0,750	0,750	0,750	0,875	0,750	0,750	0,750	0,750	0,750	0,775 +/- 0,053
SMS6	13	0,692	0,769	0,538	0,692	0,769	0,692	0,692	0,692	0,615	0,692	0,685 +/- 0,067

C CE1 + Male Young

Name	N Sample	Score1	Score2	Score3	Score4	Score5	Score6	Score7	Score8	Score9	Score10	avg.+std
511O	12	0,917	0,917	0,917	0,833	0,917	0,833	0,833	0,833	0,917	0,917	0,883 +/- 0,043
618O	7	0,857	0,857	0,857	0,857	1,000	0,857	0,857	0,857	0,857	0,857	0,871 +/- 0,045
743O	7	0,857	0,857	0,857	0,857	0,857	0,857	0,857	0,857	0,857	0,857	0,857 +/- 0
SMS1	28	0,964	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,996 +/- 0,011
SMS2	17	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1 +/- 0
SMS3	13	0,846	0,846	0,923	0,846	0,846	0,923	0,846	0,846	0,846	0,846	0,862 +/- 0,032

D CE1 + Male Old

Name	N Sample	Score1	Score2	Score3	Score4	Score5	Score6	Score7	Score8	Score9	Score10	avg.+std
11OADS	13	0,231	0,231	0,154	0,231	0,231	0,231	0,231	0,231	0,231	0,308	0,231 +/- 0,036
819O	17	0,824	0,941	0,882	0,765	0,882	0,941	0,882	0,882	0,824	0,882	0,871 +/- 0,054
842O	15	0,733	0,667	0,600	0,667	0,600	0,733	0,667	0,733	0,667	0,667	0,673 +/- 0,049
SMS7	13	0,308	0,231	0,308	0,385	0,231	0,231	0,308	0,308	0,308	0,308	0,292 +/- 0,049
SMS8	33	0,909	0,909	0,909	0,909	0,909	0,909	0,909	0,879	0,909	0,909	0,906 +/- 0,01
SMS9	13	0,385	0,385	0,385	0,308	0,385	0,385	0,308	0,385	0,308	0,385	0,362 +/- 0,037

FIGURE 6
Summary of the results for the CE1, using RF. **(A)** Detailed performance for the old female subgroup. **(B)** Detailed performance for the young female subgroup. **(C)** Detailed performance for the young male subgroup. **(D)** Detailed performance for the old male subgroup.

K-Nearest Neighbors

A CE1 + Female Old

Name	N Sample	Score1	Score2	Score3	Score4	Score5	Score6	Score7	Score8	Score9	Score10	avg.+std
10AGPC	13	0,538	0,538	0,538	0,538	0,538	0,538	0,538	0,538	0,538	0,538	0,538 +/- 0
11AAZM	13	0,923	0,923	0,923	0,923	0,923	0,923	0,923	0,923	0,923	0,923	0,923 +/- 0
12109A	6	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1 +/- 0
SMS10	13	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1 +/- 0
SMS11	52	0,654	0,654	0,654	0,654	0,654	0,654	0,654	0,654	0,654	0,654	0,654 +/- 0
SMS12	34	0,941	0,941	0,941	0,941	0,941	0,941	0,941	0,941	0,941	0,941	0,941 +/- 0

B CE1 + Female Young

Name	N Sample	Score1	Score2	Score3	Score4	Score5	Score6	Score7	Score8	Score9	Score10	avg.+std
517A	6	0,167	0,167	0,167	0,167	0,167	0,167	0,167	0,167	0,167	0,167	0,167 +/- 0
612A	12	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0 +/- 0
637A	10	0,400	0,400	0,400	0,400	0,400	0,400	0,400	0,400	0,400	0,400	0,4 +/- 0
SMS4	11	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1 +/- 0
SMS5	8	0,875	0,875	0,875	0,875	0,875	0,875	0,875	0,875	0,875	0,875	0,875 +/- 0
SMS6	13	0,769	0,769	0,769	0,769	0,769	0,769	0,769	0,769	0,769	0,769	0,769 +/- 0

C CE1 + Male Young

Name	N Sample	Score1	Score2	Score3	Score4	Score5	Score6	Score7	Score8	Score9	Score10	avg.+std
511O	12	0,917	0,917	0,917	0,917	0,917	0,917	0,917	0,917	0,917	0,917	0,917 +/- 0
618O	7	0,714	0,714	0,714	0,714	0,714	0,714	0,714	0,714	0,714	0,714	0,714 +/- 0
743O	7	0,857	0,857	0,857	0,857	0,857	0,857	0,857	0,857	0,857	0,857	0,857 +/- 0
SMS1	28	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1 +/- 0
SMS2	17	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1 +/- 0
SMS3	13	0,923	0,923	0,923	0,923	0,923	0,923	0,923	0,923	0,923	0,923	0,923 +/- 0

D CE1 + Male Old

Name	N Sample	Score1	Score2	Score3	Score4	Score5	Score6	Score7	Score8	Score9	Score10	avg.+std
11OADS	13	0,154	0,154	0,154	0,154	0,154	0,154	0,154	0,154	0,154	0,154	0,154 +/- 0
819O	17	0,941	0,941	0,941	0,941	0,941	0,941	0,941	0,941	0,941	0,941	0,941 +/- 0
842O	15	0,667	0,667	0,667	0,667	0,667	0,667	0,667	0,667	0,667	0,667	0,667 +/- 0
SMS7	13	0,462	0,462	0,462	0,462	0,462	0,462	0,462	0,462	0,462	0,462	0,462 +/- 0
SMS8	33	0,879	0,879	0,879	0,879	0,879	0,879	0,879	0,879	0,879	0,879	0,879 +/- 0
SMS9	13	0,154	0,154	0,154	0,154	0,154	0,154	0,154	0,154	0,154	0,154	0,154 +/- 0

FIGURE 7
Summary of the results for the CE1, using KNN. **(A)** Detailed performance for the old female subgroup. **(B)** Detailed performance for the young female subgroup. **(C)** Detailed performance for the young male subgroup. **(D)** Detailed performance for the old male subgroup.

The second model discussed in this study is RF. Acceptable performance is achieved with medians of 0.765 for normative and 0.884 in SMS at CE1. However, practically identical performance is observed to the previous case in CE2. Medians are between 0.787 and 0.852 for normative and SMS. It is crucial to say that the use of SMOTE does not always guarantee an improvement in model performance. In fact, it can become a problem by generating noise in situations of high dimensionality. Nevertheless, it does not rule out the possibility that the combination of the SMOTE technique with RF can improve results with other datasets. For example, in Abdar et al. (2019) four variants of DTs are proposed to predict coronary artery disease. The article proposes a multi-filtering approach based on supervised and unsupervised methods to modify the weights of the attributes, leading to a 20–30% improvement in the methods.

The two final models analyzed in this study exhibit relevant high performances. Firstly, the KNN's performance experiences a significant improvement: from medians of 0.69 and 0.9 in CE1 to 0.90 and 0.81 for normative and SMS in CE2. This improvement can be attributed to the data arrangement, as shown in Figure 2, where three out of four clusters present adequate separation. Consequently, this technique is better than the others because if the closest samples are selected then higher recognition rate are obtained. Finally, the

Support Vector Machine

A CE1 + Female Old

Name	N Sample	Score1	Score2	Score3	Score4	Score5	Score6	Score7	Score8	Score9	Score10	avg.+std
10AGPC	13	0,462	0,462	0,462	0,462	0,462	0,462	0,462	0,462	0,462	0,462	0,462 +/- 0
11AAZM	13	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1 +/- 0
12109A	6	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1 +/- 0
SMS10	13	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1 +/- 0
SMS11	52	0,654	0,654	0,654	0,654	0,654	0,654	0,654	0,654	0,654	0,654	0,654 +/- 0
SMS12	34	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1 +/- 0

B CE1 + Female Young

Name	N Sample	Score1	Score2	Score3	Score4	Score5	Score6	Score7	Score8	Score9	Score10	avg.+std
517A	6	0,167	0,167	0,167	0,167	0,167	0,167	0,167	0,167	0,167	0,167	0,167 +/- 0
612A	12	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0 +/- 0
637A	10	0,400	0,400	0,400	0,400	0,400	0,400	0,400	0,400	0,400	0,400	0,4 +/- 0
SMS4	11	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1 +/- 0
SMS5	8	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1 +/- 0
SMS6	13	0,846	0,846	0,846	0,846	0,846	0,846	0,846	0,846	0,846	0,846	0,846 +/- 0

C CE1 + Male Young

Name	N Sample	Score1	Score2	Score3	Score4	Score5	Score6	Score7	Score8	Score9	Score10	avg.+std
511O	12	0,750	0,750	0,750	0,750	0,750	0,750	0,750	0,750	0,750	0,750	0,75 +/- 0
618O	7	0,714	0,714	0,714	0,714	0,714	0,714	0,714	0,714	0,714	0,714	0,714 +/- 0
743O	7	0,857	0,857	0,857	0,857	0,857	0,857	0,857	0,857	0,857	0,857	0,857 +/- 0
SMS1	28	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1 +/- 0
SMS2	17	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1 +/- 0
SMS3	13	0,923	0,923	0,923	0,923	0,923	0,923	0,923	0,923	0,923	0,923	0,923 +/- 0

D CE1 + Male Old

Name	N Sample	Score1	Score2	Score3	Score4	Score5	Score6	Score7	Score8	Score9	Score10	avg.+std
11OADS	13	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0 +/- 0
819O	17	0,824	0,824	0,824	0,824	0,824	0,824	0,824	0,824	0,824	0,824	0,824 +/- 0
842O	15	0,467	0,467	0,467	0,467	0,467	0,467	0,467	0,467	0,467	0,467	0,467 +/- 0
SMS7	13	0,385	0,385	0,385	0,385	0,385	0,385	0,385	0,385	0,385	0,385	0,385 +/- 0
SMS8	33	0,939	0,939	0,939	0,939	0,939	0,939	0,939	0,939	0,939	0,939	0,939 +/- 0
SMS9	13	0,077	0,077	0,077	0,077	0,077	0,077	0,077	0,077	0,077	0,077	0,077 +/- 0

FIGURE 8
Summary of the results for the CE1, using SVM. **(A)** Detailed performance for the old female subgroup. **(B)** Detailed performance for the young female subgroup. **(C)** Detailed performance for the young male subgroup. **(D)** Detailed performance for the old male subgroup.

Linear Discriminant Analysis

A CE1 + Female Old

Name	N Sample	Score1	Score2	Score3	Score4	Score5	Score6	Score7	Score8	Score9	Score10	avg.+std
10AGPC	13	0,615	0,615	0,615	0,615	0,615	0,615	0,615	0,615	0,615	0,615	0,615 +/- 0
11AAZM	13	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1 +/- 0
12109A	6	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1 +/- 0
SMS10	13	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1 +/- 0
SMS11	52	0,615	0,615	0,615	0,615	0,615	0,615	0,615	0,615	0,615	0,615	0,615 +/- 0
SMS12	34	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1 +/- 0

B CE1 + Female Young

Name	N Sample	Score1	Score2	Score3	Score4	Score5	Score6	Score7	Score8	Score9	Score10	avg.+std
517A	6	0,167	0,167	0,167	0,167	0,167	0,167	0,167	0,167	0,167	0,167	0,167 +/- 0
612A	12	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0 +/- 0
637A	10	0,400	0,400	0,400	0,400	0,400	0,400	0,400	0,400	0,400	0,400	0,4 +/- 0
SMS4	11	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1 +/- 0
SMS5	8	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1 +/- 0
SMS6	13	0,846	0,846	0,846	0,846	0,846	0,846	0,846	0,846	0,846	0,846	0,846 +/- 0

C CE1 + Male Young

Name	N Sample	Score1	Score2	Score3	Score4	Score5	Score6	Score7	Score8	Score9	Score10	avg.+std
511O	12	0,750	0,750	0,750	0,750	0,750	0,750	0,750	0,750	0,750	0,750	0,75 +/- 0
618O	7	0,714	0,714	0,714	0,714	0,714	0,714	0,714	0,714	0,714	0,714	0,714 +/- 0
743O	7	0,857	0,857	0,857	0,857	0,857	0,857	0,857	0,857	0,857	0,857	0,857 +/- 0
SMS1	28	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1 +/- 0
SMS2	17	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1 +/- 0
SMS3	13	0,923	0,923	0,923	0,923	0,923	0,923	0,923	0,923	0,923	0,923	0,923 +/- 0

D CE1 + Male Old

Name	N Sample	Score1	Score2	Score3	Score4	Score5	Score6	Score7	Score8	Score9	Score10	avg.+std
11OADS	13	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0 +/- 0
819O	17	0,941	0,941	0,941	0,941	0,941	0,941	0,941	0,941	0,941	0,941	0,941 +/- 0
842O	15	0,667	0,667	0,667	0,667	0,667	0,667	0,667	0,667	0,667	0,667	0,667 +/- 0
SMS7	13	0,385	0,385	0,385	0,385	0,385	0,385	0,385	0,385	0,385	0,385	0,385 +/- 0
SMS8	33	0,939	0,939	0,939	0,939	0,939	0,939	0,939	0,939	0,939	0,939	0,939 +/- 0
SMS9	13	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0 +/- 0

FIGURE 9
Summary of the results for the CE1, using LDA. **(A)** Detailed performance for the old female subgroup. **(B)** Detailed performance for the young female subgroup. **(C)** Detailed performance for the young male subgroup. **(D)** Detailed performance for the old male subgroup.

model that yields the best results is LDA, with medians of 0.690 and 0.970 for CE1 in normative and SMS, respectively. It is accomplished medians between 0.95 and 0.90 in CE2, making it the model with the most outstanding results throughout the research work.

Tables 2, 3 present a statistical comparison using the Wilcoxon Rank Sum test to evaluate the performance of the four employed ML methods which present the following structure. Each table is divided into three concepts. On the left side, the authors detail the accuracy rates for every ML method (RF, KNN, SVM and LDA) for each subject. The next column provides the speaker identifier. Finally, on the right-hand side, the authors detail the comparisons, contrasting the results obtained in the ten iterations (e.g., RF_{score1} … $RF_{Score10}$) of each method against the ten iterations (e.g., LDA_{score1} … $LDA_{Score10}$) of another method for the same subject. The last six columns display p-values from the Wilcoxon test. A *p*-value less than or equal to 0.05 indicates statistically significant differences in accuracy rates between methods, leading to rejection of the null hypothesis that they are equal. The table occasionally shows "Not Applicable (NA)" values. This occurs when the Wilcoxon test cannot calculate a p-value because the distance between all elements of the two input methods is zero. Such scenarios mostly arise when both methods achieve 100% or 0% accuracy (particularly in Table 3) but can also occur with other values. It is likely due to the relatively small dataset size (6–13 samples per subject), which increases the chance of different models achieving identical performance.

Upon comparing the two Tables 2, 3, a disparity is observed in the number of NA values. Table 2 records 59 NA values (29 in normotypic group and 30 in non-normotypic group). Indeed, the Table 3 shows

Random Forest

A CE2 + Female Old

Name	N Sample	Score1	Score2	Score3	Score4	Score5	Score6	Score7	Score8	Score9	Score10	avg.+std
10AGPC	13	0,846	0,692	0,692	0,692	0,769	0,769	0,692	0,692	0,615	0,769	0,723 +/- 0,065
11AAZM	13	1	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1 +/- 0
12109A	6	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1 +/- 0
SMS10	13	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1 +/- 0
SMS11	52	0,538	0,577	0,577	0,519	0,577	0,596	0,635	0,538	0,577	0,481	0,562 +/- 0,043
SMS12	34	0,853	0,853	0,882	0,882	0,912	0,853	0,912	0,912	0,912	0,882	0,885 +/- 0,026

B CE2 + Female Young

Name	N Sample	Score1	Score2	Score3	Score4	Score5	Score6	Score7	Score8	Score9	Score10	avg.+std
517A	6	0,333	0,167	0,333	0,167	0,333	0,167	0,167	0,167	0,167	0,167	0,217 +/- 0,081
612A	12	0,000	0,083	0,083	0,167	0,083	0,083	0,083	0,167	0,083	0,083	0,092 +/- 0,047
637A	10	0,600	0,400	0,600	0,600	0,500	0,600	0,600	0,500	0,600	0,600	0,56 +/- 0,07
SMS4	11	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1 +/- 0
SMS5	8	0,750	0,750	0,750	0,750	0,875	0,750	0,625	0,750	0,750	0,750	0,75 +/- 0,059
SMS6	13	0,385	0,385	0,385	0,462	0,385	0,615	0,462	0,462	0,462	0,385	0,438 +/- 0,073

C CE2 + Male Young

Name	N Sample	Score1	Score2	Score3	Score4	Score5	Score6	Score7	Score8	Score9	Score10	avg.+std
511O	12	0,917	0,917	0,917	0,917	1,000	0,917	0,917	0,917	1,000	0,917	0,933 +/- 0,035
618O	7	0,857	0,857	1,000	1,000	0,857	1,000	0,857	0,857	1,000	1,000	0,929 +/- 0,075
743O	7	0,857	0,857	0,857	0,857	0,857	0,857	0,857	0,857	0,857	0,857	0,857 +/- 0
SMS1	28	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,964	0,996 +/- 0,011
SMS2	17	0,941	0,941	1,000	0,882	0,941	0,941	0,941	0,882	0,941	0,882	0,929 +/- 0,037
SMS3	13	0,846	0,846	0,846	0,923	0,846	0,846	0,846	0,846	0,846	0,923	0,862 +/- 0,032

D CE2 + Male Old

Name	N Sample	Score1	Score2	Score3	Score4	Score5	Score6	Score7	Score8	Score9	Score10	avg.+std
11OADS	13	0,385	0,308	0,385	0,308	0,308	0,308	0,308	0,385	0,385	0,308	0,338 +/- 0,04
819O	17	0,824	0,882	0,765	0,824	0,882	0,765	0,882	0,882	0,824	0,882	0,841 +/- 0,048
842O	15	0,733	0,733	0,733	0,733	0,733	0,733	0,733	0,733	0,733	0,733	0,733 +/- 0
SMS7	13	0,077	0,231	0,077	0,077	0,077	0,231	0,231	0,154	0,231	0,077	0,146 +/- 0,076
SMS8	33	0,848	0,818	0,909	0,788	0,848	0,848	0,848	0,758	0,939	0,818	0,842 +/- 0,053
SMS9	13	0,385	0,308	0,308	0,385	0,308	0,231	0,385	0,231	0,231	0,308	0,308 +/- 0,063

FIGURE 10
Summary of the results for the CE2, using RF. **(A)** Detailed performance for the old female subgroup. **(B)** Detailed performance for the young female subgroup. **(C)** Detailed performance for the young male subgroup. **(D)** Detailed performance for the old male subgroup.

K-Nearest Neighbors

A CE2 + Female Old

Name	N Sample	Score1	Score2	Score3	Score4	Score5	Score6	Score7	Score8	Score9	Score10	avg.+std
10AGPC	13	0,923	0,846	0,923	1	0,846	0,923	0,846	0,846	0,923	0,923	0,9 +/- 0,052
11AAZM	13	1	1	1	1	1	1	1	1	1	1	1 +/- 0
12109A	6	1	1	1	1	1	1	1	1	1	1	1 +/- 0
SMS10	13	1	1	1	1	1	1	1	1	1	1	1 +/- 0
SMS11	52	0,538	0,423	0,462	0,481	0,442	0,462	0,404	0,442	0,5	0,462	0,462 +/- 0,038
SMS12	34	0,882	0,882	0,912	0,853	0,941	0,941	0,853	0,882	0,882	0,853	0,888 +/- 0,033

B CE2 + Female Young

Name	N Sample	Score1	Score2	Score3	Score4	Score5	Score6	Score7	Score8	Score9	Score10	avg.+std
517A	6	0,5	0,5	0,667	0,5	0,5	0,5	0,5	0,667	0,667	0,5	0,55 +/- 0,081
612A	12	0,083	0,167	0,083	0,083	0,083	0,083	0,167	0,167	0,083	0	0,1 +/- 0,053
637A	10	0,6	0,7	0,6	0,6	0,6	0,7	0,6	0,7	0,6	0,8	0,65 +/- 0,071
SMS4	11	1,000	1,000	1	1	1	1	1	1	1	1	1 +/- 0
SMS5	8	0,625	0,625	0,75	0,625	0,75	0,75	0,75	0,75	0,625	0,75	0,7 +/- 0,065
SMS6	13	0,231	0,385	0,308	0,615	0,308	0,462	0,385	0,385	0,385	0,308	0,377 +/- 0,105

C CE2 + Male Young

Name	N Sample	Score1	Score2	Score3	Score4	Score5	Score6	Score7	Score8	Score9	Score10	avg.+std
511O	12	1	1	1	1	1	1	1	1	1	1	1 +/- 0
618O	7	1	1	1	1	1	1	1	1	1	1	1 +/- 0
743O	7	0,857	0,857	1	0,857	1	1	0,857	1	0,857	0,857	0,914 +/- 0,074
SMS1	28	0,964	0,964	1	1	0,964	1	1	0,964	0,964	1	0,982 +/- 0,019
SMS2	17	0,941	0,941	0,941	0,941	0,882	0,882	0,941	0,882	0,882	0,882	0,912 +/- 0,031
SMS3	13	0,846	0,846	0,923	0,846	0,923	0,769	0,846	0,769	0,846	0,923	0,854 +/- 0,057

D CE2 + Male Old

Name	N Sample	Score1	Score2	Score3	Score4	Score5	Score6	Score7	Score8	Score9	Score10	avg.+std
11OADS	13	0,385	0,385	0,308	0,308	0,308	0,308	0,385	0,308	0,308	0,385	0,338 +/- 0,04
819O	17	1	1	1	1	1	1	1	1	1	1	1 +/- 0
842O	15	0,733	0,733	0,733	0,733	0,733	0,800	0,733	0,733	0,733	0,733	0,74 +/- 0,021
SMS7	13	0,231	0,154	0,308	0,231	0,154	0,231	0,231	0,154	0,231	0,231	0,215 +/- 0,049
SMS8	33	0,788	0,758	0,788	0,818	0,818	0,818	0,727	0,727	0,758	0,758	0,776 +/- 0,036
SMS9	13	0,154	0,154	0,077	0,154	0,154	0,077	0,154	0,154	0,154	0,077	0,131 +/- 0,037

FIGURE 11
Summary of the results for the CE2, using KNN. **(A)** Detailed performance for the old female subgroup. **(B)** Detailed performance for the young female subgroup. **(C)** Detailed performance for the young male subgroup. **(D)** Detailed performance for the old male subgroup.

34 NA values (20 in normotypic group and 14 in non-normotypic group). This difference can be attributed to the limitation of the training dataset in CE1 (without SMOTE), which leads to the models generating identical results due to data bias. However, when SMOTE is applied, the different models can produce diverse results due to data augmentation process and the correction of bias during training. Analyzing the results reveals that some speakers, like 11AAZM and SMS04, are highly identifiable across all methods, achieving 100% accuracy and received "Not Applicable" (NA) values in all one-to-one Wilcoxon comparisons. Likewise, while most comparisons yield p-valuess below 0.05, indicating statistically significant differences, the RF vs. KNN comparison shows 12 non-significant results. This suggests similar performance for these methods, potentially making them less effective than the others. Conversely, SVM and LDA generally exhibit more statistically significant values, implying stronger distinctions in their performance compared to the other ML methods.

Another point of debate is whether the SMOTE technique can affect the performance of the different models. In Blagus and Lusa (2013), the authors applied this technique to high-dimensionality cases. However, here, it is addressed a single dimension (the CPP). The obtained results agree with those of the previously referenced work. First, the authors noted that for low-dimensionality cases, SMOTE usually represents an improvement (e.g., the RF, SVM and KNN cases) or equates the results to those of other undersampling techniques (e.g., the LDA case). These results agree with those achieved in the current study, i.e., for the four ML techniques used, the results were improved with the application of the SMOTE technique. There are techniques

Support Vector Machine

A CE2 + Female Old

Name	N Sample	Score1	Score2	Score3	Score4	Score5	Score6	Score7	Score8	Score9	Score10	avg.+std
10AGPC	13	1	1	1	1	0,923	1	1	1	1	1	0,992 +/- 0,024
11AAZM	13	1	1	1	1	1	1	1	1	1	1	1 +/- 0
12109A	6	1	1	1	1	1	1	1	1	1	1	1 +/- 0
SMS10	13	1	1	1	1	1	1	1	1	1	1	1 +/- 0
SMS11	52	0,327	0,327	0,365	0,327	0,385	0,327	0,327	0,327	0,327	0,423	0,346 +/- 0,034
SMS12	34	0,824	0,765	0,824	0,794	0,794	0,794	0,794	0,824	0,824	0,824	0,806 +/- 0,021

B CE2 + Female Young

Name	N Sample	Score1	Score2	Score3	Score4	Score5	Score6	Score7	Score8	Score9	Score10	avg.+std
517A	6	0,5	0,5	0,5	0,5	0,5	0,5	0,5	0,5	0,5	0,5	0,5 +/- 0
612A	12	0,083	0,083	0,083	0,083	0,083	0,083	0,083	0,083	0,083	0,083	0,083 +/- 0
637A	10	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,7 +/- 0
SMS4	11	1	1	1	1	1	1	1	1	1	1	1 +/- 0
SMS5	8	0,75	0,75	0,75	0,75	0,75	0,75	0,75	0,75	0,75	0,75	0,75 +/- 0
SMS6	13	0,231	0,308	0,308	0,308	0,308	0,308	0,231	0,231	0,231	0,308	0,277 +/- 0,04

C CE2 + Male Young

Name	N Sample	Score1	Score2	Score3	Score4	Score5	Score6	Score7	Score8	Score9	Score10	avg.+std
511O	12	1	1	1	1	1	1	1	1	1	1	1 +/- 0
618O	7	1	1	1	1	1	1	1	1	1	1	1 +/- 0
743O	7	1	1	1	1	1	1	1	1	1	1	1 +/- 0
SMS1	28	0,929	0,929	0,929	0,929	0,929	0,929	0,929	0,929	0,929	0,929	0,929 +/- 0
SMS2	17	0,941	0,941	0,941	0,941	0,941	0,941	0,941	0,941	0,941	0,941	0,941 +/- 0
SMS3	13	0,769	0,769	0,769	0,769	0,769	0,769	0,769	0,769	0,769	0,769	0,769 +/- 0

D CE2 + Male Old

Name	N Sample	Score1	Score2	Score3	Score4	Score5	Score6	Score7	Score8	Score9	Score10	avg.+std
11OADS	13	0,846	0,846	0,769	0,846	0,846	0,615	0,846	0,846	0,846	0,846	0,815 +/- 0,074
819O	17	1	1	1	1	1	1	1	1	1	1	1 +/- 0
842O	15	1	1	0,933	1	0,933	0,933	1	1	1	1	0,98 +/- 0,032
SMS7	13	0,077	0,077	0,077	0,077	0,077	0,077	0,077	0,077	0,077	0,077	0,077 +/- 0
SMS8	33	0,667	0,667	0,667	0,667	0,667	0,667	0,727	0,667	0,667	0,667	0,673 +/- 0,019
SMS9	13	0	0	0	0	0	0	0	0	0	0	0 +/- 0

FIGURE 12
Summary of the results for the CE2, using SVM. **(A)** Detailed performance for the old female subgroup. **(B)** Detailed performance for the young female subgroup. **(C)** Detailed performance for the young male subgroup. **(D)** Detailed performance for the old male subgroup.

Linear Discriminant Analysis

A CE2 + Female Old

Name	N Sample	Score1	Score2	Score3	Score4	Score5	Score6	Score7	Score8	Score9	Score10	avg.+std
10AGPC	13	1	0,846	0,923	0,923	0,846	0,846	0,923	0,846	0,923	0,846	0,892 +/- 0,054
11AAZM	13	1	1	1	1	1	1	1	1	1	1	1 +/- 0
12109A	6	1	1	1	1	1	1	1	1	1	1	1 +/- 0
SMS10	13	1	1	1	1	1	1	1	1	1	1	1 +/- 0
SMS11	52	0,404	0,404	0,423	0,462	0,462	0,442	0,442	0,462	0,442	0,481	0,442 +/- 0,026
SMS12	34	0,971	1,000	0,971	1,000	1,000	0,971	0,971	0,971	0,971	1	0,982 +/- 0,015

B CE2+Female Young

Name	N Sample	Score1	Score2	Score3	Score4	Score5	Score6	Score7	Score8	Score9	Score10	avg.+std
517A	6	0,333	0,333	0,333	0,333	0,333	0,333	0,333	0,333	0,5	0,333	0,35 +/- 0,053
612A	12	0	0	0	0	0	0	0	0	0	0	0 +/- 0
637A	10	0,9	0,9	0,9	0,8	0,9	0,9	0,8	0,9	0,9	0,9	0,88 +/- 0,042
SMS4	11	1	1	1	1	1	1	1	1	1	1	1 +/- 0
SMS5	8	1	1	0,875	1	1	1	1	0,875	1	1	0,975 +/- 0,053
SMS6	13	0,308	0,308	0,308	0,308	0,308	0,308	0,308	0,308	0,231	0,308	0,3 +/- 0,024

C CE2+Male Young

Name	N Sample	Score1	Score2	Score3	Score4	Score5	Score6	Score7	Score8	Score9	Score10	avg.+std
511O	12	1,000	0,917	0,917	0,917	0,917	1,000	1,000	1,000	0,917	1,000	0,958 +/- 0,044
618O	7	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1 +/- 0
743O	7	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1 +/- 0
SMS1	28	1,000	1,000	1,000	1,000	0,964	0,964	0,964	1,000	1,000	1,000	0,989 +/- 0,017
SMS2	17	0,941	0,941	0,941	0,941	0,941	0,941	0,941	0,941	0,941	0,941	0,941 +/- 0
SMS3	13	0,692	0,692	0,769	0,692	0,769	0,692	0,769	0,769	0,692	0,769	0,731 +/- 0,041

D CE2+Male Old

Name	N Sample	Score1	Score2	Score3	Score4	Score5	Score6	Score7	Score8	Score9	Score10	avg.+std
11OADS	13	0,769	0,615	0,538	0,538	0,538	0,538	0,538	0,615	0,538	0,538	0,577 +/- 0,075
819O	17	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1 +/- 0
842O	15	0,933	1,000	0,933	1,000	0,933	0,933	0,933	1,000	0,933	1,000	0,96 +/- 0,034
SMS7	13	0,231	0,231	0,231	0,231	0,231	0,231	0,231	0,231	0,231	0,231	0,231 +/- 0
SMS8	33	0,848	0,879	0,879	0,879	0,848	0,879	0,879	0,848	0,848	0,879	0,867 +/- 0,016
SMS9	13	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0 +/- 0

FIGURE 13
Summary of the results for the CE2, using LDA. **(A)** Detailed performance for the old female subgroup. **(B)** Detailed performance for the young female subgroup. **(C)** Detailed performance for the young male subgroup. **(D)** Detailed performance for the old male subgroup.

that can be regarded as more beneficial than others while others may be less beneficial (e.g., high-dimensionality cases). For example, a secondary effect of SMOTE is that the new samples from the minority class exhibit variances one-third smaller than those of the original distribution. This result implies that this technique is not as effective in methods that use variance as an indicator, such as the LDA. RF, SVM, and KNN are the methods that offer better results in cases of low dimensionality. In the case of SVM, it has meant an improvement, but it has not quite reached the expected performance. The reason for this behavior may be due to the combination of the increase in the dimensionality of the SVM itself along with the use of SMOTE. Likewise, the interaction between LDA and KNN methods with SMOTE is negligible, since the Euclidean distance between the classes is the same, before and after the use of SMOTE with low dimensionality, as demonstrated by Blagus and Lusa (2013).

Interestingly, in this research work, the average accuracy across ML methods is similar for every single method. In CE1 (without applied SMOTE technique – see Table 2), all methods achieved values: RF (71.1%), KNN (70.6%), SVM (68.6%), and LDA (70.1%). Notably, RF performed best with 71.1% accuracy.

For CE2 (with SMOTE technique – see Table 3), average accuracy increased across all methods compared to CE1, reaching 70.6% for RF, 72.9% for KNN, 73.5% for SVM, and 75.3% for LDA. Notably, LDA emerged as the best performer in CE2 with an average accuracy of 75.3%. This finding suggests that the data augmentation techniques used in CE2 led to overall improved performance.

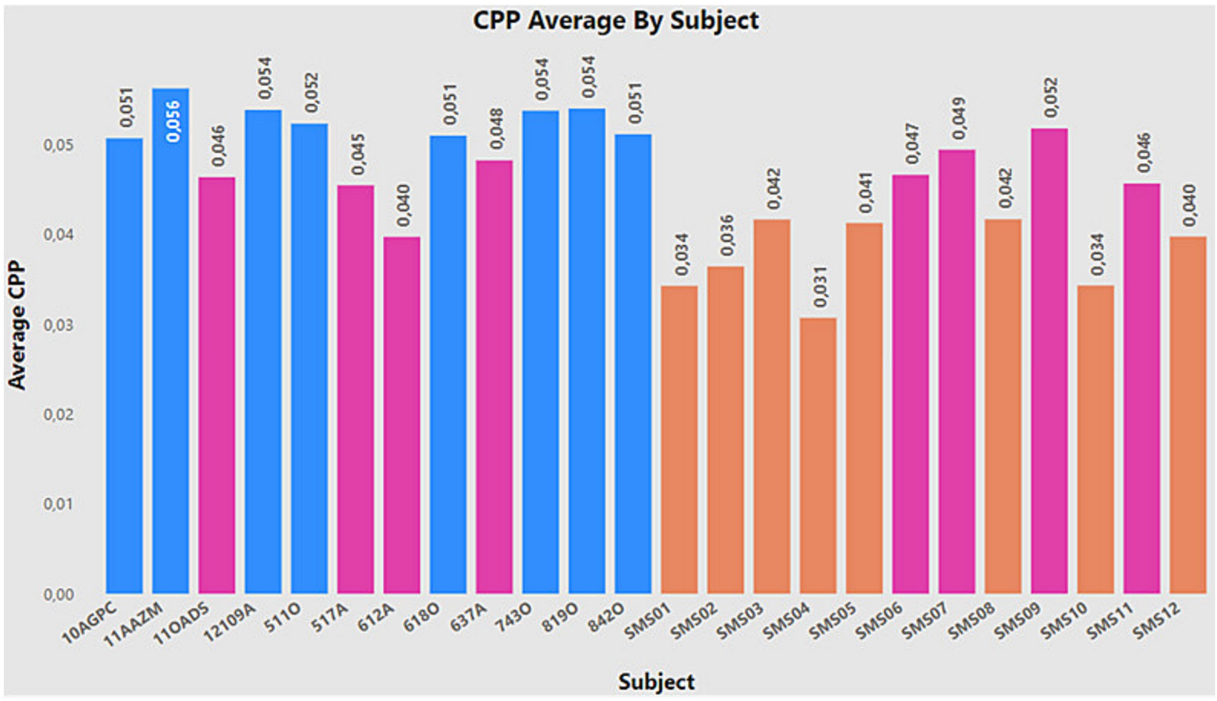


FIGURE 14
CPP average by subject.

5 Conclusion

Two objectives have been achieved in the work. The first objective showed that, due to the application of correct data preprocessing, the performance of the models can be improved, as demonstrated through different case studies. Furthermore, the outcomes of CE2 are more reliable and robust compared to the results of CE1, owing to the application of data augmentation techniques. While it may appear that CE1 has a superior classification rate, this is primarily due to the class imbalance, with a greater number of SMS samples compared to normotypical ones. The second goal of the work was to study whether the CPP is a suitable metric for the identification of SMS vs. normotypic individuals, and, according to the results obtained in the last case study, it can be confirmed that this metric fulfils this function. The main limitation of the study is the number of individuals with SMS currently available. However, this situation opens the opportunity to explore different data augmentation methods and compare their performance to find the most suitable one for the study context. A similar process will be carried out with the machine learning algorithms, using different variants of them. Another interesting approach would be the inclusion of cost-sensitive algorithms. As explained in Figure 14, individuals with outlier values have been identified compared to their respective groups. Therefore, it may be beneficial to implement counterfactual methods to decrease the biased caused by those outliers.

Regarding the supervised learning models used, no attempts were made to identify the ideal iteration that would yield a very high result. This is because when such a model is applied in a real-world context, it tends to underperform due to its adaptation to a specific data combination for achieving the results. As a result, the initial case study reveals models that are biased toward the target class (SMS), while the final case study presents models with less bias and a high precision rate. The results also indicate that performance improves following a series of transformations on complex initial data. However, to enhance and solidify these results, it is essential to obtain samples from new subjects.

Furthermore, it is important to highlight the presence of certain individuals who show significantly low detection rates in most models, considering CE2 as a reference. These individuals include 11OADS, 517A, 612A, 637A (the latter shows good performance in LDA and SVM, but not in the rest), as well as SMS6, SMS7, SMS9, and SMS11. Figure 14 presents the average CPP value for everyone stored in the database, remembering that the normative group should exhibit higher CPP values, while the non-normative group should show lower values. The bars marked in pink correspond to the individuals mentioned above, showing how they present higher or lower values than their respective groups. In other words, these individuals constitute the decision boundary of the problem. This finding raises possible future approaches, such as the application of synthetic data augmentation methods on the decision boundary, assigning weights to the problem samples, opening new possibilities to improve model performance.

Finally, two potential avenues of research are proposed. The first involves replicating the same machine learning procedures with other rare diseases, such as WS. The goal would be to compare performance and potentially conduct a case study where different models are trained to distinguish between SW and SMS individuals, thereby extracting the similarities and differences between both pathologies. The second avenue of research would focus on the application of deep learning techniques. However, to develop more robust models, it would first be necessary to increase the number of SMS samples. It

should be noted that authors explore several new methods based on SMOTE techniques and data augmentation methods in future research works.

Data availability statement

The datasets presented in this article are not readily available because the data collected in this research are subject to data protection law due to their biometric and sensitive nature. Furthermore, the study population is minors. Requests to access the datasets should be directed to DP-A, daniel.palacios@urjc.es.

Ethics statement

The studies involving humans were approved by Universidad Politécnica de Madrid. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

Author contributions

RF-R: Conceptualization, Formal analysis, Software, Writing – original draft, Writing – review & editing. EN-V: Investigation, Supervision, Writing – original draft, Writing – review & editing. IH-d: Data curation, Validation, Writing – review & editing. EG-H: Data curation, Validation, Writing – review & editing. AÁ-M: Supervision, Validation, Writing – review & editing. RM-O: Formal analysis, Software, Writing – review & editing. DP-A: Conceptualization, Funding acquisition, Project administration, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was supported by the University Rey Juan Carlos, under grant 2023/00004/039-M3002.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abdar, M., Nasarian, E., Zhou, X., Bargshady, G., Wijayaningrum, V. N., and Hussain, S. (2019). Performance improvement of decision trees for diagnosis of coronary artery disease using multi filtering approach, In 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS) 26–30. doi: 10.1109/CCOMS.2019.8821633,

Alabi, R. O., Elmusrati, M., Sawazaki-Calone, I., Kowalski, L. P., Haglund, C., Coletta, R. D., et al. (2020). Comparison of supervised machine learning classification techniques in prediction of locoregional recurrences in early oral tongue cancer. *Int. J. Med. Inform.* 136:104068. doi: 10.1016/j.ijmedinf.2019.104068

Albertini, G., Bonassi, S., Dall'Armi, V., Giachetti, I., Giaquinto, S., and Mignano, M. (2010). Spectral analysis of the voice in down syndrome. *Res. Dev. Disabil.* 31, 995–1001. doi: 10.1016/J.RIDD.2010.04.024

Ali, L., Zhu, C., Zhang, Z., and Liu, Y. (2019). Automated detection of Parkinson's disease based on multiple types of sustained phonations using linear discriminant analysis and genetically optimized neural network. *IEEE J. Trans. Eng. Health Med.* 7, 1–10. doi: 10.1109/JTEHM.2019.2940900

Antonell, A., Del Campo, M., Flores, R., Campuzano, V., and Pérez-Jurado, L. A. (2006). Síndrome de Williams: Aspectos clínicos y bases moleculares. *Rev. Neurol.* 42:S069. doi: 10.33588/rn.42s01.2005738

Ayvaz, U., Gürüler, H., Khan, F., Ahmed, N., Whangbo, T., and Bobomirzaevich, A. (2022). Automatic speaker recognition using Mel-frequency cepstral coefficients through machine learning. *CMC-Comp. Mater. Continua* 71, 5511–5521. doi: 10.32604/cmc.2022.023278

Blagus, R., and Lusa, L. (2013). SMOTE for high-dimensional class-imbalanced data. *BMC Bioinform.* 14:106. doi: 10.1186/1471-2105-14-106

Boslaugh, S. (2012). *Statistics in a nutshell: a desktop quick reference*. USA: O'Reilly Media, Inc.

Bozhilova, N., Welham, A., Adams, D., Bissell, S., Bruining, H., Crawford, H., et al. (2023). Profiles of autism characteristics in thirteen genetic syndromes: a machine learning approach. *Mol. Autism.* 14:3. doi: 10.1186/s13229-022-00530-5

Brendal, M. A., King, K. A., Zalewski, C. K., Finucane, B. M., Introne, W., Brewer, C. C., et al. (2017). Auditory phenotype of Smith–Magenis syndrome. *J. Speech Lang. Hear. Res.* 60:1076. doi: 10.1044/2016_JSLHR-H-16-0024

Brinca, L. F., Batista, A. P. F., Tavares, A. I., Gonçalves, I. C., and Moreno, M. L. (2014). Use of cepstral analyses for differentiating Normal from dysphonic voices: a comparative study of connected speech versus sustained vowel in European Portuguese female speakers. *J. Voice* 28, 282–286. doi: 10.1016/j.jvoice.2013.10.001

Cai, H., Huang, X., Liu, Z., Liao, W., Dai, H., Wu, Z., et al. (2023). Exploring multimodal approaches for Alzheimer's disease detection using patient speech transcript and audio data. *arXiv*. doi: 10.48550/arXiv.2307.02514

Calà, F., Frassineti, L., Sforza, E., Onesimo, R., D'Alatri, L., Manfredi, C., et al. (2023). Artificial intelligence procedure for the screening of genetic syndromes based on voice characteristics. *Bioengineering* 10:1375. doi: 10.3390/bioengineering10121375

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 16, 321–357. doi: 10.1613/jair.953

Cifci, M. A., and Hussain, S. (2018). Data mining usage and applications in health services. *Int. J. Inform. Visual.* 2, 225–231. doi: 10.30630/joiv.2.4.148

Edelman, E. A., Girirajan, S., Finucane, B., Patel, P. I., Lupski, J. R., Smith, A. C. M., et al. (2007). Gender, genotype, and phenotype differences in Smith-Magenis syndrome: a meta-analysis of 105 cases. *Clin. Genet.* 71, 540–550. doi: 10.1111/J.1399-0004.2007.00815.X

Elsea, S. H., and Girirajan, S. S. (2008). Smith-Magenis syndrome. *Euro. J. Human Genet.* 16, 412–421. doi: 10.1038/SJ.EJHG.5202009

Firouzi, F., Rahmani, A. M., Mankodiya, K., Badaroglu, M., Merrett, G. V., Wong, P., et al. (2018). Internet-of-things and big data for smarter healthcare: from device to architecture, applications and analytics. *Futur. Gener. Comput. Syst.* 78, 583–586. doi: 10.1016/j.future.2017.09.016

Frassineti, L., Zucconi, A., Calà, F., Sforza, E., Onesimo, R., Leoni, C., et al. (2021). Analysis of vocal patterns as a diagnostic tool in patients with genetic syndromes. *Proc. Rep.* doi: 10.36253/978-88-5518-449-6

Girirajan, S., Truong, H. T., Blanchard, C. L., and Elsea, S. H. (2009). A functional network module for Smith-Magenis syndrome. *Clin. Genet.* 75, 364–374. doi: 10.1111/j.1399-0004.2008.01135.x

Górriz, J. M., Álvarez-Illán, I., Álvarez-Marquina, A., Arco, J. E., Atzmueller, M., Ballarini, F., et al. (2023). Computational approaches to explainable artificial intelligence: advances in theory, applications and trends. *Inform. Fus.* 100:101945. doi: 10.1016/j.inffus.2023.101945

Górriz, J. M., Ramírez, J., Ortíz, A., Martínez-Murcia, F. J., Segovia, F., Suckling, J., et al. (2020). Artificial intelligence within the interplay between natural and artificial computation: advances in data science, trends and applications. *Neurocomputing* 410, 237–270. doi: 10.1016/j.neucom.2020.05.078

Greenberg, F., Lewis, R. A., Potocki, L., Glaze, D., Parke, J., Killian, J., et al. (1996). Multi-disciplinary clinical study of Smith-Magenis syndrome (deletion 17p11. 2). *Am. J. Med. Genet.* 62, 247–254. doi: 10.1002/(SICI)1096-8628(19960329)62:3<247::AID-AJMG9>3.0.CO;2-Q

Heman-Ackah, Y. D., Michael, D. D., Baroody, M. M., Ostrowski, R., Hillenbrand, J., Heuer, R. J., et al. (2003). Cepstral peak prominence: a more reliable measure of dysphonia. *Ann. Otol. Rhinol. Laryngol.* 112, 324–333. doi: 10.1177/000348940311200406

Hidalgo, I., Gómez Vilda, P., and Garayzábal, E. (2018). Biomechanical description of phonation in children affected by Williams syndrome. *J. Voice* 32, 515.e15–515.e28. doi: 10.1016/J.JVOICE.2017.07.002

Hidalgo-De la Guía, I., Garayzábal, E., Gómez-Vilda, P., and Palacios-Alonso, D. (2021a). Specificities of phonation biomechanics in down syndrome children. *Biomed. Sig. Process. Control* 63:102219. doi: 10.1016/J.BSPC.2020.102219

Hidalgo-De la Guía, I., Garayzábal-Heinze, E., Gómez-Vilda, P., Martínez-Olalla, R., and Palacios-Alonso, D. (2021b). Acoustic analysis of phonation in children with Smith–Magenis syndrome. *Front. Hum. Neurosci.* 15:259. doi: 10.3389/FNHUM.2021.661392/BIBTEX

Izenman, A. J. (2008). "Linear discriminant analysis" in *Modern multivariate statistical techniques: Regression, classification, and manifold learning*. ed. A. J. Izenman (New York, NY: Springer (Springer Texts in Statistics)).

Jakkula, V. (2006) 'Tutorial on support vector machine (svm)'. *School of EECS, Washington State University*, 37(2.5), 3. Available at: https://course.ccs.neu.edu/cs5100f11/resources/jakkula.pdf (Accessed February 23, 2024).

Jeffery, T., Cunningham, S., and Whiteside, S. P. (2018). Analyses of sustained vowels in down syndrome (DS): a case study using spectrograms and perturbation data to investigate voice quality in four adults with DS. *J. Voice* 32, 644.e11–644.e24. doi: 10.1016/j.jvoice.2017.08.004

Jia, J., Wang, R., An, Z., Guo, Y., Ni, X., and Shi, T. (2018). RDAD: a machine learning system to support phenotype-based rare disease diagnosis. *Front. Genet.* 9:587. doi: 10.3389/fgene.2018.00587

Joloudari, J. H., Marefat, A., Nematollahi, M. A., Oyelere, S. S., and Hussain, S. (2023). Effective class-imbalance learning based on SMOTE and convolutional neural networks. *Appl. Sci.* 13:4006. doi: 10.3390/app13064006

Lee, J. Y. (2021). Experimental evaluation of deep learning methods for an intelligent pathological voice detection system using the Saarbruecken voice database. *Appl. Sci.* 11:7149. doi: 10.3390/APP11157149

Li, X., Wang, Y., Wang, D., Yuan, W., Peng, D., and Mei, Q. (2019). Improving rare disease classification using imperfect knowledge graph. *BMC Med. Inform. Decis. Mak.* 19:238. doi: 10.1186/s12911-019-0938-1

Linders, C. C., van Eeghen, A. M., Zinkstok, J. R., van den Boogaard, M.-J., and Boot, E. (2023). Intellectual and behavioral phenotypes of Smith–Magenis syndrome: comparisons between individuals with a 17p11.2 deletion and pathogenic RAI1 variant. *Genes* 14:1514. doi: 10.3390/genes14081514

Moers, C., Möbius, B., Rosanowski, F., Nöth, E., Eysholdt, U., and Haderlein, T. (2012). Vowel- and text-based cepstral analysis of chronic hoarseness. *J. Voice* 26, 416–424. doi: 10.1016/j.jvoice.2011.05.001

Moore, J., and Thibeault, S. (2012). Insights into the role of elastin in vocal fold health and disease. *J. Voice* 26, 269–275. doi: 10.1016/J.JVOICE.2011.05.003

Orozco-Arroyave, J.R., Arias-Londoño, J.D., Vargas-Bonilla, J.F., González-Rátiva, M.C., and Nöth, E. (2014) 'New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease', in N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard and J. Mariani, (eds) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). LREC 2014, Reykjavik, Iceland: European Language Resources Association (ELRA).

Orphanet (2023). Available at: https://rb.gy/98xds (Accessed February 15, 2024).

Pachange, S., Joglekar, B., and Kulkarni, P. (2015) 'An ensemble classifier approach for disease diagnosis using random Forest. In 2015 Annual IEEE India Conference (INDICON), New Delhi, India: IEEE.

Peterson, E. A., Roy, N., Awan, S. N., Merrill, R. M., Banks, R., and Tanner, K. (2013). Toward validation of the cepstral spectral index of dysphonia (CSID) as an objective treatment outcomes measure. *J. Voice* 27, 401–410. doi: 10.1016/j.jvoice.2013.04.002

Rasmussen, C. (1999). "The infinite Gaussian mixture model" in *Advances in neural information processing systems* (Cambridge, Massachusetts: MIT Press).

Rother, A.-K., Schwerk, N., Brinkmann, F., Klawonn, F., Lechner, W., and Grigull, L. (2015). Diagnostic support for selected Paediatric pulmonary diseases using answer-pattern recognition in questionnaires based on combined data mining applications--a monocentric observational pilot study. *PLoS One* 10:e0135180. doi: 10.1371/journal.pone.0135180

Rusko, M., Sabo, R., Trnka, M., Zimmermann, A., Malaschitz, R., Ružický, E., et al. (2023). EWA-DB, Slovak database of speech affected by neurodegenerative diseases'. medRxiv.10.13.23296810. doi: 10.1101/2023.10.13.23296810,

Shen, F., Liu, S., Wang, Y., Wang, L., Afzal, N., and Liu, H. (2017). Leveraging collaborative filtering to accelerate rare disease diagnosis. *Annu. Symp. Proc.* 2017, 1554–1563. https://pubmed.ncbi.nlm.nih.gov/29854225/

Shorten, C., and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *J. Big Data* 6:60. doi: 10.1186/s40537-019-0197-0

Sinaga, K. P., and Yang, M.-S. (2020). Unsupervised K-means clustering algorithm. *IEEE Access* 8, 80716–80727. doi: 10.1109/ACCESS.2020.2988796

Slager, R. E., Newton, T. L., Vlangos, C. N., Finucane, B., and Elsea, S. H. (2003). Mutations in RAI1 associated with Smith-Magenis syndrome. *Nat. Genet.* 33, 466–468. doi: 10.1038/NG1126

Spiga, O., Cicaloni, V., Fiorini, C., Trezza, A., Visibelli, A., Millucci, L., et al. (2020). Machine learning application for development of a data-driven predictive model able to investigate quality of life scores in a rare disease. *Orphanet J. Rare Dis.* 15:46. doi: 10.1186/s13023-020-1305-0

Uddin, S., Haque, I., Lu, H., Moni, M. A., and Gide, E. (2022). Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. *Sci. Rep.* 12:6256. doi: 10.1038/s41598-022-10358-x

Vlangos, C. N., Yim, D. K. C., and Elsea, S. H. (2003). Refinement of the Smith–Magenis syndrome critical region to ~950 kb and assessment of 17p11.2 deletions. Are all deletions created equally? *Mol. Genet. Metab.* 79, 134–141. doi: 10.1016/S1096-7192(03)00048-9

Warule, P., Mishra, S. P., and Deb, S. (2023). Time-frequency analysis of speech signal using Chirplet transform for automatic diagnosis of Parkinson's disease. *Biomed. Eng. Lett.* 13, 613–623. doi: 10.1007/s13534-023-00283-x

Watts, C. R., Awan, S. N., and Marler, J. A. (2008). An investigation of voice quality in individuals with inherited elastin gene abnormalities. *Clin. Linguist. Phon.* 22, 199–213. doi: 10.1080/02699200701803361

Zhang, S., Poon, S. K., Vuong, K., Sneddon, A., and Loy, C. T. (2019). A deep learning-based approach for gait analysis in Huntington disease. *Stud. Health Technol. Inform.* 264, 477–481. doi: 10.3233/SHTI190267

Zolnoori, M., Zolnour, A., and Topaz, M. (2023). ADscreen: a speech processing-based screening system for automatic identification of patients with Alzheimer's disease and related dementia. *Artif. Intell. Med.* 143:102624. doi: 10.1016/j.artmed.2023.102624

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 12 April 2024
DOI 10.3389/fnins.2024.1363930

OPEN ACCESS

EDITED BY
Roberto Maffulli,
Italian Institute of Technology (IIT), Italy

REVIEWED BY
Guokai Zhang,
University of Shanghai for Science and Technology, China
Ebrahim Elsayed,
Mansoura University, Egypt

*CORRESPONDENCE
Yan Kang
✉ kangyan@sztu.edu.cn
Yingwei Guo
✉ guoyingwei8801@163.com

RECEIVED 31 December 2023
ACCEPTED 04 March 2024
PUBLISHED 12 April 2024

CITATION
Zaman A, Hassan H, Zeng X, Khan R, Lu J, Yang H, Miao X, Cao A, Yang Y, Huang B, Guo Y and Kang Y (2024) Adaptive Feature Medical Segmentation Network: an adaptable deep learning paradigm for high-performance 3D brain lesion segmentation in medical imaging.
Front. Neurosci. 18:1363930.
doi: 10.3389/fnins.2024.1363930

Adaptive Feature Medical Segmentation Network: an adaptable deep learning paradigm for high-performance 3D brain lesion segmentation in medical imaging

Asim Zaman[1,2,3,4], Haseeb Hassan[2], Xueqiang Zeng[2,3], Rashid Khan[3,4,5], Jiaxi Lu[2,3], Huihui Yang[2,3], Xiaoqiang Miao[2,6], Anbo Cao[2,3], Yingjian Yang[7], Bingding Huang[5], Yingwei Guo[2,8]* and Yan Kang[1,2,3,4,6]*

[1]School of Biomedical Engineering, Shenzhen University Medical School, Shenzhen University, Shenzhen, China, [2]College of Health Science and Environmental Engineering, Shenzhen Technology University, Shenzhen, China, [3]School of Applied Technology, Shenzhen University, Shenzhen, China, [4]Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, School of Biomedical Engineering, Medical School, Shenzhen University, Shenzhen, China, [5]College of Big Data and Internet, Shenzhen Technology University, Shenzhen, China, [6]College of Medicine and Biological Information Engineering, Northeastern University, Shenyang, China, [7]Shenzhen Lanmage Medical Technology Co., Ltd, Shenzhen, China, [8]School of Electrical and Information Engineering, Northeast Petroleum University, Daqing, China

Introduction: In neurological diagnostics, accurate detection and segmentation of brain lesions is crucial. Identifying these lesions is challenging due to its complex morphology, especially when using traditional methods. Conventional methods are either computationally demanding with a marginal impact/enhancement or sacrifice fine details for computational efficiency. Therefore, balancing performance and precision in compute-intensive medical imaging remains a hot research topic.

Methods: We introduce a novel encoder-decoder network architecture named the Adaptive Feature Medical Segmentation Network (AFMS-Net) with two encoder variants: the Single Adaptive Encoder Block (SAEB) and the Dual Adaptive Encoder Block (DAEB). A squeeze-and-excite mechanism is employed in SAEB to identify significant data while disregarding peripheral details. This approach is best suited for scenarios requiring quick and efficient segmentation, with an emphasis on identifying key lesion areas. In contrast, the DAEB utilizes an advanced channel spatial attention strategy for fine-grained delineation and multiple-class classifications. Additionally, both architectures incorporate a Segmentation Path (SegPath) module between the encoder and decoder, refining segmentation, enhancing feature extraction, and improving model performance and stability.

Results: AFMS-Net demonstrates exceptional performance across several notable datasets, including BRATs 2021, ATLAS 2021, and ISLES 2022. Its design aims to construct a lightweight architecture capable of handling complex segmentation challenges with high precision.

Discussion: The proposed AFMS-Net addresses the critical balance issue between performance and computational efficiency in the segmentation of brain lesions. By introducing two tailored encoder variants, the network adapts to

varying requirements of speed and feature. This approach not only advances the state-of-the-art in lesion segmentation but also provides a scalable framework for future research in medical image processing.

KEYWORDS

medical image analysis, brain lesion segmentation, adaptive feature extraction, attention mechanism, encoder-decoder architecture, computer-aided diagnosis, deep learning, neurological diagnostics

1 Introduction

Artificial intelligence (AI) in medical imaging has led to a new era in the healthcare system (Hassan et al., 2022). AI-based medical imaging diagnosis facilitates doctors to detect abnormalities earlier, allowing for early control of diseases (Yang and Yu, 2021). One example is the various imaging modalities, such as magnetic resonance imaging (MRI), computed tomography (CT), and ultrasound machines, which enable detailed visualization of structures within the body (Hurlock et al., 2009). To fully utilize these abilities, the detailed medical image segmentation (MIS) process requires careful marking of organs and lesions, slice by slice. This step is essential in radiology, particularly for identifying and monitoring disease conditions. It is a big challenge due to the varied nature of brain lesions and stroke data, the complex structure of the brain itself, as well as significant amounts of MRI and CT scans (Siuly and Zhang, 2016). The precision of segmentation has an impact on diagnosing, treating, and combating nervous system disorders, which account for many deaths around the world (Stoyanov et al., 2018). In recent years, Deep Learning (DL) techniques have greatly simplified medical segmentation. Consequently, there is more research into automating brain lesion detection and segmentation (Wang et al., 2022; Ma et al., 2023). Because of such technological progress, manual and semi-manual are greatly improved. These improved experiences resulted in earlier interventions and better patient results.

Advances in DL approaches have greatly improved the segmentation of medical images, providing significant performance and adaptability to different medical image applications (Greenspan et al., 2023). However, using these methodologies can also pose several challenges. Due to most DL networks' intricate layers and parameters, training takes a long processing time and computational cost. Additionally, consider applying these approaches in a specific imaging situation, such as a brain lesion with split pixel imbalances and complex structures. The segmentation process becomes more complex and less efficient (Shatnawi et al., 2018). Considering these minor errors can significantly affect the performance of these techniques, designing and configuring them for specific problems needs a high level of expertise (Li et al., 2020). Image modalities, imagine size, voxel spacing, and class ratio can all have a substantial impact on 3D medical imaging performance (Vedaei et al., 2023). In addition, to effectively use these approaches, memory requirements, processing capability, and task-specific expertise must be addressed (Celaya et al., 2022).

To address these issues in 3D medical images, we propose the Adaptive Feature Medical Segmentation Network (AFMS-Net). AFMS-Net consists of two encoder modules: Single Adaptive Encoder Block (SAEB) and Dual Adaptive Encoder Block (DAEB). Both versions aim to improve feature extraction and model interpretation. SAEB uses a squeeze-and-excite technique to improve feature representation while reducing model parameters. It is ideal for initial screenings and applications where computational efficiency is a priority. Conversely, DAEB integrates advanced attention mechanisms to capture local and global features, resulting in a comprehensive and precise representation of feature information. The DAEB is designed to address multi-class segmentation challenges datasets such as BRATS, where accurate segmentation with fine-grained and multiple-class labels is essential. This module is particularly useful in cases involving multi-class lesions, where the size, shape, and location of each lesion may significantly influence the diagnosis and treatment plan. Then, incorporate a novel SegPath module between the encoder and the decoder to eliminate the semantic gap and boost feature refinement. The AFMS-decoder utilizes simple convolutional layers and transpose layers to illustrate the respective encoder's features. The proposed AFMS-Net strikes the stability between computational efficiency and segmentation performance, demonstrating impressive findings across three diverse medical datasets in single and multiclass segmentation tasks. Therefore, the suggested segmentation framework shows a significant benchmark for future research in medical image diagnosis.

The key contributions of our research are summarized as follows:

1 We designed an encoder-decoder framework called the Adaptive Feature Medical Segmentation Network (AFMS-Net) framework for brain lesion segmentation.
2 We propose two different encoder modules, a Single Adaptive Encoder Block (SAEB) and a Dual Adaptive Encoder Block (DAEB). SAEB, designed for efficiency, employs a Squeeze-and-Excitation mechanism to capture sufficient primary features from the input images. In contrast, DAEB, is embedded in our AFMS-Net targeting complex cases like BRATS, uses a detailed attention mechanism that considers advanced channel-wise and spatial data.
3 The strategic placement of the new SegPath between the network's encoder-decoder modules addresses the problem of gradient vanishing, boosting feature refinement, and aggregation for enhanced segmentation features. The introduction of an AFMS decoder illustrates the respective encoder's features.
4 Comprehensive experimental analysis was conducted across three standard MIS datasets (BraTS, ALTAS, and ISLES), and seven different state-of-the-art approaches were compared. Our findings show that the AFMS-Net's robust performance

and generalization capability across different datasets emphasize its potential as a new benchmark for segmenting medical images based on standard evaluation metrics.

2 Related work

2.1 Brain lesion segmentation

There has been significant progress in brain lesion segmentation and advanced imaging techniques in recent years. However, accurate segmentation still poses a challenge. Traditional approaches mainly incorporate model-driven techniques, which rely on handcrafted features such as intensity distributions, gradients, morphological attributes, and texture characteristics. Using a voxel probability estimation approach, Anbeek et al. (2004) segmented white matter lesions from brain MRI images. Furthermore, Gooya et al. (2012) combined multi-channel MRI with probabilistic models to show the adaptability of conventional techniques. Moreover, Islam et al. (2013) presented an advanced method of brain tumor segmentation based on spatial and intensity characteristics.

Recently, deep learning has made a significant contribution to brain lesion segmentation. Numerous automated techniques have been proposed, including fully-supervised, supervised unsupervised, and atlas-based methods. So far, convolutional neural network (CNN) based deep learning techniques have demonstrated exceptional performance in medical imaging. The U-Net (Ronneberger et al., 2015) model's efficient encoder-decoder structure has become a starting point for many advanced medical segmentation methodologies. Çiçek et al. (2016) expanded the U-Net architecture into 3D to handle the volumetric data. Based on U-Net, Zhou et al. (2018) developed nested U-Net (Unet++), which minimizes the loss of semantic information between the encoder and decoder. During the 2018 BRATS challenge, Myronenko (2019) proposed a densely connected convolutional blocks auto-encoder model for enhanced brain tumor segmentation. Huang et al. (2020) introduced a full-scale skip connection method through the integration of high-resolution and low-resolution data at various scales. In the Double U-Net network (Guo et al., 2021), two U-Net networks are sequentially organized, in which an Atrous Spatial Pyramid Pooling (ASPP) is placed after every down sample layer in the encoder. In the evaluation, Double U-Net segments nuclei and lesion boundaries well. A gradient vanishing problem has been observed during the converging process of deeper networks. To overcome this problem, Limonova et al. (2021) developed the ResNet-like architecture model. As a contribution to this growing research, Isensee et al. (2021) developed nnU-Net, a self-configuring method for medical image segmentation that adapts based on the provided dataset. According to Rashid et al. (2021) deep learning can automatically segment cerebral microbleeds from structural brain MRI scans. Furthermore, Kermi et al. (2022) developed a multi-view CNN combining the advantages of 2D and 3D networks for glioma segmentation. These findings highlight the various and constantly developing uses of deep learning for medical image segmentation. This research aims to gradually increase segmentation performance, boost efficiency, and address specific issues related to lesion patterns across various illnesses.

Despite all of the advancements made, some issues still need to be resolved in this field. Precisely identifying lesion boundaries remains a challenge for appropriate diagnosis and treatment planning. Secondly, the class imbalance issue often leads to suboptimal model performance in medical imaging datasets, where lesions are considerably smaller than the non-lesion areas. In addition, multi-class lesions, where a single brain scan might reveal several different types of lesions that must be segmented concurrently, remain an open issue. The aim should be to overcome these challenges to design more accurate, effective, and reliable techniques for brain lesion segmentation. The proposed framework addresses these issues using an advanced attention-based deep-learning approach.

2.2 3D attention mechanism in medical imaging data

Attention mechanisms recently gained popularity in computer vision, particularly in medical image segmentation (Gao et al., 2023). This technique, which is well-known for its precise feature selection, enhances the effectiveness of CNNs for a wide range of complex tasks, including detection and classification problems. Squeeze-and-Excitation Network (SENet) (Hu et al., 2018) is a well-illustrated example of an attention mechanism. In SENet, Squeeze-and-Excitation modules determine how feature map channels interact to gather global spatial information. Inspired by SENet, Oktay et al. (2018) designed attention U-Net architecture. This approach reduced the need for extra computational resources or model parameters by accurately targeting regions and highlighting valuable features using a novel bottom-up attention gate. As the field progressed, more sophisticated models began to emerge. Wang et al. (2019) introduced the Volumetric Attention (VA) mechanism, capable of creating 3D enhanced attention maps across spatial and channel dimensions, specifically targeting areas of interest like liver tumors in CT scans. Taking a different approach, Zhang et al. (2020) developed employing attention guidance to enhance segmentation decoders' ability to perceive 3D contexts. Mou et al. (2021) proposed self-attention mechanism, particularly effective in segmenting curved structures such as nerves and blood vessels. This proposal opened new avenues for future research and advancements in the field. In the most recent developments, Zeng et al. (2023) introduced the Multi-Scale Reverse Attention modules (MSRAM) to capture fine-grained features in 3D brain vessel images at different scales. Several promising methods (Nie et al., 2022; Mehrani and Tsotsos, 2023) have developed due to the advancement of attention mechanisms in 3D medical image segmentation. As the field progresses, we optimize existing attention architectures and propose a lightweight, enhanced attention-based model to segment 3D medical images precisely.

3 Methodology

3.1 Overall architecture

We introduced two versions of AFMS-Net for segmenting brain lesions using the proposed SAEB, DAEB, SegPath, and decoder, as demonstrated in Figure 1. Different encoders (SAEB/DAEB) are used in each version, allowing for capturing global and local feature

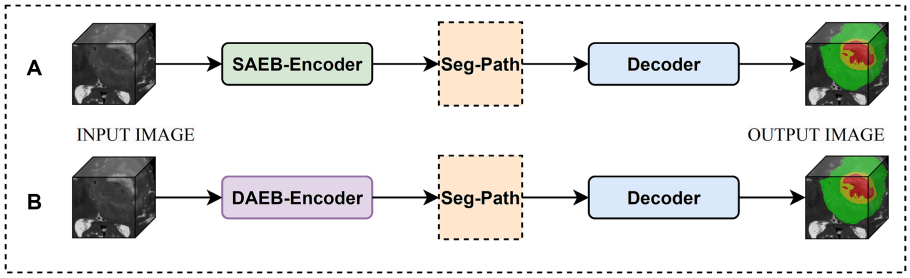


FIGURE 1
Overview of the AFMS-Net. **(A)** Encoder-decoder with SAEB. **(B)** Encoder-decoder with DAEB.

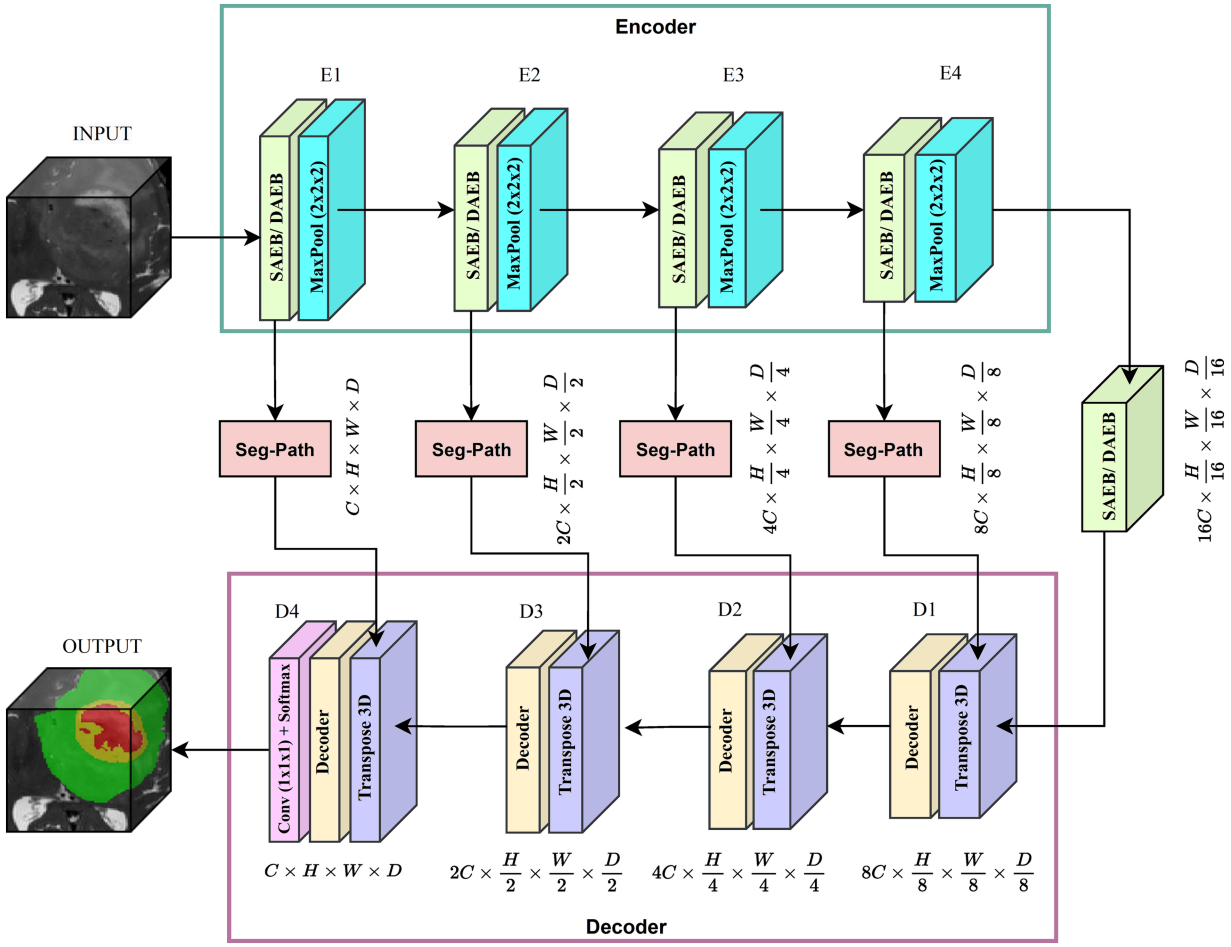


FIGURE 2
Proposed brain lesion segmentation pipeline. Adaptive encoder, SegPath and decoder.

information, enhancing the network's representative ability and feature extraction process. Both versions follow the encoder-decoder structural design illustrated in Figure 2. The SAEB encoder block first uses the Squeeze-and-Excitation (SE) block to extract low-level features. It achieves this by recalibrating channel responses, thereby highlighting crucial details. Additionally, the fusion of $3\times3\times3$ convolutional along with $1\times1\times1$ convolutions serves to synthesize these features, further refining the high-level feature understanding. The DAEB module applies a dual-attention mechanism that emphasizes meaningful semantic features.

Initially, channel-wise attention is achieved through Global Average Pooling (GAP), reshaping, and convolutional layers. This approach enables the network to highlight features in specific channels selectively. The network then learns to focus on essential spatial regions by processing max-pooled and average-pooled information through a convolutional layer. Combining these two attention mechanisms results in a more focused and relevant feature map highlighting channel-specific and spatial information. Each SAEB and DAEB is followed by a $3\times3\times3$ max pooling with stride 2 for a down-sampling operation. The SegPath module is strategically

placed between the encoder and decoder, addressing gradient vanishing and increasing feature refining and aggregation for improved segmentation features. The AFMS-Net decoder gradually up samples the feature maps obtained by the encoder to correspond with the resolution of the input image. The final output of the AFMS-Net is a segmentation probability map obtained from a 3D convolutional layer followed by a softmax activation function, accurately identifying brain lesions. The distinguishing feature of AFMS-Net is its dynamic feature refinement, ensuring superior model results while maintaining computational efficiency. The two versions of the model allow us to evaluate and compare the efficiency and effectiveness of SAEB and DAEB in brain lesion segmentation. More detailed information about the components and operations of AFMS-Net are provided in the subsequent sections.

3.2 AFMS-Net encoder

3.2.1 Single Adaptive Encoder Block

Medical image analysis presents unique challenges that require efficient and robust network architectures. While several network architectures like MobileNet (Howard et al., 2017), EfficientNet (Tan and Le, 2019), and PocketNet (Celaya et al., 2022) have contributed valuable approaches to handling complex features, they often grapple with a trade-off between performance and computational efficiency. Such as, Deeplabv3 (Yurtkulu et al., 2019) captures complex image features that demand significant computational resources. Deeplabv3 parallel convolutional pathways handle multi-scale features but at the cost of a complex architecture and high parameters count. MobileNet and EfficientNet introduced solutions used depth-wise separable convolutions and compound scaling. However, the goal for optimal efficiency and real-time processing continues.

In response to these challenges, proposed network balance computational efficiency with the capacity for effective feature extraction in medical image analysis. Inspired by the Squeeze-and-Excitation (SE) mechanism, SAEB begins the feature extraction process with a single 3D convolution layer. This approach initiates the feature extraction process with a single 3D convolution layer. An intermediate GAP operation follows, leading to the application of two $1\times1\times1$ convolution layers. These layers act as channel-wise transformation agents within the SE mechanism, effectively managing dimensionality reduction and restoration. Figure 3 illustrates the transformations and operations performed within the SAEB, which are especially useful when interpreting complex patterns, such as segmenting brain lesions. The integration of $3\times3\times3$ and $1\times1\times1$ convolutions synthesizes and refines features, enhancing the model's high-level feature understanding and representative ability.

For instance, we initiate this discussion with the examination of the 3D convolution layer which allows the model to handle the width, height, and depth dimensions of the input data, which is crucial in medical image analysis. Mathematically, the convolution operation involves an input tensor $X \in \mathbb{R}^{(H\times W\times D\times C)}$ and filter $F \in \mathbb{R}^{(H\times W\times D\times C)}$, where each position (i,j,k) in the output feature map $Y \in \mathbb{R}^{(H'\times W'\times D'\times C')}$ is computed as follows.

$$Y(i,j,k,c') = \sum_{a=0}^{h-1}\sum_{b=0}^{w-1}\sum_{c=0}^{d-1}\sum_{d=0}^{C-1} X(i+a,j+b,k+c,d)\times F(a,b,c,d,c') \quad (1)$$

In Eq. 1, $Y(i,j,k,c')$ represents the value at the position (i,j,k,c') in the output tensor Y. The four nested summations are indexed by variable a,b,c and d iterate over the ranges $[0,h-1][0,w-1][0,d-1]$ and $[0,C-1]$ respectively. These indices correspond to the spatial dimensions and channels of the input tensor X. The $X(i+a,j+b,k+c,d)$ represents the value at the position $(i+a,j+b,k+c,d)$ in the input tensor X and $F(a,b,c,d,c')$ describes the learnable parameters of the convolutional filter, where c' denotes the output channel index. The SAEB incorporates a Batch Normalization (BN) operation to ensure model stability and efficient training. BN normalizes the input feature maps, mitigating the issue of internal covariate shift and improving model stability and performance. The BN operation calculates the batch mean $E(Y)$, variance $\mathrm{Var}(Y)$ and utilizes learnable scale (γ) and shift (β)

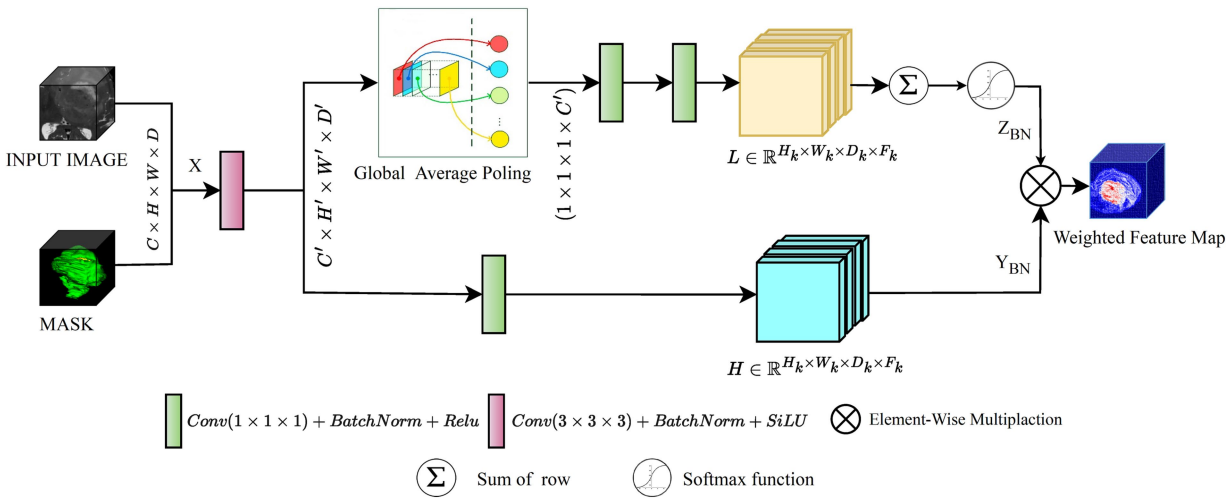


FIGURE 3
An illustration of the proposed SAEB module, the yellow rectangles representing low-level features and the blue rectangles representing high-level features.

parameters to produce batch-normalized output Y_{BN}. The following combined equation can represent the BN operation.

$$Y_{BN} = \gamma \times \left(\frac{Y - E(Y)}{\sqrt{\mathrm{Var}(Y) + \in}} \right) + \beta \tag{2}$$

In Eq. 2, initially, the batch mean $E(Y)$ is calculated as the mean of the input tensor across the mini-batch for each channel, ensuring the normalization process considers the distribution of inputs, as formalized in Eq. 3

$$E(Y) = \frac{1}{m} \sum_{i=0}^{m-1} Y[i] \tag{3}$$

Following the computation of the $E(Y)$, the batch variance $\mathrm{Var}(Y)$ is calculated as the average of the squared differences between each element in the mini-batch and the batch mean, as described by Eq. 4.

$$\mathrm{Var}(Y) = \frac{1}{m} \sum_{i=0}^{m-1} \left(Y[i] - E(Y) \right)^2 \tag{4}$$

Eq. 5, describes how the normalized output $\hat{y}$ is obtained by subtracting the batch mean from the input tensor Y and dividing it by the square root of the batch variance plus a small constant $\in$ for numerical stability.

$$\hat{Y} = \gamma \times \left(\frac{Y - E(Y)}{\sqrt{\mathrm{Var}(Y) + \in}} \right) \tag{5}$$

The final batch-normalized output Y_{BN} is obtained by scaling the normalized output $\hat{y}$ with the learnable scale (β) and shift parameters as depicted in Eq. 6. This step customizes the normalization to the specifics of the data being processed.

$$Y_{BN} = \left(\gamma \times \hat{Y} \right) + \beta \tag{6}$$

following the BN, the SAEB applies a GAP operation to the batch-normalized output Y_{BN} as encapsulated in Eq. 7, which summarizes the presence of each feature across the spatial dimensions, resulting in a tensor $S \in \mathbb{R}^{(1\times1\times1\times c')}$ that captures the global information of the feature maps. The $(c\mathrm{th})$ element of S can be expressed as:

$$[S_c] = \frac{1}{H' \times W' \times D'} \sum_{i=0}^{H'-1} \sum_{j=0}^{W'-1} \sum_{k=0}^{D'-1} Y_{BN}[i,j,k,c] \tag{7}$$

The GAP operation provides a global summary of each channel, capturing the overall presence of features across the spatial dimensions. Following the GAP operation, the SAEB applies a reshape operation to transform the GAP output into a suitable shape for subsequent operations. It is then passed through two $1 \times 1 \times 1$ convolutions to perform channel-wise transformations. The first $1 \times 1 \times 1$ convolution reduces the number of channels, while the second $1 \times 1 \times 1 \times 1$ convolution restores the original number of channels. The softmax activation operation is then applied to generate attention weights $A[c]$ that represent the importance assigned to each channel. This operation calculates a probability distribution across the channel dimension, yielding attention weights $A[c]$ given as follows:

$$A[c] = \frac{\exp(S(c))}{\sum_{d=0}^{C'-1} \exp(S(d))} \tag{8}$$

In Eq. 8, $S(d)$ denotes the value of the GAP output at the $(d\mathrm{th})$ channel. These attention weights $A[c]$ derived from $S(d)$ are pivotal for recalibrating the feature responses. As illustrated in Eq. 9, these weights are applied by element-wise multiplying with the batch-normalized output feature maps Y_{BN}, resulting in the recalibrated feature map Z_{BN}. This step is crucial for enhancing the network's focus on pertinent features within the data.

$$Z_{BN}[i,j,k,c] = A[c] \times Y_{BN}[i,j,k,c] \tag{9}$$

Subsequent to recalibration, the SAEB's final output V is generated by applying an activation function $ReLU$ to the recalibrated feature map Z_{BN}, as formulated in Eq. 10. This transformation introduces non-linearity, enabling the extraction of complex patterns from the recalibrated feature map and preparing the model for further processing layers.

$$V[i,j,k,c] = \max\left(0, Z_{BN}[i,j,k,c]\right) \tag{10}$$

The SAEB final output V is a recalibrated version of the initial input feature map based on channel-wise attention mechanism. This process allows the model to focus on the more relevant features of the task at hand. The model is then used as the final output as the input to the next layer. SAEB recalibrates its output feature maps by using attention weights. This technique allows the model to focus on areas of interest and provide contextually relevant information. In time-sensitive clinical settings or with limited computing resources, this model excels at doing rapid initial screenings.

3.2.2 Dual Adaptive Encoder Block

In 3D data processing, Deep learning algorithms present substantial challenges in 3D data processing, such as extracting prominent spatial and channel dimension features. Primarily, CNN models relied heavily on typical convolution operations and activation functions, which frequently fail to highlight the most critical regions of interest within the data. Attention approaches have emerged as practical solutions that focus on more significant features dynamically. Among these attention methods, Hu et al. (2018) introduced Squeeze-and-Excitation (SE) attention, which plays a critical role in recalibrating channel-wise elements of data. This technique is effective but overlooks the spatial dependencies within feature maps. To address this problem, Woo et al. (2018) proposed spatial attention mechanisms, further refined by Li et al. (2020). However, these approaches largely neglect the interaction between

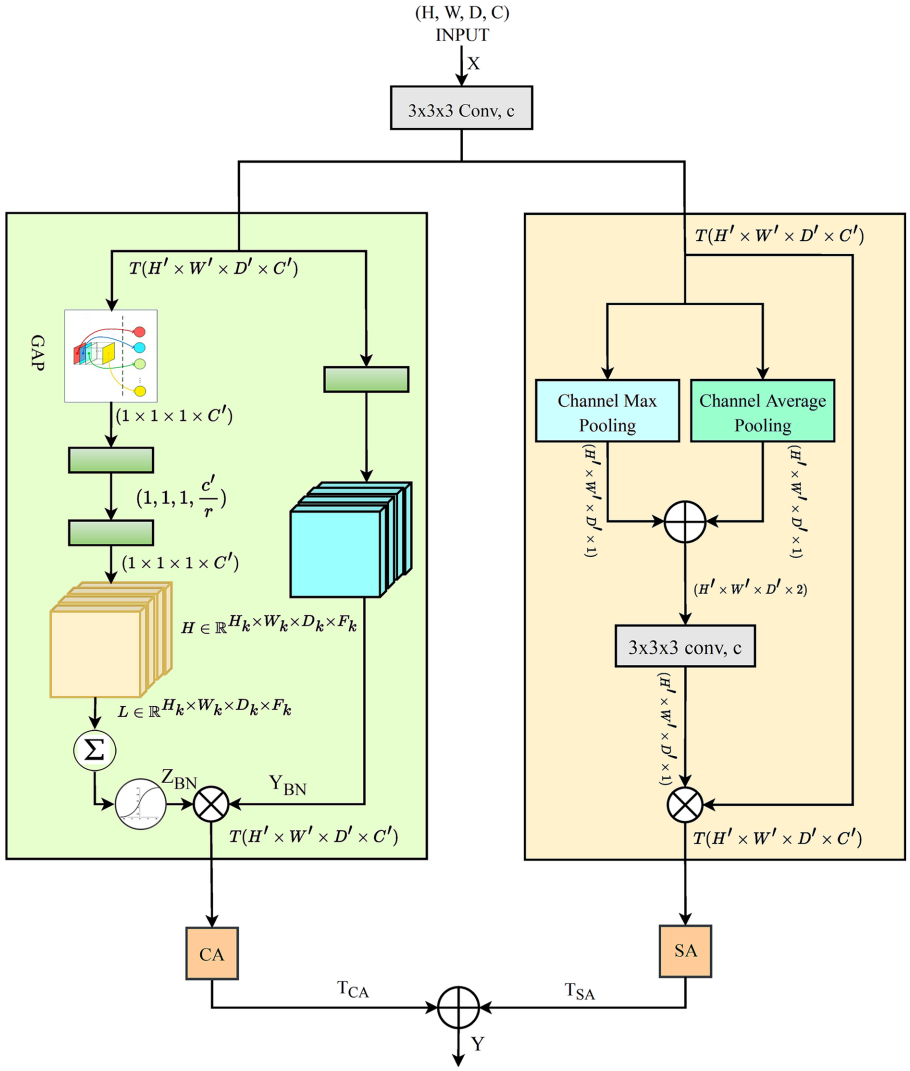


FIGURE 4
The architecture of the proposed DAEB.

channel-wise dependencies. This oversight reveals a compelling opportunity: by integrating both channel-wise and spatial dependencies, model performance could be significantly enhanced. Recognizing this potential, we introduced the DAEB as a proposed solution. The DAEB presents a dual-attention mechanism that significantly extends the suggested network's capability to highlight fine-grain semantic features. By applying channel-specific and spatial attention mechanisms, the DAEB module offers a comprehensive approach to feature refinement. This dual attention is achieved through the integration of global average-pooled information and subsequent convolutional layer processing, which ensure a more focused and relevant feature map. A visual representation of the DAEB and its operations is shown in Figure 4.

We start by applying a 3D convolution operation, denoted by the function F, to the input tensor X, where $X \in \mathbb{R}^{(H \times W \times D \times C)}$ and H, W, D, and C represent the height, width, depth, and channel dimensions of the tensor, respectively. This operation transforms X into an intermediate tensor T, and the transformation can be denoted as:

$$T(i',j',k',c') = \sum_{i_n} \sum_{j_n} \sum_{k_n} \sum_{k} W(i,j,k,c,c') \times X(i' + i_n, j' + j_n, k' + k_n, c) \quad (11)$$

In Eq. 11, the variables (i',j',k',c') represent coordinates in the new tensor T, and the non-primed ones (i,j,k,c) represent coordinates in the original tensor X. The variables (i_n, j_n, k_n) iterate over the kernel dimensions and W represents the kernel weights. This operation extracts localized features from the input tensor X based on the filter weights. Subsequently, we introduce a channel-wise focus through the GAP mechanism, which is applied to the tensor T. This yields the global descriptor $\mathrm{CA} \in R^{C'}$, where C' representing the channels in the transformed tensor:

$$\mathrm{CA}(c') = \left(\frac{1}{(H' \times W' \times D')} \right) \times \sum_{i} \sum_{j} \sum_{k} T(i,j,k,c') \quad (12)$$

In Eq. 12, (H', W', D') represents the height, width, and depth dimensions of T, respectively. The global descriptor CA gives importance to informative channels and suppresses the less relevant ones in tensor T. Then, two-step transformation process is implemented on the global descriptor CA, yielding a new descriptor CA′:

$$\mathrm{CA}'(i',j',k',c'') = \sum_{i_n}\sum_{j_n}\sum_{k_n}\sum_{k} W'(i,j,k,c',c'') \times \mathrm{CA}(i'+i_n, j'+j_n, k'+k_n, c') \quad (13)$$

In Eq. 13, W' represents the transformation weights, and the c'' term denotes the channels in the newly transformed descriptor. This transformation helps to highlight channel-wise dependencies in the global descriptor CA.

$$\mathrm{CA}(i',j',k',c'') = \sigma(\mathrm{CA}'(i',j',k',c'')) \quad (14)$$

The transformation process is further refined by applying a sigmoid activation function (σ) to the descriptor CA′, which generates the channel-wise attention map CA as detailed in Eq. 14, effectively scaling each channel's values within the interval [0, 1]. This step is essential for determining the significance of each channel in terms of the spatial features of the input tensor. After obtaining the channel-wise attention map CA, reweight the tensor T through an element-wise multiplication operation, yielding tensor T_{CA}

$$T_{\mathrm{CA}}(i',j',k',c') = T(i',j',k,c') \times \mathrm{CA}(i',j',k',c'') \quad (15)$$

Eq. 15 describes the application of the channel-wise attention map CA, where the recalibrated tensor T_{CA} is produced by an element-wise multiplication with the tensor T. This operation enables the model to adaptively emphasize informative features and suppress irrelevant ones in the tensor T. Then, we compute the spatial attention map $SA \in \mathbb{R}^{(H' \times W' \times D' \times 2)}$ as illustrated in Eq. 16, by concatenating the maximum and average pooling maps derived from T.

$$\mathrm{SA} = [\mathrm{MaxPool}(T) \oplus \mathrm{AvgPool}(T)] \quad (16)$$

This step captures spatial dependencies in the feature maps. The spatial attention map SA is then transformed through a 3D convolution operation denoted by Conv.

$$\mathrm{SA} = \mathrm{Conv}(\mathrm{SA}) \quad (17)$$

In Eq. 17, the spatial attention map SA undergoes a 3D convolution transformation, which enhances the model's capability to capture spatial dependencies within the feature maps. This convolution operation consolidates the various spatial features into a more coherent structure that is crucial for accurate segmentation.

Following this convolution, Eq. 18, details how the spatial attention map SA scaled by a sigmoid activation function, assigning a value between 0 and 1 to each position. This scaling effectively ranks the spatial features by their relevance. The resulting map is then utilized to modulate the tensor T, with an element-wise multiplication producing the reweighted tensor T_{SA}.

$$T_{\mathrm{SA}} = T \otimes \mathrm{SA} \quad (18)$$

The process, affiliated with channel-wise reweighting, allows the model to emphasize informative features and suppress irrelevant ones adaptively. Finally, we combine the outputs of the channel-wise and spatial attention mechanisms applied separately to the input tensor. The resultant tensors (not the attention maps) are fused to generate the final output tensor Y:

$$Y = T_{\mathrm{CA}} \oplus T_{\mathrm{SA}} \quad (19)$$

Eq. 19 represents the fusion of the channel-wise and spatial attention mechanisms, resulting in the new output tensor Y. The model leverages informative channels and spatially relevant regions by integrating these outputs, thereby effectively understanding and classifying complex multi-dimensional data. The DAEB's dual-attention mechanism addresses the need for extracting prominent features across both spatial and channel dimensions, effectively overcoming the limitations of traditional CNN models that may overlook critical regions of interest within the data. By implementing the DAEB, it is anticipated that models can learn more effectively from 3D data, potentially leading to enhanced performance across various tasks and domains. The DAEB consistently outperformed existing models through rigorous experimental analysis, solidifying its standing as an optimized solution for 3D data segmentation.

3.3 SegPath

Semantic segmentation has various approaches for enhancing the connectivity between encoders and decoders. In this regard, the skip connection is an outstanding solution that has gained recognition, particularly in architectures such as U-Net. This method enables encoder features to be directly associated with corresponding decoder layers, thereby ensuring the preservation and recovery of spatial details, which is vital for accurate segmentation. Merging the encoder features (low-level features) with decoder features (high-level features) results in a semantic gap.

In recent research on connectivity strategies, the ResPath architecture emerged, integrating residual connections reminiscent of the ResNet strategy within the skip pathways. This fusion improves the model's ability to learn refined residual feature representations. Moreover, Mubashar et al. (2022) combines dense and skip connections in a significant way. This architecture ensures that all feature maps are densely connected via a skip connection structure. Drawing from these improvements, we present the SegPath module, a sophisticated modification to the skip connection structure, as shown in Figure 5. SegPath enhances segmentation performance through two fundamental processes: adaptive feature accumulation and the integration of multi-scale contextual information. Adaptive feature accumulation works by iteratively accumulating enhanced feature maps through element-wise addition, enabling SegPath to form a comprehensive representation of the input data. This process allows for the adaptive refinement of feature maps, customized to meet the specific requirements of the segmentation task.

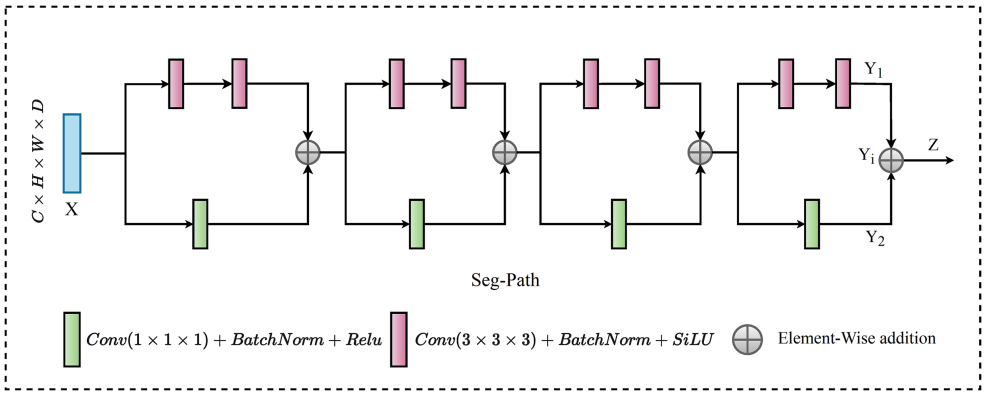


FIGURE 5
The framework of SegPath.

Concurrently, SegPath employs parallel transformations to capture a wide range of aspects from the input feature map, including detailed textures and broader contextual information. These transformations involve convolving the input feature map with filters of different sizes ($1\times1\times1$ and $3\times3\times3$), followed by batch normalization and ReLU activation. It incorporates a series of parallel transformations to capture various aspects of the input feature map X. In the first transformation, X undergoes a $1\times1\times1$ convolution with a filter F_1, resulting in a tensor X_1. This operation can be formulated as shown in Eq. 20,

$$X_1(i,j,k,l) = \sum_m X(i,j,k,m) \times F_1(1,1,1,m) \tag{20}$$

In Eq. 20, i, j and, k are spatial locations in the 3D feature maps and 1 denotes the feature channel at each spatial location. The index m is used to iterate over the feature channels in the input feature map X and the convolution filter F_1. Simultaneously, X is convolved with a $3\times3\times3$ filter F_2, leading to tensor X_2, as expressed in Eq. 21.

$$X_2 = (i,j,k,l) = \sum_m \sum_{a=-1^1} \sum_{k=-1^1} \sum_{x=-1^1} X(i+a,j+b,k+c,m) \times F_2(a+2,b+2,c+2,m) \tag{21}$$

where a,b, c used to traverse the 3D convolution filter's spatial extent during the convolution operation, ranging from -1 to 1 to cover the $3\times3\times3$ spatial extent of the filter F_2. After each convolution, batch normalization is applied to normalize the tensor, creating X_1 normalized and X_2 normalized tensors. Following the normalization step, the ReLU activation function is applied element-wise to X_1 and X_2 normalized, resulting in tensors Y_1 and Y_2, respectively, detailed in Eq. 22 and Eq. 23.

$$Y_1(i,j,k,l) = \max(0, X_1\text{Norm}(i,j,k,1)) \tag{22}$$

and,

$$Y_2(i,j,k,l) = \max(0, X_2\text{Norm}(i,j,k,1)) \tag{23}$$

These enhanced feature maps Y_1 and Y_2 are accumulated through element-wise addition to create an enhanced representation Y_i for each iteration i, as outlined in Eq. 24.

$$Y_i(i,j,k,l) = Y_1(i,j,k,1) + Y_2(i,j,k,1) \tag{24}$$

This step is repeated n times, where $[i = 1,2,\ldots,n]$ and the outputs are summed together to obtain the final output tensor Z, encapsulated in Eq. 25.

$$Z = \sum_{i=1}^{n} Y_i \tag{25}$$

The accumulation of adaptive features enriches the information carried by the final output tensor, Z allowing for better capture of complex patterns and variations in the input data. This is particularly crucial in medical image analysis tasks, where detailed and accurate feature extraction is key to successful segmentation. This approach ensures a general understanding of the samples, significantly improving segmentation outcomes by using the strengths of both detailed and contextual information processing within the model. The adaptive feature accumulation of the SegPath block allows for learning more critical features for the specific task, thus enhancing its representative capacity. Furthermore, it provides an additional path for gradient flow through the adaptive features, improving the mitigation of the vanishing gradient problem.

4 Materials and experimental setup

4.1 Materials

To demonstrate the broad utility and effectiveness of our proposed model, AFMS-Net, we have used three appreciated, publicly accessible datasets, each supporting a distinct medical image segmentation task. The details of these datasets are summarized in Table 1. Specifically, the Brain Tumor Segmentation BRATS2021 dataset (Baid et al., 2021), facilitates brain tumor segmentation. For ischemic stroke lesion identification and tracing of lesions after a stroke, we have employed

TABLE 1 Details of the medical segmentation datasets used in our experiments.

Dataset	Images	Voxel size	Input size	Train	Valid	Test
BraTS 2021	1,151	1×1×1	128×128×128	874	115	162
ATLAS v2.0	655	1×1×1	160×160×160	458	105	092
ISLES 2022	246	2×2×2	128×128×128	196	024	026

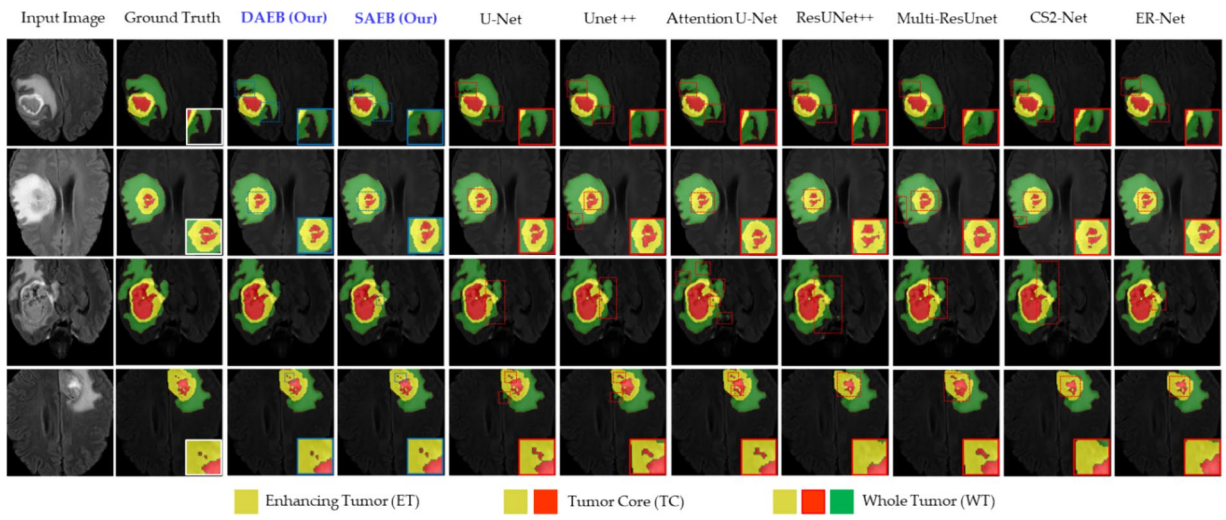


FIGURE 6
Illustrative examples highlighting diverse challenges in the BraTS Dataset. The set of samples is organized from left to right as follows: the original input image, the associated ground truth, and the segmentation outputs generated by our models (SAEB and DAEB), U-Net, Unet++, Attention U-Net, ResUNet++, Multi-ResUnet, CS2-Net, and ER-Net. Four distinct samples demonstrate specific challenges: boundary delineation, fine-grained analysis, and lesion variability. This comprehensive comparison aims to underscore each segmentation technique's relative strengths and limitations in addressing these challenges.

the Anatomical Tracings of Lesions after Stroke (ATLAS v2.0) 2021 datasets (Liew et al., 2017), and the Ischemic Stroke Lesion Segmentation (ISLES) 2022 datasets (Hernandez Petzsche et al., 2022), respectively. In addition, we used rigid registration and affine transformation techniques to register ISLES datasets according to the standard Montreal Neurological Institute (MNI) space (Chau and McIntosh, 2005).

4.1.1 Brain tumor segmentation datasets

The Proposed framework utilized the BraTS-2021 benchmark dataset, which includes a training set comprising 1,251 patients with both High-Grade Gliomas (HGG) and Low-Grade Gliomas (LGG). Each patient dataset consists of four MRI sequences: T1-weighted (T1), contrast-enhanced T1-weighted (T1ce), T2-weighted (T2), and fluid-attenuated inversion recovery (FLAIR). These sequences offer a detailed and multidimensional view of the tumor, aiding in more precise segmentation. Images in the dataset were collected following various clinical guidelines, using MRI machines of differing specifications and magnetic intensities, contributing to its heterogeneity. The image preprocessing steps were critical in ensuring data consistency across all datasets. It included co-registration of each patient's MRI modalities, skull stripping, and voxel resampling to a 1 mm^3 isotropic resolution, resulting in a uniform MRI volume size of 155 × 240 × 240. The ground truth segmentation for each MRI volume was categorized into four segments: background, Necrotic and Non-enhancing Tumor (NCR), Peritumoral Edema (ED), and Enhancing Tumor (ET). However, for evaluation, the three nested sub-regions, namely enhancing tumor (ET), tumor core (TC—i.e., the union of ED and NCR/NET), and whole tumor (WT), are used (see the sample ground truths in Figure 6). In order to enhance computational efficiency and concentrate the suggested model's attention on the most pertinent areas, resized the original volume to dimensions of 128 × 128 × 128.

Moreover, we fused the FLAIR, T1ce, and T2 modalities into a single multi-channel image, which provided proposed framework with the most comprehensive information about each tumor's characteristics. In data preprocessing step, we implemented a filtering mechanism to disregard less informative samples. Specifically, any volume where less than 1% of labels were non-zero (indicative of tumor presence) was deemed "useless" and discarded. This helped reduce noise in the training data, thereby enhancing the learning efficiency of our model. For a comprehensive model evaluation, we systematically divided the data by allocating 80% for model training, allowing the model to learn from diverse information. The remaining 20% was equally divided into validation and test sets. The validation set helped fine-tune our model's hyper-parameters. In contrast, the test set assessed our model's performance on unseen data, providing a more reliable evaluation of its effectiveness.

4.1.2 ATLAS v2.0 dataset

The ATLAS v2.0 dataset, a meticulously composed repository of MRI scans and lesion segmentation masks, has been methodically organized into three subsets: training, testing, and a holdout set. The training subset comprises 655 T1-weighted MRI scans from multiple cohorts, each linked

with its corresponding lesion segmentation mask. The test subset includes 300 T1-weighted MRI scans drawn from the same cohorts, with their respective lesion segmentation masks intentionally hidden. The holdout test set encapsulates 316 entirely obscured T1-weighted MRI scans and lesion segmentation masks, each originating from an independent set. This dataset was utilized strategically through a comprehensive preprocessing pipeline in the experimental process. The initial step involved performing a central cropping operation on the image data to a size of 160 × 160 × 160 voxels. Focusing on the region of interest reduced superfluous peripheral information, thereby enhancing computational efficiency. Standardizing voxel size across the dataset involved resampling the cropped image data, contributing to consistent and reliable outcomes in subsequent machine-learning tasks. The image data was normalized to diminish the impact of intensity variations across different MRI scans. Gaussian smoothing was implemented to mitigate the influence of noise on the MRI scans. This technique not only reduced noise but also augmented the visibility of the lesions, thereby improving detection accuracy. Simultaneously, lesion segmentation masks were resampled to match the size of the corresponding image and converted into a one-hot encoded format, facilitating their integration into subsequent machine-learning tasks. The 655 T1-weighted MRI scans were then divided into training, validation, and testing sub-sets, comprising approximately 70, 16, and 14% samples, respectively. This stratified splitting strategy balanced the representation of different lesion sizes across all subsets, circumventing potential bias in the model training phase. This rigorous approach guarantees the validity and robustness of the experimental procedures.

4.1.3 ISLES 2022 dataset

The ISLES dataset is designed to evaluate automated acute and subacute stroke lesion segmentation methods in 3D multi-modal MRI data. For our experiments, we used a series of preprocessing steps. The dataset consists of DWI, ADC, and FLAIR images. The FLAIR image was registered to the standard Montreal Neurological Institute (MNI) space (Chau and McIntosh, 2005) using an affine transformation, creating a transformation matrix. This transformation matrix was then used to register the DWI and ADC images first to the original FLAIR images and then to the standard MNI space. In other words, the FLAIR image was registered to the DWI space utilizing rigid registration and affine transformation techniques. After registration, the ADC, DWI, and FLAIR data were consolidated into a multi-channel image. Each image was cropped to a size of 128 × 128 × 128, improving computational efficiency by removing non-essential regions. The dataset encompasses a total of 246 samples. To ensure an unbiased evaluation of our developed model, we randomized the data and divided it into training, validation, and testing sets, adhering to an 80-10-10 split.

4.2 Experimental setup

The proposed approach was implemented and trained using the TensorFlow and Keras frameworks, and all experiments conducted on NVIDIA RTX A5000 GPUs. This setup offered the computational power necessary for handling the intensive demands of training deep learning models on complex medical image datasets. The choice of hardware reflects a balance between computational efficiency and the capability to process large volumes of data, characteristic of medical imaging tasks.

4.2.1 Model optimization and hyperparameter selection

Our experimental strategy employment the Adam optimizer, chosen for its effectiveness in handling sparse gradients and adaptively adjusting learning rates, which is crucial for deep learning applications in medical imaging. We set the learning rate to a modest 0.0001, a decision informed by preliminary trials that indicated it as optimal for balancing training speed with convergence stability. Similarly, a weight decay of 0.0005 was applied as a regularization measure to mitigate the risk of overfitting—a common challenge in deep learning models. This weight decay introduces a minor penalty to the loss function, proportional to the L2 norm of the model weights, encouraging the model to learn more generalizable features.

4.2.2 Computational resources and model complexity

Training the AFMS-Net required significant computational resources. Specifically, the training process was executed over approximately 8–16 h on NVIDIA RTX A5000 GPUs, utilizing around 16GB of GPU memory per model instance. These figures highlight the computational demands of training AFMS-Net, emphasizing the need for powerful hardware to achieve optimal performance. To provide a comparative insight into AFMS-Net's model complexity versus traditional segmentation networks, we reference Figure 7, which illustrates the computational performance trade-offs by comparing mIoU with the number of parameters. This comparison reveals that AFMS-Net achieves a commendable balance between model complexity and segmentation performance. Unlike traditional segmentation networks such as U-Net and its variants, AFMS-Net demonstrates enhanced computational efficiency, achieving competitive or superior performance metrics with a reduced number of parameters. This efficiency is pivotal for deploying advanced segmentation models in real-world medical imaging scenarios, where computational resources might be limited.

4.2.3 Custom loss function

A distinctive feature of our experimental setup is the incorporation of a custom loss function that combines dice loss and categorical focal loss. This approach was designed to address the challenges of class imbalance and ensure accurate segmentation across varying medical image characteristics. The Dice loss, formulated in Eq. 26, is particularly effective in promoting overlap between the predicted segmentation maps and the ground truth, thereby enhancing the model's precision in delineating lesion boundaries.

$$L_{\text{Dice}}(G,P) = 1 - \frac{2\left(\sum_c \sum_i w_c\, G_{ci} P_{ci}\right) + \in}{\left(\sum_c \sum_i w_c\, G_{ci} + \sum_c \sum_i w_c P_{ci}\right) + \in} \quad (26)$$

In Eq. 26, G and P are the ground truth and predicted probability map, c denotes each class, i stands for individual voxels, w_c refers to the weight of each class, and $\in$ is a small constant used to prevent division by zero.

Further refining the model's predictive accuracy, the categorical focal loss—described in Eq. 27, adjusts the model's focus towards difficult-to-classify examples, thereby improving overall classification accuracy.

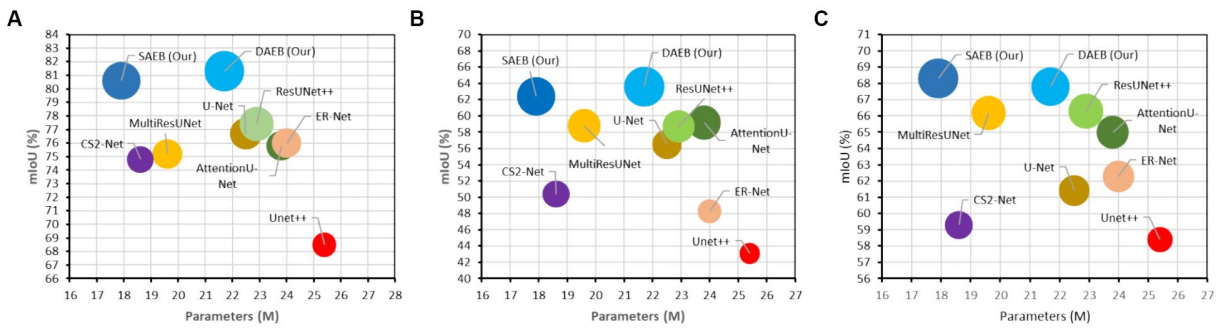


FIGURE 7
Computational performance trade-offs illustrated by mIoU versus the number of parameters of various models across multiple datasets. **(A)** BraTS 2021, **(B)** ATLAS, and **(C)** ISLES 2022.

$$L_{\text{focal}}(G,P) = -\sum_{c}\sum_{i} G_{ci}\log(P_{ci})(1-P_{ci})^{\gamma} \tag{27}$$

where P_{ci} represents the model's estimated probability for the true class, γ is a tuning parameter (typically set at 1.0), and the sum is calculated over all classes. Each class was assigned an equal weight for dice loss calculation.

Ultimately, the total loss utilized for training the model is computed as the sum of the Dice loss and the categorical focal loss, as shown in Eq. 28. This combined loss function leverages the strengths of both components to provide a balanced optimization criterion.

$$L_{\text{total}}(G,P) = L_{\text{Dice}}(G,P) + L_{\text{focal}}(G,P) \tag{28}$$

4.3 Evaluation metrics

This section outlines the key metrics used to assess the model's effectiveness comprehensively. The proposed brain lesion segmentation model is rigorously evaluated using a comprehensive set of metrics, all at a threshold of 0.5, to provide a thorough understanding of its performance. Accuracy is calculated as the proportion of true predictions, both correct lesion identifications (true positives) and correct non-lesion identifications (true negatives), over the total number of cases, as specified in Eq. 29.

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \tag{29}$$

where TP are true positives, TP are true negatives, FP are false positives, and FN are false negatives. Precision, defined as the ratio of true positives to the sum of true positives and false positives, reflects the model's accuracy in predicting lesion instancesinstances, outlined in Eq. 30.

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{30}$$

Meanwhile, Recall measures the model's ability to identify all actual lesion cases, calculated as the ratio of true positives to the sum of true positives and false negatives, as depicted in Eq. 31.

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{31}$$

The Dice Score (DSC) expressed in Eq. 32, is used to measure the similarity between the predicted segmentation and the ground truth. It is particularly useful for evaluating models where the class distribution is imbalanced. The DSC is calculated as:

$$\text{DSC} = \frac{2\text{TP}}{2\text{TP}+\text{FP}+\text{FN}} \tag{32}$$

Intersection over Union (IoU), presented in Eq. 33, also known as the Jaccard index, measures the overlap between the predicted segmentation and the ground truth. It is defined as:

$$\text{IoU} = \frac{\text{TP}}{\text{TP}+\text{FP}+\text{FN}} \tag{33}$$

The average Hausdorff distance (AHD), uniquely considering voxel location and defined as:

$$\text{AHD} = \frac{1}{2}\left(\frac{1}{P}\sum_{p\in P}\min_{l\in L} d(p,l) + \frac{1}{L}\sum_{l\in L}\min_{p\in P} d(p,l)\right) \tag{34}$$

In Eq. 34, P represents the point set of segmentation results, and L denotes the point set of labels, enabling reflection on the edge error of segmentation results. These metrics provide a balanced and comprehensive assessment of the efficacy of the suggested framework in brain lesion segmentation.

5 Experimental results

5.1 Comparative segmentation performance on diverse datasets

In this section, we present a thorough comparison between seven different state-of-the-art 3D MIS techniques and our suggested approach for brain lesion segmentation. We compare our approach with U-Net, Unet++ (Zhou et al., 2019), AttentionU-Net, ResUNet++

(Jha et al., 2019), Multi-ResUNet (Ibtehaz and Rahman, 2020), CS2-Net (Mou et al., 2021) and ER-Net (Xia et al., 2022). We follow a uniform protocol across all methodologies to ensure a fair and comprehensive comparison. Every baseline model follows the default settings specified by their respective original authors. The structure of each model is based either on the associated codes available on GitHub or descriptions provided by the original authors. We also maintain consistency in preprocessing and post-processing steps across all models. This standardization eliminates potential bias, ensuring the comparative results accurately reflect the performance of each method.

5.1.1 Qualitative and quantitative results on BraTS 2021 dataset

In this section, we evaluate the performance of SAEB and DAEB models on the BraTS 2021 dataset. The qualitative results, illustrated in Figure 6, offer visual insights into the performance of various segmentation methods—the first-row centers on the model's proficiency in edge detection within tumors. Certainly, most approaches perform similarly well in distinguishing important regions of enhancing tumor (ET), tumor core (TC), and whole tumor (WT). However, differences become noticeable when defining the edges of the tumor. Selected comparison methods, such as U-Net, Unet++, Attention U-Net, ResUnet++, Multi-ResUnet, CS2-Net, and ER-Net, effectively detect larger tumor structures but falter when identifying precise edges. This results in noticeable under-segmentation or over-segmentation. In comparison, The SAEB and DAEB models precisely outline the tumor edges. The blue-red, dotted rectangles and their magnified views highlight the differences. The second row demonstrates the proficiency of SAEB and DAEB in recognizing intricate tumor sub-structures. In contrast, notable methods like U-Net and its variants misrepresent subtle elements such as necrosis or non-enhancing tumor cores. The third row of Figure 6 illustrates the ability of the SAEB and DAEB to emphasize the uniformity of regions within the tumor while simultaneously identifying subtle variations in texture. In the fourth row, we address the fine-grained analysis problem. Interestingly, all the baseline approaches failed to identify these tiny features. However, the suggested framework can identify minute structures and lesions. The visualizations demonstrate the proposed models' adaptability and precision, highlighting their ability to tackle the intricate challenges presented by the BraTS dataset. For the detailed quantitative analysis, this work is divided into two main sections: overall segmentation performance, presented in Table 2, and segmentation by tumor regions, illustrated in Table 3; this comprehensive analysis evaluates the proposed framework's effectiveness. The evaluation metrics in Table 2 reinforce the superior performance of our proposed models. Our AFMS-DAEB registers impressive results with an accuracy of 99.01%, precision of 90.80%, recall of 89.06%, DSC of 90.20%, mIoU of 82.23%, and an AHD value of 6.079, respectively. These metrics indicate an approximate 1% enhancement in DSC and mIoU over the AFMS-SAEB model.

When benchmarked against state-of-the-art models, our models exhibit a considerable edge. While U-Net, with its 86.7% DSC and 76.7% mIoU, is commendable, it's surpassed by AFMS-DAEB, particularly in DSC and mIoU. Unet++ shows room for improvement, especially with its 81.2% DSC. Attention U-Net and ResUNet++ deliver DSC values around 85%, yet are outperformed by our models. Similarly, despite their respective merits, Multi-ResUNet, C2Net, and ErNet fall short compared to AFMS-DAEB's segmentation efficacy. In principle, AFMS-DAEB not only refines the capabilities of AFMS-SAEB but also delineates itself as a potent tool among established segmentation techniques, showcasing its aptitude for nuanced medical image segmentation. Further examining the segmentation performance across models, we assess three critical tumor categories: WT, TC, and ET, as detailed in Table 3. In the WT segmentation, our AFMS-SAEB model emerges as a front-runner, boasting an accuracy of 98.9%, a precision of 91.3%, a DSC of 88.4%, and mIoU of 79.2%. These metrics showcase the superiority of some models over others. For instance, U-Net achieved a DSC of 87.4% and a mIoU of 77.7%.

On the other hand, Unet++ is behind with a DSC of 45.4%, while AttentionU-Net and ResUNet++ have better results, with mIoU scores of 73.2 and 73.5%. The mIoU score of SAEB is 79.2%, which matches closely with the ground truths. For TC, DAEB has an excellent performance. Its accuracy is 99%, precision is 88.6% and DSC score reaches up to 87.5% and mIoU value of around 77%. For the ET, AFMS-DAEB showcases a commendable DSC of 85.1% and a mIoU score of 74%. Compared to other base models, ER-Net and MultiResUNet demonstrate promising results. In conclusion, based on evaluation metrics, SAEB and DAEB show promising tumor segmentation capabilities. The combination of insights from both tables provides a comprehensive evaluation of each model's segmentation performance and specialization inside various tumor locations.

5.1.2 Qualitative and quantitative results on ATLAS R2.0 dataset

Precise lesion segmentation can significantly aid stroke diagnosis and treatment. Our proposed method demonstrates this precision across four diverse stroke cases, which are visually presented in Figure 8. These cases

TABLE 2 Performance metrics of various methods evaluated on 1,251 cases from the Brats 2021 dataset.

Method	Accuracy	Precision	Recall	DSC	mIoU	AHD
U-Net	0.988 ± 0.003	0.870 ± 0.101	0.865 ± 0.178	0.867 ± 0.123	0.767 ± 0.165	7.010
Unet++	0.927 ± 0.031	0.747 ± 0.187	0.891 ± 0.165	0.812 ± 0.157	0.685 ± 0.176	12.146
AttentionU-net	0.986 ± 0.022	0.899 ± 0.153	0.883 ± 0.127	0.860 ± 0.179	0.758 ± 0.181	8.725
ResUNet++	0.988 ± 0.020	0.882 ± 0.148	0.889 ± 0.186	0.869 ± 0.158	0.774 ± 0.161	6.916
MultiResUNet	0.985 ± 0.025	0.848 ± 0.183	0.868 ± 0.121	0.856 ± 0.153	0.752 ± 0.158	9.125
CS2-Net	0.985 ± 0.033	0.853 ± 0.141	0.858 ± 0.116	0.855 ± 0.172	0.748 ± 0.187	10.165
ER-Net	0.987 ± 0.025	0.860 ± 0.161	0.866 ± 0.194	0.861 ± 0.145	0.761 ± 0.209	8.126
SAEB (Our)	0.989 ± 0.026	0.913 ± 0.118	0.885 ± 0.171	0.894 ± 0.123	0.806 ± 0.117	6.266
DAEB (Our)	0.990 ± 0.031	0.908 ± 0.189	0.896 ± 0.132	0.902 ± 0.151	0.813 ± 0.195	6.079

TABLE 3 Comparative performance metrics for whole tumor (WT), tumor core (TC), and enhancing tumor (ET) in 1,251 cases from the Brats 2021 dataset.

Model	Whole tumor					Tumor core					Enhancing tumor				
	ACC	PRE	REC	DSC	IoU	ACC	PRE	REC	DSC	IoU	ACC	PRE	REC	DSC	IoU
U-Net	0.98	0.89	0.85	0.87	0.77	0.98	0.81	0.84	0.85	0.74	0.98	0.86	0.71	0.79	0.66
Unet++	0.92	0.84	0.81	0.82	0.70	0.92	0.29	0.55	0.45	0.29	0.92	0.83	0.69	0.75	0.60
AttentionU-Net	0.98	0.74	0.87	0.80	0.67	0.98	0.83	0.85	0.84	0.73	0.98	0.86	0.79	0.82	0.70
ResUNet++	0.98	0.85	0.87	0.86	0.76	0.98	0.79	0.83	0.84	0.73	0.98	0.90	0.74	0.81	0.68
MultiResUnet	0.98	0.81	0.84	0.83	0.71	0.98	0.73	0.87	0.81	0.68	0.98	0.82	0.79	0.81	0.68
CS2-Net	0.98	0.89	0.84	0.86	0.76	0.98	0.80	0.84	0.84	0.73	0.98	0.83	0.71	0.78	0.65
ER-Net	0.98	0.89	0.85	0.87	0.77	0.98	0.81	0.84	0.85	0.74	0.98	0.85	0.74	0.79	0.66
SAEB (Our)	0.98	0.91	0.85	0.88	0.79	0.98	0.86	0.85	0.85	0.75	0.98	0.87	0.78	0.83	0.71
DAEB (Our)	0.99	0.84	0.91	0.87	0.78	0.99	0.88	0.86	0.87	0.77	0.99	0.87	0.83	0.85	0.74

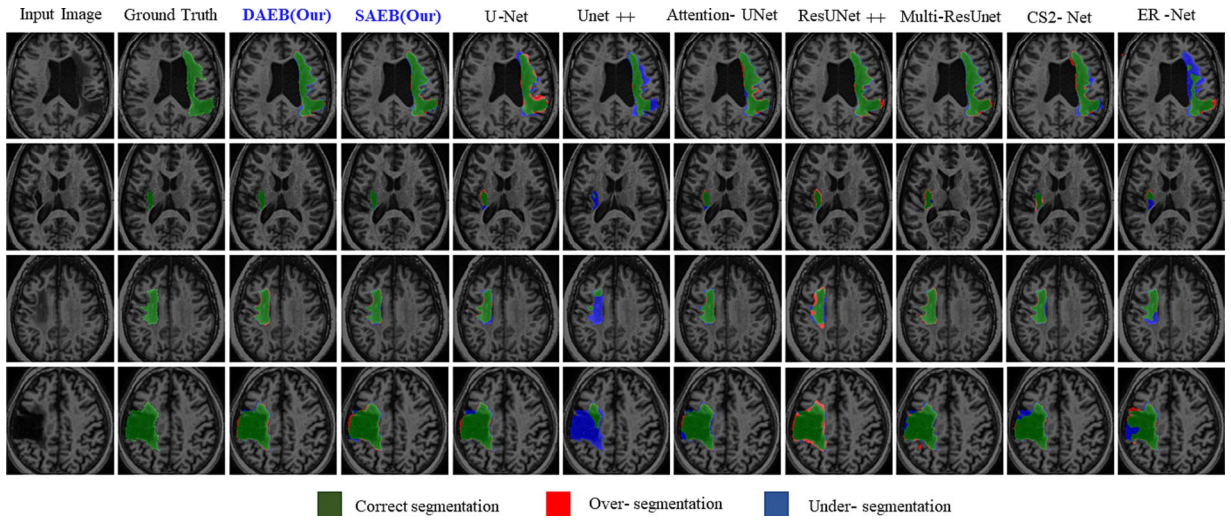


FIGURE 8
Visual comparison of segmentation challenges in four representative samples from the ATLAS R2.0 dataset. Arranged from left to right are: The original input image, the ground truth, followed by outputs from our models (DAEB, SAEB), U-Net, Unet++, Attention U-Net, ResUNet++, Multi-ResUNet, CS2-Net, and ER-Net. The samples are chosen to highlight distinct challenges inherent to the ATLAS dataset.

vary in lesion location, shape, and size within the brain, highlighting the adaptability of our approach. In the first row, the lesion is located in the anterior limb and genu of the internal capsule. AFMS-DAEB and AFMS-SAEB, predict almost the entire lesion completely, achieving a remarkable advantage over the benchmark models. While U-Net, Attention U-Net, ResUNet++, and Multi-ResUNet manage to identify most of the lesions, but they tend to over-segment the affected area.

On the other hand, Unet++, C2Net, and ER-Net only delineate a small fraction of the lesion. The second row examine a lesion in the internal capsule's posterior limb. Here, Unet++ and ER-Net struggle to mark the lesion accurately. U-Net and AttentionU-Net identify only portions of it. Although closer to the mark, ResUNet++, Multi-ResUNet, and CS2-Net present evident over-segmentations. However, the proposed framework captures this lesion clearly, highlighting its adeptness at processing boundary information. In the third row, the lesion, with its regular shape and precise location, presents a more straightforward segmentation target. Both AFMS-Net variants demonstrate superior performance in delineating the lesion accurately.

Among the benchmark models, AttentionU-Net stands out as the most effective for this particular case. Conversely, ResUNet++ and Multi-ResUNet exhibit over-segmentation issues, while the other models tend to under-segment the designated region. The lesion in the fourth row is large and irregular and located near the junction of the central and superior temporal sulcus. Only AFMS-Net adeptly captures previously overlooked regions of all models, ensuring a thorough and accurate segmentation. Meanwhile, the benchmark methods vary, with some showing marked over-segmentation or under-segmentation tendencies. Across all scenarios, Unet++ and ER-Net consistently lean towards conservative segmentations, resulting in substantial under-segmentation. Conversely, ResUNet++ and U-Net tend to produce aggressive segmentation, often mistakenly classifying cerebrospinal fluid in the lateral ventricles as target lesions. While ResUNet++ and Multi-ResUNet demonstrate commendable consistency regarding region similarity and boundary delineation, they do not surpass the benchmark models in all aspects. However, our proposed AFMS-Net excels in identifying areas that benchmark methods either

TABLE 4 Performance metrics of various segmentation methods evaluated on 655 cases from the ATLAS dataset.

Method	Accuracy	Precision	Recall	DSC	mIoU	AHD
U-Net	0.996 ± 0.014	0.727 ± 0.134	0.718 ± 0.154	0.721 ± 0.153	0.565 ± 0.171	11.850
Unet++	0.996 ± 0.025	0.866 ± 0.032	0.460 ± 0.018	0.602 ± 0.014	0.431 ± 0.017	13.798
AttentionU-Net	0.997 ± 0.026	0.782 ± 0.156	0.709 ± 0.123	0.742 ± 0.021	0.592 ± 0.165	11.407
ResUNet++	0.995 ± 0.027	0.769 ± 0.143	0.711 ± 0.154	0.738 ± 0.176	0.587 ± 0.169	12.232
MultiResUNet	0.993 ± 0.027	0.770 ± 0.176	0.713 ± 0.121	0.740 ± 0.153	0.588 ± 0.146	11.621
CS2-Net	0.997 ± 0.028	0.650 ± 0.137	0.692 ± 0.123	0.670 ± 0.175	0.504 ± 0.189	12.950
ER-Net	0.997 ± 0.025	0.716 ± 0.162	0.597 ± 0.175	0.653 ± 0.137	0.483 ± 0.212	13.396
SAEB (Our)	0.997 ± 0.028	0.820 ± 0.117	0.732 ± 0.165	0.772 ± 0.135	0.624 ± 0.102	10.642
DAEB (Our)	0.997 ± 0.034	0.839 ± 0.145	0.736 ± 0.131	0.782 ± 0.136	0.636 ± 0.175	10.416

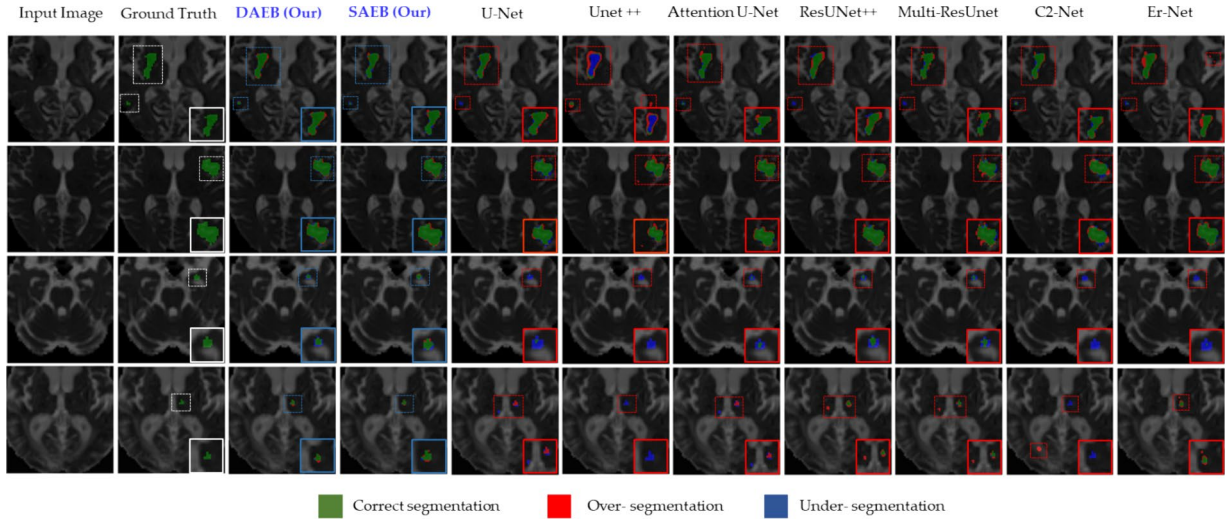


FIGURE 9
Visual examination of segmentation challenges in four selected samples from the ISLES 2022 Dataset. The arrangement from left to right comprises the original input image, the ground truth, and the segmentation outputs from our models (SAEB, DAEB), U-Net, U-Net++, Attention U-Net, ResUNet++, Multi-ResUnet, CS2-Net, and ER-Net. The samples are specifically chosen to clarify unique challenges such as multiple lesions, small lesions, and varying lesion sizes (median and large).

under-segmented or over-segmented, ensuring improved region alignment and enhanced boundary precision. While visual analysis provides insights into segmentation performance, a comprehensive quantitative assessment is essential for conclusive determinations. Accordingly, we subjected our proposed AFMS-Net and other prominent methods to rigorous evaluation metrics, with the detailed outcomes reported in Table 4. In a comparative assessment against prevailing methods, the proposed AFMS-DAEB distinctively achieves an impressive DSC of 78.20% and a mIoU of 63.60%. When benchmarked in mIoU scores, AFMS-DAEB consistently outperforms-surpassing U-Net by 8%, Unet++ by 19%, Attention U-Net by 4.4%, etc. This noticeable edge emphasizes our model's finesse in lesion segmentation and its proficiency in differentiating lesions from the intricate background noise typically found in medical imaging. In evaluating Precision and Recall, apparent differences emerge among the methods. Unet++ performs notably well in precision, with a score of 86.6%, reflecting its accuracy in detecting true positives.

On the other hand, our AFMS-DAEB leads in the recall, scoring 73.60%, highlighting its ability to detect most lesions effectively. Additionally, the DSC metric, essential for assessing the spatial overlap accuracy between the predicted segmentation and the ground truth, highlights the superior performance of AFMS-DAEB. Specifically, it leads by a 4–5% margin compared to the top benchmark model. Conclusively, these quantitative analyses demonstrate the excellent performance of our proposed network and highlight AFMS-DAEB's adeptness in complex tasks, notably boundaries and edge detection, a consistent challenge in medical image segmentation.

5.1.3 Qualitative and quantitative results on ISLES 2022 dataset

Similarly to section 5.2, we used the ISLES'22 dataset to evaluate our proposed variants further. This rigorous assessment emphasizes our model's efficacy (presented in Figure 9). This figure comprises four distinct rows, each corresponding to a specific stroke patient case. These cases encompass a range of complexities, from large infarct lesions to multiple embolic and cortical infarcts, which vary remarkably in location, size, and shape. In the first row, an apparent large lesion is accompanied by a smaller one. A group of benchmark models, specifically U-Net,

TABLE 5 A comprehensive evaluation of segmentation performance metrics for various methods across 246 cases in the ISLES 2022 dataset.

Method	Accuracy	Precision	Recall	DSC	mIoU	AHD
U-Net	0.994 ± 0.021	0.828 ± 0.130	0.704 ± 0.132	0.761 ± 0.123	0.614 ± 0.193	11.514
Unet++	0.994 ± 0.023	0.806 ± 0.021	0.680 ± 0.014	0.724 ± 0.013	0.584 ± 0.160	15.130
AttentionU-Net	0.995 ± 0.015	0.798 ± 0.130	0.768 ± 0.132	0.778 ± 0.021	0.650 ± 0.224	11.961
ResUNet++	0.995 ± 0.032	0.814 ± 0.132	0.771 ± 0.128	0.789 ± 0.152	0.663 ± 0.213	11.386
MultiResUNet	0.995 ± 0.026	0.843 ± 0.124	0.749 ± 0.121	0.787 ± 0.101	0.662 ± 0.195	12.534
CS2-Net	0.995 ± 0.022	0.775 ± 0.132	0.722 ± 0.123	0.733 ± 0.121	0.593 ± 0.190	13.312
ER-Net	0.995 ± 0.023	0.776 ± 0.136	0.755 ± 0.143	0.760 ± 0.139	0.623 ± 0.240	12.403
SAEB (Our)	0.995 ± 0.025	0.839 ± 0.132	0.780 ± 0.148	0.818 ± 0.136	0.680 ± 0.202	9.855
DAEB (Our)	0.995 ± 0.029	0.860 ± 0.139	0.761 ± 0.138	0.802 ± 0.129	0.673 ± 0.158	10.041

Unet++, AttentionUnet, ResUnet++, Multi-ResUnet, and ER-Net, failed to accurately segment the minor lesion. However, ER-net and ResUnet++ tended to over-segment, whereas Unet++ could not segment both lesions effectively. The delineated regions of interest are highlighted using a dotted rectangular line, and a zoomed view is provided for enhanced clarity. Ground truths are distinctly represented in white, our proposed models in blue, and benchmark models in red. The second row demonstrates that all segmentation methods identified the lesion's location. However, some inconsistencies were noted among the benchmark models. U-Net, AttentionUnet, and ResUnet showed tendencies of over-segmentation.

On the contrary, Unet++ and Er-Net leaned towards under-segmentation. In this context, our AFMS-Net demonstrated superior accuracy in delineating the lesion's shape, achieving remarkable regional overlap. The third and fourth rows present additional challenges, especially concerning smaller lesions. The benchmark models— U-Net, UNet++, AttentionUnet, ResUnet++, Multi-ResUnet, and ER-Net—all struggled with accurately segmenting the minor lesion. In stark contrast, our AFMS-Net showcased its competency by confidently segmenting all lesions, highlighting its distinct advantage in handling diverse lesion types. Quantitative analysis offers an objective perspective on the efficacy of segmentation models. Our evaluation of the ISLES'22 dataset, presented in Table 5, outlines the performance of AFMS-SAEB and AFMS-DAEB compared to other prominent models.

Most models demonstrate an impressive accuracy of around 99.5%, indicating a generally consistent segmentation accuracy across the board. In terms of precision, AFMS-DAEB achieves an outstanding 86.0%, outstripping all other models. Close behind is the Multi-ResUNet, with 84.3%. U-Net and Unet++ demonstrate 82.8 and 80.6% precision scores, respectively. When evaluating recall, the proposed AFMS-SAEB leads with a score of 78.0%. ResUNet++ and AttentionU-Net follow closely with 77.1 and 76.8% recalls, respectively. AFMS-DAEB further asserts its robustness with a recall of 76.1%. The DSC offers a holistic perspective on the overlap between the segmented output and the ground truth. AFMS-SAEB scores 81.8% in DSC, ResUNet++ has a DSC of 78.9%, and AFMS-DAEB reaches up to 80.2%. Regarding mIoU, AFMS-SAEB, which scores 68.0%, AFMS-DAEB closely follows this at 67.3%, and the competing models ResUNet++ and Multi-ResUNet are in the 66% range. In conclusion, each model has its strengths in specific domains. The proposed framework demonstrate an adept balance across all key metrics. This broad examination highlights the proficiency and capabilities of the proposed approach in medical image processing.

5.1.4 Generalizability across different imaging modalities and datasets

Our study mainly focuses on MRI datasets, which are crucial for brain lesion segmentation due to their high resolution and contrast between different brain tissues. We acknowledge the importance of assessing our model's generalizability across different imaging modalities to ensure its applicability in diverse clinical settings. However, our current investigation is confined to MRI data, considering its relevance and specificity to brain lesion analysis. The datasets utilized in our study encompass a range of MRI images with varying voxel sizes, which are as follows: BraTS 2021 and ATLAS v2.0 datasets have a voxel size of $1 \times 1 \times 1$ mm, providing high-resolution images for precise segmentation. Conversely, the ISLES 2022 dataset has a larger voxel size of $2 \times 2 \times 2$ mm, demonstrating our model's adaptability to images with lower resolution and potentially different characteristics. By evaluating AFMS-Net across these datasets, we aim to demonstrate its robustness not only to different lesion types but also to variations in image resolution, which is a step toward generalizability. However, we recognize that further studies are necessary to evaluate the model's performance across other imaging modalities, such as computed tomography (CT) scans or positron emission tomography (PET) images. Future work will involve extending our framework to include these modalities, thereby enhancing its diagnostic versatility and clinical utility.

5.2 Ablation studies

In this study, we introduce two encoder modules, SAEB and DAEB, in addition to a SegPath. We propose two different encoders to balance performance efficiency and computational cost. While SAEB offers competitive performance with fewer parameters, DAEB, although computationally more demanding, delivers slightly superior results. To evaluate the effectiveness of these components, we performed ablation studies using one brain tumor dataset and two-stroke datasets, specifically the BraTS 2021, ATLAS R2.0, and ISLES 2022 datasets. Initially, we evaluated the performance impact of substituting the original encoder in the 3D U-Net with our proposed SAEB encoder, resulting in the modified model termed AFMS-SAEB. This adaptation led to incremental gains in DSC and IoU by 0.36 and 0.09% for the BraTS 2021, 0.66 and 0.75% for the ATLAS, and 0.66 and 0.75% for the ISLES 2022. These results can be referenced in Table 6. Motivated by these initial findings, we explored the DAEB encoder as an alternative, creating the

TABLE 6 Ablation study assessing the incremental impact of SAEB and DAEB encoders and SegPath on segmentation metrics (DSC, IOU, AHD) across BRATS 2021, ATLAS R2.0, and ISLES 2022 datasets.

Network	DSC	IoU	AHD
Brats 2021 dataset			
Baseline (U-Net)	0.86	0.76	7.01
Baseline + SegPath	0.87	0.76	6.99
Baseline + SAEB	0.87	0.78	6.65
Baseline + SAEB + SegPath	0.89	0.80	6.26
Baseline + DAEB	0.88	0.80	6.35
Baseline + DAEB + SegPath	0.90	0.81	6.07
ATLAS R2.0 Dataset			
Baseline (U-Net)	0.72	0.56	11.8
Baseline + SegPath	0.73	0.57	11.6
Baseline + SAEB	0.75	0.60	11.2
Baseline + SAEB + SegPath	0.77	0.62	10.6
Baseline + DAEB	0.76	0.61	11.0
Baseline + DAEB + SegPath	0.78	0.63	10.4
ISLES 2022 Dataset			
Baseline (U-Net)	0.76	0.61	11.51
Baseline + SegPath	0.77	0.61	11.23
Baseline + SAEB	0.79	0.66	10.24
Baseline + SAEB + SegPath	0.81	0.68	9.855
Baseline + DAEB	0.79	0.65	10.25
Baseline + DAEB + SegPath	0.80	0.67	10.04

AFMS-DAEB model. The DAEB encoder exhibited superior performance, boosting DSC and IoU by 0.96 and 1% on the BraTS 2021, 1.3 and 1.5% on the ATLAS, and 1.5 and 2.7% on the ISLES 2022 dataset. These enhancements are also detailed in Table 6. Aside from qualitative improvements in segmentation, we also examined the computational performance of our proposed models. A comparative analysis between mIoU and the number of parameters for AFMS-SAEB and AFMS-DAEB and benchmark models has been depicted in Figure 7. This figure provides a balanced perspective on performance versus computational complexity.

In summary, our ablation studies, built on the baseline 3D U-Net model, attest to the efficacy of our proposed encoders. The summarized results and conclusions can be found in Table 6. By offering these two encoder alternatives, we allow users to choose between SAEB's computational efficiency or DAEB's slightly superior performance, depending on their specific requirements.

5.2.1 Ablation study for SAEB

We have conducted a comprehensive ablation study to evaluate the impact of integrating the Single Adaptive Encoder Block (SAEB) with the SegPath module. As seen in the results presented in Table 6, the fusion of SAEB with SegPath, as summarized by the AFMS-SAEB configuration, demonstrates substantial improvements across all examined datasets. Reviewing the BraTS 2021 dataset shows a marked enhancement in DSC and IoU metrics by integrating the SAEB and SegPath modules. In particular, the IoU increased from 76.7 to 81.6%, and the DSC score increased from the starting value of 86.7 to 89.4%. Same for ISLES 2022 and ATLAS R2.0 datasets. The outcomes show that the AFMS-SAEB model can accurately represent the edges of lesions and other small features, which are critical for medical image segmentation. The AFMS-SAEB's precise ability results from the SAEB module's feature extraction power and SegPath's capability in contextual capture, which precisely detects intricate anatomical and clinical characteristics. To sum up, Table 6 presents compelling evidence about the efficacy of SAEB and SegPath's combined competence inside the AFMS-SAEB model. Our ablation research demonstrates that AFMS-SAEB has considerable efficiency in fine-grain identification and segmentation and enhances the accuracy of image segmentation.

5.2.2 Ablation study for DAEB

The AFMS-DAEB is designed for the Dual-Dimension Attention mechanism purpose by the strategic integration of DAEB, SegPath, and decoder module (Table 6), demonstrates the performance and robustness of AFMS-DAEB for complex anatomical and pathological structures across various medical imaging datasets. For the Brats dataset, the proposed AFMS-DAEB significantly improved over the baseline method in DSC and IoU, from 86 to 90% and 76 to 81%, respectively. Due to the dual attention mechanism, the DAEB module can detect subtle lesions that most models may overlook. Improvements in the ATLAS R2.0 and ISLES 2022 datasets further validate the model's efficacy. The AFMS-DAEB emphasizes the importance of extracting details-oriented features. DAEB and SegPath modules, ensures that the model preserves and maintains a holistic understanding of a spatial context while extracting finer details, edges, and complex contrasts. Because of the DAEB's robustness, the model can extract the most contextual information from medical images, which helps it overcome the difficulties presented by subtle variances in medical imaging. In the meantime, the SegPath improves this by supporting the processing and hierarchical structuring of the learned features.

5.2.3 Ablation study for SegPath

To evaluate the effectiveness of SegPaths, we integrate the SegPaths with the base model U-Net to conduct quantitative analysis. The results are shown in Table 6. All three datasets had an improvement in DSC scores; BraTS, ATLAS, and ISLES registered scores of 87.1, 73.6, and 77.1%, respectively. In the Baseline + SAEB versus Baseline + SAEB + SegPath (AFMS-SAEB) comparison, the combination of SAEB and SegPath performed better. The DSC score increased from 88.7 to 89.4%, and the IoU score increased from 80.8 to 81.6% for the BraTS 2021. Notable improvements were also observed in the ATLAS and ISLES 2022 datasets, demonstrating the cooperative effect of the SAEB and SegPath. The model with Baseline + DAEB + SegPath (AFMS-SAEB) showed remarkable results at the end of our investigation, particularly when compared to the Baseline + DAEB. For example, the BraTS 2021 outperformed all previous architectures with DSC and IoU ratings of 90.2 and 82.3%, respectively. In Summary, SegPath dramatically improves the model's capacity for feature refinement.

6 Conclusion and future work

Deep learning models must effectively capture local and global features to perform accurate and efficient brain lesion segmentation. Previously, many state-of-the-art methods such as U-Net, VGG-Net, ResNet, and DenseNet have set the foundation. However, these methods may fail in precisely segmenting brain lesions due to the brain's complex structure. Moreover, these methods could face computational overload.

Thus, we introduce a novel network AFMS-Net to optimize segmentation accuracy and computational efficiency. Our proposed network has an encoder-decoder-like architecture that includes SAEB and DAEB modules. These encoder structures represent a notable shift in feature extraction, enhanced by techniques such as squeeze-and-excite and channel-spatial attention. The SAEB and DAEB utilized SegPath by combining residual and traditional skip connections for adaptive feature accumulation, which is further responsible for capturing and enhancing detailed features and multi-scale context for improved segmentation outcomes. Thus, it is suitable for limited computational resources, or the primary target is identifying and segmenting the most prominent features. SAEB is ideal for fast and efficient segmentation in scenarios prioritizing speed, unsuitable for complex, detailed analysis. DAEB excels in precise, intricate segmentation tasks, especially with multi-class lesions, not recommended for rapid, less detailed screenings.

The experimental findings of the AFMS-SAEB module demonstrated impressive performance in terms of Dice and IoU scores. For the BraTS dataset, 89.4% of Dice and 80.6% of IoU scores were achieved. The ATLAS scores were recorded as 77.2 and 62.4%, while on the ISLES dataset, the Dice and IoU scores were 81.8 and 68.0%, respectively. Compared to other models, it achieved a 2.7% improvement in Dice and 3.9% in IoU compared to U-Net, surpassing Attention U-Net by 3.4 and 4.8%, ResUNet++ by 2.5 and 3.2%, Multi-ResUNet by 3.8 and 5.4%, CS2-Net by 3.9 and 5.8%, and ER-Net by 3.3 and 4.5% on BRATS. Conversely, the proposed AFMS-DAEB module is suitable for fine-grained and complex segmentation tasks that utilize GAP, channel spatial, and weighted channel attention. It emphasizes information channels and integrates spatial attention to identify and classify various lesion types. AFMS-DAEB's effectiveness is validated through rigorous experiments on several datasets. On BraTS, it achieved remarkable Dice and IoU scores of 90.2% and 0.81.3%, respectively, showcasing its capability in handling complex brain tumor segmentation tasks. For ATLAS and ISLES, it achieved 78.2 and 80.2% (Dice scores) and 63.6 and 67.3% (IoU scores), supporting the model's robustness and versatility across different medical imaging challenges. Results across all datasets show that AFMS-DAEB performs better than the baseline U-Net model. Regarding Dice and IoU, it improved by 3.5 and 4.6% on BraTS, respectively. Performances were considerably greater on ATLAS, with an increase of 7% in IoU and 6.1% in Dice. The model demonstrated outstanding results: a rise of 5.9% in IoU and 4.1% in Dice on the ISLES dataset.

Furthermore, our study has some limitations because it only used high-resolution MRI scans, which may not accurately reflect the range of clinical circumstances that are seen in real-world settings. To be more specific, the performance of the AFMS-Net on datasets such as BraTS 2021, ATLAS v2.0, and ISLES 2022, which have voxel sizes of $1 \times 1 \times 1$ mm and $2 \times 2 \times 2$ mm respectively, demonstrates its ability in high-resolution context setting. When applied to lower-resolution images or other imaging modalities, which are often used in a variety of diagnostic contexts, this approach may raise concerns regarding the model's efficacy and flexibility. This limitation highlights the possibility of bias in the model towards the high-resolution features included in the datasets that were utilized, and it may raise the possibility of a compromise in the generalizability of the model. In order to overcome these issues, future research will focus on AFMS-Net's usefulness across various imaging modalities in addition to evaluating and improving its ability to adapt to images of various resolutions.

We will also refine our approach to parameter tuning and explore the potential of leveraging unsupervised learning for 3D medical image segmentation. In our forthcoming work, we aim to expand interdisciplinary collaborations that will augment the clinical applicability of our models. Through these collaborative efforts, we anticipate that AFMS-Net will profoundly influence clinical decision-making by facilitating precise and efficient lesion segmentation. In conclusion, AFMS-Net represents a significant advancement in medical image segmentation.

Data availability statement

The datasets used in this study can be found in online repositories, and the names of the repository/repositories and accession number(s) are provided in the article.

Author contributions

AZ: Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing, Conceptualization, Data curation, Formal analysis, Investigation. HH: Data curation, Investigation, Writing – review & editing. XZ: Data curation, Investigation, Software, Writing – review & editing. RK: Formal analysis, Validation, Visualization, Writing – review & editing. JL: Data curation, Software, Validation, Writing – review & editing. HY: Data curation, Formal analysis, Software, Writing – review & editing. XM: Data curation, Investigation, Software, Writing – review & editing. AC: Investigation, Software, Validation, Writing – review & editing. YY: Data curation, Formal analysis, Investigation, Software, Writing – review & editing. BH: Investigation, Software, Validation, Writing – review & editing. YG: Formal analysis, Investigation, Methodology, Software, Writing – review & editing, Validation, Visualization. YK: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. The National Key Research and Development Program of China, Grant Nos. 2022YFF0710800 and 2022YFF0710802; the National Natural Science Foundation of China, Grant Number 62071311; the Special Program for Key Fields of Colleges and Universities in Guangdong Province (Biomedicine and Health) of China, Grant Number 2021ZDZX2008; and the Stable Support Plan for Colleges and Universities in Shenzhen of China, Grant Number SZWD2021010.

Conflict of interest

YY was employed by company Shenzhen Lanmage Medical Technology Co., Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Anbeek, P., Vincken, K., Vanosch, M., Bisschops, R., and Vandergrond, J. (2004). Automatic segmentation of different-sized white matter lesions by voxel probability estimation. *Med. Image Anal.* 8, 205–215. doi: 10.1016/j.media.2004.06.019

Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., et al. (2021). The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv*. Available at: https://doi.org/10.48550/arXiv.2107.02314

Celaya, A., Actor, J. A., Muthusivarajan, R., Gates, E., Chung, C., Schellingerhout, D., et al. (2022). PocketNet: a smaller neural network for medical image analysis. *IEEE Trans. Med. Imaging* 42, 1172–1184. doi: 10.1109/TMI.2022.3224873

Chau, W., and McIntosh, A. R. (2005). The Talairach coordinate of a point in the MNI space: how to interpret it. *NeuroImage* 25, 408–416. doi: 10.1016/j.neuroimage.2004.12.007

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O.. (2016). 3D U-Net: learning dense volumetric segmentation from sparse annotation. Medical Image Computing and Computer-Assisted Intervention—MICCAI 2016: 19th International Conference. Athens, Greece, October 17–21, 2016. Springer

Gao, L., Li, J., Zhang, R., Bekele, H. H., Wang, J., Cheng, Y., et al. (2023). MMGan: a multimodal MR brain tumor image segmentation method. *Front. Hum. Neurosci.* 17:1275795. doi: 10.3389/fnhum.2023.1275795

Gooya, A., Pohl, K. M., Bilello, M., Cirillo, L., Biros, G., Melhem, E. R., et al. (2012). GLISTR: glioma image segmentation and registration. *IEEE Trans. Med. Imaging* 31, 1941–1954. doi: 10.1109/TMI.2012.2210558

Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., et al. (2023). Medical Image Computing and Computer Assisted Intervention–MICCAI 2023: 26th International Conference. Vancouver, BC, Canada: October 8–12, 2023. 14224. Springer Nature.

Guo, W., Zhou, H., Gong, Z., and Zhang, G. (2021). Double U-Nets for image segmentation by integrating the region and boundary information. *IEEE Access* 9, 69382–69390. doi: 10.1109/ACCESS.2021.3075294

Hassan, H., Ren, Z., Zhao, H., Huang, S., Li, D., Xiang, S., et al. (2022). Review and classification of AI-enabled COVID-19 CT imaging models based on computer vision tasks. *Comput. Biol. Med.* 141:105123. doi: 10.1016/j.compbiomed.2021.105123

Hernandez Petzsche, M. R., de la Rosa, E., Hanning, U., Wiest, R., Valenzuela, W., Reyes, M., et al. (2022). ISLES 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. *Sci. Data* 9:762. doi: 10.1038/s41597-022-01875-5

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., et al. (2017). MobileNets: efficient convolutional neural networks for mobile vision applications. *arXiv*. Available at: https://doi.org/10.48550/arXiv.1704.04861

Hu, J., Shen, L., Albanie, S., Sun, G., and Wu, E.. (2018). Squeeze-and-excitation networks Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., et al. (2020). UNet 3+: A full-scale connected UNet for medical image segmentation. ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE.

Hurlock, G. S., Higashino, H., and Mochizuki, T. (2009). History of cardiac computed tomography: single to 320-detector row multislice computed tomography. *Int. J. Cardiovasc. Imaging* 25, 31–42. doi: 10.1007/s10554-008-9408-z

Ibtehaz, N., and Rahman, M. S. (2020). MultiResUNet: rethinking the U-Net architecture for multimodal biomedical image segmentation. *Neural Netw.* 121, 74–87. doi: 10.1016/j.neunet.2019.08.025

Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., and Maier-Hein, K. H. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* 18, 203–211. doi: 10.1038/s41592-020-01008-z

Islam, A., Reza, S. M., and Iftekharuddin, K. M. (2013). Multifractal texture estimation for detection and segmentation of brain tumors. *IEEE Trans. Biomed. Eng.* 60, 3204–3215. doi: 10.1109/TBME.2013.2271383

Jha, D., Smedsrud, P. H., Riegler, M. A., Johansen, D., de Lange, T., Halvorsen, P., et al. (2019). ResUNet++: an advanced architecture for medical image segmentation. 2019 IEEE International Symposium on Multimedia (ISM). IEEE.

Kermi, A., Behaz, M. K. N., Benamar, A., and Khadir, M. T.. (2022). A deep learning-based 3D-GAN for glioma subregions detection and segmentation in multimodal brain MRI volumes 2022 International Symposium on Innovative Informatics of Biskra (ISNIB). IEEE.

Li, Z., Kamnitsas, K., and Glocker, B. (2020). Analyzing overfitting under class imbalance in neural networks for image segmentation. *IEEE Trans. Med. Imaging* 40, 1065–1077. doi: 10.1109/TMI.2020.3046692

Li, H., Qiu, K., Chen, L., Mei, X., Hong, L., and Tao, C. (2020). SCAttNet: semantic segmentation network with spatial and channel attention mechanism for high-resolution remote sensing images. *IEEE Geosci. Remote Sens. Lett.* 18, 905–909. doi: 10.1109/LGRS.2020.2988294

Liew, S.-L., Anglin, J. M., Banks, N. W., Sondag, M., Ito, K. L., Kim, H., et al. (2017). The Anatomical Tracings of Lesions After Stroke (ATLAS) Dataset—Release 1.1. *bioRxiv*. Available at: https://doi.org/10.1101/179614

Limonova, E., Alfonso, D., Nikolaev, D., and Arlazarov, V. V.. (2021). ResNet-like architecture with low hardware requirements 2020 25th International Conference on Pattern Recognition (ICPR). IEEE.

Ma, J., Yuan, G., Guo, C., Gang, X., and Zheng, M. (2023). SW-UNet: a U-Net fusing sliding window transformer block with CNN for segmentation of lung nodules. *Front. Med.* 10:1273441. doi: 10.3389/fmed.2023.1273441

Mehrani, P., and Tsotsos, J. K.. (2023). Self-attention in vision transformers performs perceptual grouping, not attention. *arXiv*. Available at: https://doi.org/10.48550/arXiv.2303.01542

Mou, L., Zhao, Y., Fu, H., Liu, Y., Cheng, J., Zheng, Y., et al. (2021). CS2-Net: deep learning segmentation of curvilinear structures in medical imaging. *Med. Image Anal.* 67:101874. doi: 10.1016/j.media.2020.101874

Mubashar, M., Ali, H., Grönlund, C., and Azmat, S. (2022). R2U++: a multiscale recurrent residual U-net with dense skip connections for medical image segmentation. *Neural Comput. Appl.* 34, 17723–17739. doi: 10.1007/s00521-022-07419-7

Myronenko, A.. (2019). 3D MRI brain tumor segmentation using autoencoder regularization. Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II. Springer

Nie, X., Zhou, X., Tong, T., Lin, X., Wang, L., Zheng, H., et al. (2022). N-Net: a novel dense fully convolutional neural network for thyroid nodule segmentation. *Front. Neurosci.* 16:872601. doi: 10.3389/fnins.2022.872601

Oktay, O., Schlemper, J., Le Folgoc, L., Lee, M., Heinrich, M., Misawa, K., et al. (2018). Attention U-Net: learning where to look for the pancreas. *arXiv*. Available at: https://doi.org/10.48550/arXiv.1804.03999

Rashid, T., Abdulkadir, A., Nasrallah, I. M., Ware, J. B., Liu, H., Spincemaille, P., et al. (2021). DEEPMIR: a deep neural network for differential detection of cerebral microbleeds and iron deposits in MRI. *Sci. Rep.* 11:14124. doi: 10.1038/s41598-021-93427-x

Ronneberger, O., Fischer, P., and Brox, T.. (2015). U-Net: convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference. Munich, Germany: October 5–9, 2015. Springer

Shatnawi, A., Al-Bdour, G., Al-Qurran, R., and Al-Ayyoub, M.. (2018). A comparative study of open source deep learning frameworks 2018 9th International Conference On Information And Communication Systems (ICICS). IEEE.

Siuly, S., and Zhang, Y. (2016). Medical big data: neurological diseases diagnosis through medical data analysis. *Data Sci. Eng.* 1, 54–64. doi: 10.1007/s41019-016-0011-3

Stoyanov, D., Taylor, Z., Kainz, B., Maicas, G., Beichel, R. R., Martel, A., et al. (2018). Image Analysis for Moving Organ, Breast, and Thoracic Images: Third International Workshop, RAMBO 2018, Fourth International Workshop, BIA 2018, and First International Workshop, TIA 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16 and 20, 2018, Proceedings. 11040. Springer.

Tan, M., and Le, Q.. (2019). EfficientNet: rethinking model scaling for convolutional neural networks. International Conference on Machine Learning. 21013–21036

Vedaei, F., Mashhadi, N., Zabrecky, G., Monti, D., Navarreto, E., Hriso, C., et al. (2023). Identification of chronic mild traumatic brain injury using resting state functional MRI and machine learning techniques. *Front. Neurosci.* 16:1099560. doi: 10.3389/fnins.2022.1099560

Wang, X., Han, S., Chen, Y., Gao, D., and Vasconcelos, N.. (2019). Volumetric attention for 3D medical image segmentation and detection. Medical Image Computing and Computer Assisted Intervention—MICCAI 2019: 22nd International Conference. Shenzhen, China, October 13–17, 2019. Springer

Wang, R., Lei, T., Cui, R., Zhang, B., Meng, H., and Nandi, A. K. (2022). Medical image segmentation using deep learning: a survey. *IET Image Process.* 16, 1243–1267. doi: 10.1049/ipr2.12419

Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S.. (2018). CBAM: convolutional block attention module. Proceedings of the European Conference on Computer Vision (ECCV).

Xia, L., Zhang, H., Wu, Y., Song, R., Ma, Y., Mou, L., et al. (2022). 3D vessel-like structure segmentation in medical images by an edge-reinforced network. *Med. Image Anal.* 82:102581. doi: 10.1016/j.media.2022.102581

Yang, R., and Yu, Y. (2021). Artificial convolutional neural network in object detection and semantic segmentation for medical imaging analysis. *Front. Oncol.* 11:638182. doi: 10.3389/fonc.2021.638182

Yurtkulu, S. C., Şahin, Y. H., and Unal, G.. (2019). Semantic segmentation with extended DeepLabv3 architecture. 2019 27th Signal Processing and Communications Applications Conference (SIU). IEEE.

Zeng, X., Guo, Y., Zaman, A., Hassan, H., Lu, J., Xu, J., et al. (2023). Tubular structure segmentation via multi-scale reverse attention sparse convolution. *Diagnostics* 13:2161. doi: 10.3390/diagnostics13132161

Zhang, J., Xie, Y., Wang, Y., and Xia, Y. (2020). Inter-slice context residual learning for 3D medical image segmentation. *IEEE Trans. Med. Imaging* 40, 661–672. doi: 10.1109/TMI.2020.3034995

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., and Liang, J.. (2018). UNet++: A Nested U-Net Architecture for Medical Image Segmentation: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings. Springer.

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., and Liang, J. (2019). UNet++: redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Trans. Med. Imaging* 39, 1856–1867. doi: 10.1109/TMI.2019.2959609

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 11 June 2024
DOI 10.3389/fnins.2024.1365141

OPEN ACCESS

EDITED BY
Anees Abrol,
Georgia State University, United States

REVIEWED BY
Zhijun Liu,
Shandong Second Medical University, China
Chien-Yuan Lin,
GE Healthcare, Taiwan
Xiang Feng,
The University of Sydney, Australia

*CORRESPONDENCE
Xiaoan Zhang
✉ zxa@zzu.edu.cn

RECEIVED 03 January 2024
ACCEPTED 21 May 2024
PUBLISHED 11 June 2024

CITATION
Zhang P, Yang J, Shu Y, Cheng M, Zhao X, Wang K, Lu L, Xing Q, Niu G, Meng L, Wang X, Zhou L and Zhang X (2024) The value of synthetic MRI in detecting the brain changes and hearing impairment of children with sensorineural hearing loss.
Front. Neurosci. 18:1365141.
doi: 10.3389/fnins.2024.1365141

The value of synthetic MRI in detecting the brain changes and hearing impairment of children with sensorineural hearing loss

Penghua Zhang[1], Jinze Yang[1], Yikai Shu[2], Meiying Cheng[1], Xin Zhao[1], Kaiyu Wang[3], Lin Lu[1], Qingna Xing[1], Guangying Niu[1], Lingsong Meng[1], Xueyuan Wang[1], Liang Zhou[1] and Xiaoan Zhang[1]*

[1]Third Affiliated Hospital of Zhengzhou University, Zhengzhou, Henan, China, [2]Henan University of Science and Technology, Luoyang, Henan, China, [3]MRI Research, GE Healthcare, Beijing, China

Introduction: Sensorineural hearing loss (SNHL) can arise from a diverse range of congenital and acquired factors. Detecting it early is pivotal for nurturing speech, language, and cognitive development in children with SNHL. In our study, we utilized synthetic magnetic resonance imaging (SyMRI) to assess alterations in both gray and white matter within the brains of children affected by SNHL.

Methods: The study encompassed both children diagnosed with SNHL and a control group of children with normal hearing {1.5-month-olds (n = 52) and 3-month-olds (n = 78)}. Participants were categorized based on their auditory brainstem response (ABR) threshold, delineated into normal, mild, moderate, and severe subgroups.Clinical parameters were included and assessed the correlation with SNHL. Quantitative analysis of brain morphology was conducted using SyMRI scans, yielding data on brain segmentation and relaxation time.Through both univariate and multivariate analyses, independent factors predictive of SNHL were identified. The efficacy of the prediction model was evaluated using receiver operating characteristic (ROC) curves, with visualization facilitated through the utilization of a nomogram. It's important to note that due to the constraints of our research, we worked with a relatively small sample size.

Results: Neonatal hyperbilirubinemia (NH) and children with inner ear malformation (IEM) were associated with the onset of SNHL both at 1.5 and 3-month groups. At 3-month group, the moderate and severe subgroups exhibited elevated quantitative T1 values in the inferior colliculus (IC), lateral lemniscus (LL), and middle cerebellar peduncle (MCP) compared to the normal group. Additionally, WMV, WMF, MYF, and MYV were significantly reduced relative to the normal group. Additionally, SNHL-children with IEM had high T1 values in IC, and LL and reduced WMV, WMF, MYV and MYF values as compared with SNHL-children without IEM at 3-month group. LL-T1 and WMF were independent risk factors associated with SNHL. Consequently, a prediction model was devised based on LL-T1 and WMF. ROC for training set, validation set and external set were 0.865, 0.806, and 0.736, respectively.

Conclusion: The integration of T1 quantitative values and brain volume segmentation offers a valuable tool for tracking brain development in children affected by SNHL and assessing the progression of the condition's severity.

KEYWORDS

sensorineural hearing loss, white matter, synthetic MRI, magnetic resonance imaging, brain volume

Introduction

Congenital sensorineural hearing loss (SNHL) denotes deafness occurring before language development, typically during pregnancy, impacting auditory neural pathways. Approximately 1.2–1.7 cases per 1,000 live births lead to permanent childhood hearing loss due to SNHL (Korver et al., 2010). Delayed diagnosis in infants and young children with SNHL can profoundly hinder learning, affecting language acquisition, memory formation, and cognitive development (Surprenant and Didonato, 2014; Slade et al., 2020; Johnson et al., 2021; Shende and Mudar, 2023).

While the auditory brainstem response (ABR) test is commonly utilized for hearing screening in newborns, more quantitative and sensitive measures are necessary for early and precise diagnosis. Magnetic resonance imaging (MRI) is pivotal in diagnosing and monitoring disease progression and treatment responses (Van Der Weijden et al., 2023). Techniques such as Diffusion Tensor Imaging (DTI), Diffusion Kurtosis Imaging (DKI), and Functional Magnetic Resonance Imaging (fMRI) have been instrumental in diagnosing SNHL and studying brain development in affected infants (Wang et al., 2019). However, conventional MRI methods (T1WI, T2WI) lack the ability for quantitative analysis of brain region changes. Moreover, most studies involve subjects older than 2 years (Wang et al., 2023), potentially limiting the efficacy of interventions aimed at improving language discrimination abilities. While techniques like DTI, DKI, and fMRI offer quantitative analysis, they often necessitate longer scan durations.

Synthetic Magnetic Resonance Imaging (SyMRI) is an innovative technology for relaxation quantification imaging, delivering T1 and T2 relaxation times along with proton density (PD) in a single scan within clinically acceptable acquisition times (Chari and Chan, 2017; Goncalves et al., 2018). This approach offers absolute measurements of tissue microstructure, enhancing the objectivity of disease assessment. Unlike traditional methods, SyMRI allows adjustments of parameters like repetition time, echo time, and inversion time based on mathematical calculations rather than predefined settings (Gulani et al., 2004; Ji et al., 2022). This advancement reduces brain diagnostic study durations to ∼5 min with SyMRI, potentially enhancing throughput and minimizing the need for rescans, while delivering valuable quantitative data (Warntjes et al., 2008). SyMRI software, such as Synthetic MR from Linköping, Sweden, streamlines the generation of synthetic quantitative images. It offers fully automated volumetric parameters based on anticipated quantitative values for various brain tissue types (West et al., 2012). Integrated into radiology picture archiving and communication systems, this software enables rapid brain volume analysis in under 1 min (Granberg et al., 2016; Vanderhasselt et al., 2020).

Utilizing SyMRI technology, each voxel within an MRI scan can be categorized into four components: white matter (WM), gray matter (GM), cerebrospinal fluid (CSF), and non-WM/GM/CSF (NON). The measurement of SyMRI volume has been extensively investigated in both pediatric and adult populations. Previous case reports have highlighted the efficacy of SyMRI in diagnosing conditions like Sturge-Weber syndrome (Andica et al., 2016). Moreover, SyMRI enables the synthesis of Gd-enhanced FLAIR images post-acquisition, while Gd-enhanced synthetic Double Inversion Recovery (DIR) can aid in accentuating subtle meningeal enhancements (Andica et al., 2017). SyMRI scans have demonstrated superior plaque detection in multiple sclerosis (MS) compared to conventional MRI (Granberg et al., 2016). Additionally, the utilization of synthetic DIR and Phase-Sensitive Inversion Recovery (PSIR) images may facilitate the identification of intra-cortical or mixed WM-GM lesions (Miller et al., 1998). Studies by Vagberg et al. (2013) have validated SyMRI volumetric analysis as a reliable method for determining brain parenchymal fraction (BPF) in MS, showing that BPF is notably lower in pediatric MS cases, primarily due to GM loss (Yeh et al., 2009). These quantitative values are invaluable in evaluating brain tumors, aiding in differentiation between glioblastomas and metastases (Badve et al., 2017), as well as revealing the internal structure of tumors and lesions in MS (Granberg et al., 2016; Chen et al., 2021; Nunez-Gonzalez et al., 2022). While research on brain relaxation time in SNHL, particularly in children within the first year, is lacking, the potential for SyMRI in exploring this area remains untapped.

In our study, we employed SyMRI to examine the quantitative T1, T2, and PD values across 10 brain regions and 12 brain segmentations in children with SNHL at 1.5 and 3 months of age. Our results offer significant insights for clinical diagnosis and early developmental research in children affected by SNHL.

Materials and methods

Participants and clinical assessments

The study received approval from the local ethics committee. A discovery cohort of 80 children diagnosed with SNHL participated and 33 children have normal ABR threshold, in which 52 children tested at 1.5 months and 61 tested at 3 months. An external cohort included 17 children tested at 3 months, comprising nine children diagnosed with SNHL and eight children were normal. All participants underwent ABR testing to determine their hearing thresholds. The severity of hearing loss was categorized as mild (31–50 dB), moderate (51–70 dB), or severe (>70 dB) for each ear.

Inclusion criteria encompassed right-handed children with SNHL identified through hearing screening tests at 1.5 and 3 months post-birth, with bilateral ABR thresholds exceeding 30 dB. Exclusion criteria involved the presence of severe neurological disorders such as epilepsy and congenital leukodystrophy, cognitive impairments like autism and severe hyperactivity syndrome, and a history of treatment for ear-related infections.

Imaging examinations

SyMRI was conducted on a 3.0 T scanner (SIGNA Pioneer; GE Healthcare, Waukesha, WI, USA) equipped with a 21-channel head coil for all participants. Prior to scanning, children were sedated with midazolam (intramuscular or intravenous administration: 0.05–0.1 mg/kg/time via slow injection for 5 min) and immobilized using a MedVac vacuum device (CFI Medical Solutions, Fenton, Michigan). Ear protection was ensured with neonatal earmuffs covered by headphones. Parental consent was obtained before

TABLE 1 Summary of participant characteristics in the 1.5-month group.

	Normal (n =14)	Mild (n = 16)	Moderate (n =13)	Severe (n = 9)	p-value
Sex (female, male)	6, 8	8, 8	6, 7	3, 6	0.96
Birth method (natural delivery, cesarean section)	8, 6	10, 6	9, 4	5, 4	0.868
Birth weight (g)	3,325 (3,112.5, 3,662.5)	3,050 (2,900, 3,512.5)	3,200 (3,000, 3,800)	3,175 (3,075, 3,487.5)	0.559
Gestational age at birth (weeks)	39.29 (38.25, 40.75)	39.93 (39, 40.21)	39.43 (37.86, 40.43)	39.43 (38.79, 40.29)	0.651
dB hearing loss: left ear	\	43.75 ± 6.191	63.85 ± 6.50	92.22 ± 8.33	<0.001
dB hearing loss: right ear	\	45.63 ± 5.12	64.62 ± 6.60	94.44 ± 5.27	<0.001

TABLE 2 Summary of participant characteristics in the 3-month group.

	Normal (n =19)	Mild (n = 15)	Moderate (n =19)	Severe (n = 8)	p-value
Sex (female, male)	10, 9	7, 8	11, 8	6, 2	0.655
Birth method (natural delivery, cesarean section)	9, 10	9, 6	9, 10	4, 4	0.871
Birth weight (g)	3,286.84 ± 421.26	3,265.67 ± 542.65	3,323.68 ± 390.29	3,481.25 ± 465.17	0.953
Gestational age at birth (weeks)	39.71 (38.64, 40)	39.57 (38.86, 39.93)	39.29 (38.29, 40)	39.29 (38.79, 40.18)	0.719
dB hearing loss: left ear	\	44.00 ± 6.33	62.11 ± 7.13	87.50 ± 10.35	<0.001
dB hearing loss: right ear	\	45.33 ± 5.16	62.11 ± 7.13	93.75 ± 7.44	<0.001

MRI and sedation. The sequence parameters for SyMRI were set as follows: Field of View (FOV) = 200 mm, slice thickness = 3 mm, slice gap = 0.5 mm, number of slices = 36, TR/TE = 4,230/20.4 ms, NEX=1, with an acquisition time of 5 min and 8 s. Quantification maps (T1, T2, and PD) were generated using the vendor-provided program (SyMRI 8.0; SyntheticMR, Linköping, Sweden).

Measurements of quantitative values

Following the scans, two neurology specialists meticulously reviewed all scan sequences to eliminate any macroscopic pathology. The SyMRI sequence image guide supplier's program (SyMRI 8.0, Synthetic MR, Linköping, Sweden) was employed to automatically generate T1 and T2 mapping diagrams. The regions of interest (ROIs) for this study were primarily delineated by the co-first author, possessing 7 and 6 years of experience in imaging diagnosis, respectively. All findings underwent thorough review and verification by the corresponding authors and imaging instructors of this study, each with 20 years of imaging diagnosis expertise. For manual operations on T1 and T2 mapping diagrams, the ITK-SNAP 3.8.0 software was utilized. Ten ROIs were sketched, including the semioval center (SC), frontal lobe (FL), posterior limb of the internal capsule (PLIC), genu of the corpus callosum (GCC), splenium of the corpus callosum (SCC), caudate nucleus (CN), globus pallidus (GP), inferior colliculus (IC), lateral lemniscus (LL), and middle cerebellar peduncle (MCP). Each ROI was meticulously placed to ensure precise anatomical positioning, minimizing interference from cerebrospinal fluid and surrounding anatomical structures. T1 and T2 values for each ROI were measured thrice, and their averages were computed. Subsequently, the mean values of symmetrical parts from both brain hemispheres were calculated post-measurement.

MR volumetric calculations

The raw data obtained from SyMRI underwent further processing with the SyMRI 8.0 post-processing software to derive brain segmentation volume and relaxation values. This included parameters such as white matter volume (WMV), gray matter volume (GMV), cerebrospinal fluid volume (CSF), myelin volume (MYV), brain parenchymal volume (BPV), intracranial volume (ICV), non-WM/GM/CSF (NON), white matter fraction (WMF = WMV/BPV), myelin fraction (MYF = MYV/BPV), gray matter fraction (GMF = GMV/BPV), NONF = NON/BPV, and cerebrospinal fluid fraction (CSFF = CSF/ICV).

Construction and validation of the prediction model

Parameters including IC-T1, LL-T1, MCP-T1, WMV, WMF, MYV, and MYF were chosen for children examined at

TABLE 3 Correlation of clinical parameters and SNHL at 1.5-month group.

Clinical parameters	Total ($n = 52$)	Normal ($n = 14$)	SNHL ($n = 38$)	p
Premature birth, *n* (%)				1
No	36 (69)	10 (71)	26 (68)	
Yes	16 (31)	4 (29)	12 (32)	
NH, *n* (%)				0.03
No	30 (58)	12 (86)	18 (47)	
Yes	22 (42)	2 (14)	20 (53)	
GDM, *n* (%)				0.746
No	34 (65)	10 (71)	24 (63)	
Yes	18 (35)	4 (29)	14 (37)	
HDP, *n* (%)				1
No	42 (81)	11 (79)	31 (82)	
Yes	10 (19)	3 (21)	7 (18)	
CMV infection, *n* (%)				0.729
No	39 (75)	10 (71)	29 (76)	
Yes	13 (25)	4 (29)	9 (24)	
IEM, *n* (%)				0.002
No	36 (69)	14 (100)	22 (58)	
Yes	16 (31)	0 (0)	16 (42)	

NH, neonatal hyperbilirubinemia; HDP, pregnancy-induced hypertension; GDM, gestational diabetes mellitus; CMV, cytomegalovirus; IEM, inner ear malformation.

TABLE 4 Correlation of clinical parameters and SNHL at 3-month group.

Clinical parameters	Total ($n = 52$)	Normal ($n = 14$)	SNHL ($n = 38$)	p
Premature birth, *n* (%)				0.803
No	42 (69)	14 (74)	28 (67)	
Yes	19 (31)	5 (26)	14 (33)	
NH, *n* (%)				0.022
No	30 (49)	14 (74)	16 (38)	
Yes	31 (51)	5 (26)	26 (62)	
GDM, *n* (%)				0.436
No	39 (64)	14 (74)	25 (60)	
Yes	22 (36)	5 (26)	17 (40)	
HDP, *n* (%)				1
No	47 (77)	15 (79)	32 (76)	
Yes	14 (23)	4 (21)	10 (24)	
CMV infection, *n* (%)				0.707
No	52 (85)	17 (89)	35 (83)	
Yes	9 (15)	2 (11)	7 (17)	
IEM, *n* (%)				0.003
No	47 (77)	19 (100)	28 (67)	
Yes	14 (23)	0 (0)	14 (33)	

NH, neonatal hyperbilirubinemia; HDP, pregnancy-induced hypertension; GDM, gestational diabetes mellitus; CMV, cytomegalovirus; IEM, inner ear malformation.

3 months. The discovery cohort of 61 samples was randomly divided into training and validation sets in a ratio of 55–45%, respectively. The external validation set contained 17 samples, including eight normal samples and nine SNHL samples. Univariate analysis was conducted, and variables with p-values < 0.05 were included for multivariate analysis using the bidirectional stepwise regression method in training set. A generalized linear model was then employed to build the prediction model. Evaluation of the model's efficacy was performed using a ROC curve, and visualization of a nomogram was facilitated using the R packages "pROC" and "regplot".

Statistical analysis

Data analysis was conducted utilizing R software (version 4.0.1). Analysis of variance (ANOVA) was employed to assess differences among variables across the normal, mild, moderate, and severe groups. The Wilcoxon test was utilized for non-normally distributed data to compare differences between two groups, while the Student's t-test was applied for normally distributed data. A significance level of $p < 0.05$ was considered statistically significant for all analyses.

Results

Correlation of clinical parameters and onset of SNHL

To assess the diagnostic efficacy of SyMRI for SNHL, we conducted evaluations on a cohort of 52 children at 1.5 months and 61 children at 3 months. The sample was categorized into four groups based on disease severity: normal, mild, moderate, and severe. Tables 1, 2 provide comprehensive clinical details of these children. Notably, no significant differences were detected in age, birth method, birth weight, or sex across the normal, mild, moderate, and severe subgroups. Next, we evaluated the correlation of clinical complications of newborns and pregnant women and onset of SNHL. Results demonstrated that neonatal hyperbilirubinemia (NH) and children with inner ear malformation (IEM) were associated with high incidence of SNHL (Tables 3, 4) both at 1.5 and 3-month group. Next, we used SyMRI to calculate the T1, T2 and PD values as well as automatic whole-brain volume segmentation. Our analysis focused on 10 ROIs, including SC, FL, PLIC, GCC, SCC, CN, GP, IC, LL, and MCP (Figures 1A–D). Additionally, Figures 1E, F depict representative T1 and T2 quantitative maps, respectively.

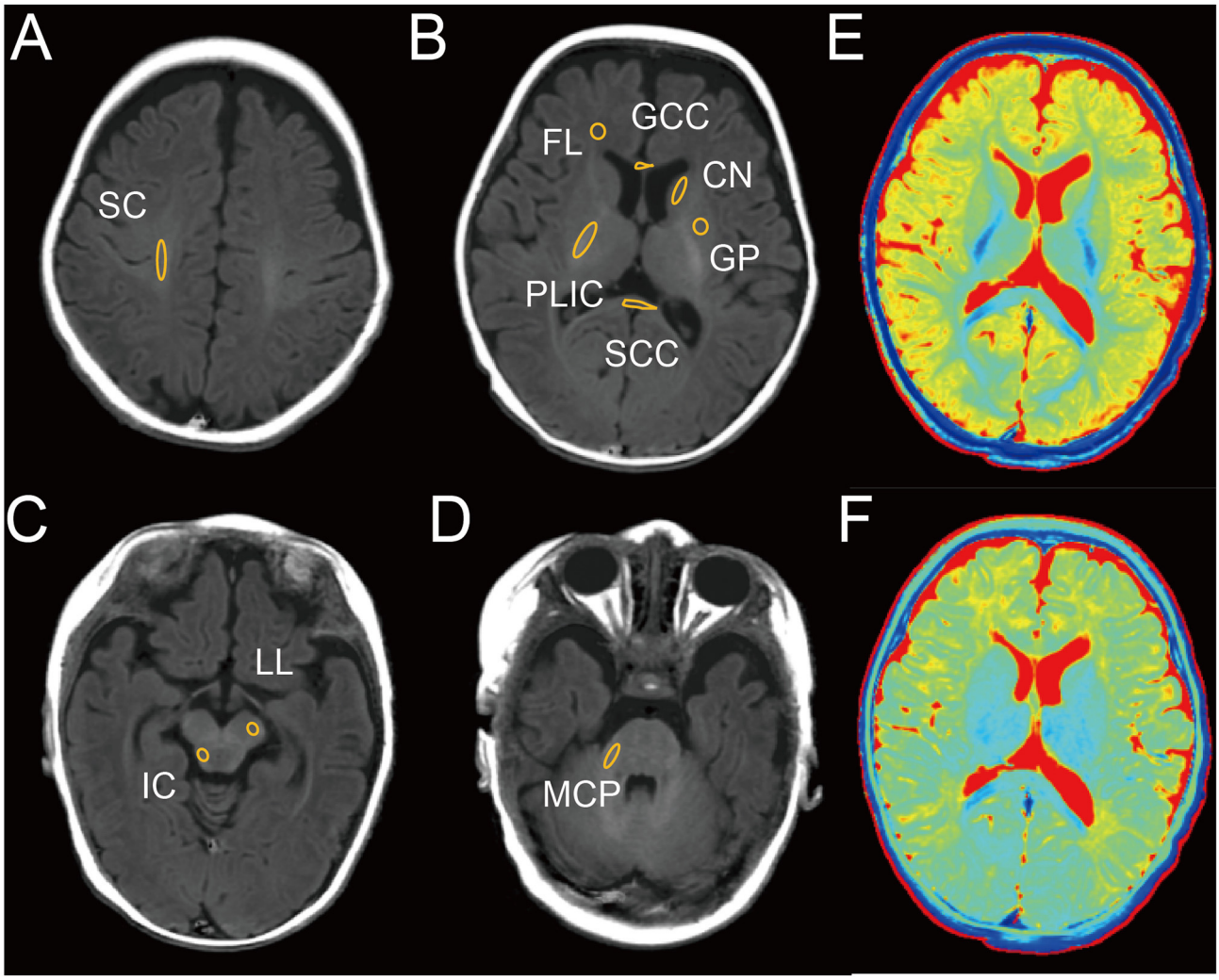


FIGURE 1
Representative image of male infant brain at 3 months with delineation of ROIs on T1 image. **(A)** Semioval center (SC). **(B)** Frontal lobe (FL), posterior limb of the internal capsule (PLIC), genu of the corpus callosum (GCC), splenium of the corpus callosum (SCC), caudate nucleus (CN), globus pallidus (GP). **(C)** Inferior colliculus (IC), lateral lemniscus (LL). **(D)** Middle cerebellar peduncle (MCP). **(E)** T1 map. **(F)** T2 map.

Measurements of quantitative parameters correlated with SNHL

Initially, we conducted an analysis of T1, T2, and PD values in the brains of children tested at 1.5- and 3-month groups across normal, mild, moderate, and severe subgroups. At 1.5-month group, no significant changes were observed in T1, T2, and PD values across the four subgroups (Table 5). However, at 3-month group, significant differences were noted in T1 values within the IC, LL, and MCP regions across the four subgroups, while T2 and PD values remained relatively stable (Table 6). Subsequent pairwise comparisons of T1 values within IC, LL, and MCP between the groups at 1.5 and 3 months revealed no significant differences at 1.5 month-group (Figures 2A–C). However, at 3 month-group, while no significant change in T1 values was observed between the normal and mild subgroups in IC, LL, and MCP, there was a notable progressive increase in T1 values from moderate to severe subgroups compared to the normal subgroup in IC and LL, with a similar trend observed in MCP, albeit only significantly in the severe group (Figures 2D–F). These findings underscore the potential of T1 values to serve as a more sensitive indicator of SNHL progression by 3 months.

Detection of brain volume segmentation correlated with SNHL

Subsequently, we examined 12 brain segmentation parameters, comprising WMV, GMV, CSF, BPV, ICV, MYV, NON, WMF, GMF, CSFF, NONF, and MYF, across the normal, mild, moderate, and severe subgroups at 1.5 and 3 months. In line with the T1, T2, and PD findings, at 1.5 month-group, these parameters exhibited no significant differences among the four subgroups

10.3389/fnins.2024.1365141

TABLE 5 T1, T2, and PD values in different regions of the brain in the 1.5-month group.

Variables	Total ($n = 52$)	Normal ($n = 14$)	Mild ($n = 16$)	Moderate ($n = 13$)	Severe ($n = 9$)	p-value
SCC-T1	2,217 (2,159, 2,343.25)	2,206 (2,163, 2,310)	2,182 (2,136.75, 2,368.25)	2,203 (2,164, 2,338)	2,294 (2,218, 2,295)	0.487
SCC-T2	278.56 ± 16.04	278.71 ± 14.52	285.06 ± 14.88	269.23 ± 17.21	280.22 ± 14.47	0.063
SCC-PD	108.26 ± 2.83	108.43 ± 2.93	106.99 ± 2.77	108.97 ± 3.02	109.23 ± 1.99	0.16
GCC-T1	2,865 (2,749, 2,912)	2,867.5 (2,814, 2,902)	2,793.5 (2,662, 2,873.25)	2,877 (2,783, 2,912)	2,832 (2,774, 3,000)	0.397
GCC-T2	211 (204, 227)	214 (204, 230.75)	211.5 (205, 229.75)	205 (204, 221)	208 (203, 211)	0.517
GCC-PD	128.85 (127.07, 130.57)	128.05 (124.82, 129)	128.5 (127.6, 129.82)	129.7 (129, 131.3)	128.7 (128.2, 131)	0.271
IC-T1	1,905.79 ± 149.91	1,935 ± 130.07	1,887 ± 122.25	1,841.69 ± 150.92	1,986.33 ± 193.43	0.121
IC-T2	163.35 ± 9.84	167.21 ± 7.94	163.62 ± 10.61	159.92 ± 10.13	161.78 ± 10.1	0.269
IC-PD	115.73 ± 4.27	115.65 ± 5.36	116.18 ± 4.64	114.97 ± 3.92	116.14 ± 2.2	0.885
LL-T1	1,552.56 ± 93.02	1,557.36 ± 107.49	1,544.88 ± 78.77	1,556.38 ± 71.93	1,553.22 ± 129.65	0.984
LL-T2	158.19 ± 8.67	158.71 ± 8.19	160.44 ± 7.96	155.46 ± 9.84	157.33 ± 9.07	0.486
LL-PD	161.97 ± 7.46	161.84 ± 8.24	159.89 ± 6.68	164.18 ± 7.09	162.66 ± 8.28	0.495
PLIC-T1	2,437 (2,343.5, 2,571.5)	2,546 (2,404.5, 2,698.5)	2,397.5 (2,295, 2,494.5)	2,404 (2,349, 2,472)	2,523 (2,364, 2,614)	0.252
PLIC-T2	179.6 ± 12.17	183 ± 14.1	177.44 ± 10.34	179.31 ± 13.24	178.56 ± 11.26	0.656
PLIC-PD	108.89 ± 4.71	109.9 ± 5.51	106.81 ± 3.82	110.03 ± 4.86	109.37 ± 4.04	0.203
FL-T1	3,687.5 (3,524.75, 3,868)	3,777.5 (3,587.5, 3,909.5)	3,630 (3,377.75, 3,810.75)	3,646 (3,535, 3,836)	3,696 (3,529, 4,313)	0.376
FL-T2	282 (258.5, 316.75)	301.5 (275.25, 326)	279 (255.25, 313.75)	282 (259, 316)	276 (257, 343)	0.628
FL-PD	135.7 (134, 136.6)	136.5 (135.45, 137.2)	135.1 (133.18, 136.27)	135.6 (133.4, 136.5)	136.5 (135.6, 136.8)	0.183
CN-T1	2,417.85 ± 164.97	2,476.14 ± 162.35	2,354.88 ± 170.53	2,403.62 ± 129.14	2,459.67 ± 186.98	0.191
CN-T2	194.27 ± 16.56	202.14 ± 18.81	188.62 ± 13.76	190.54 ± 15.92	197.44 ± 15.21	0.108
CN-PD	140.15 (137.85, 142.12)	141.05 (139.22, 141.9)	139.65 (136.8, 141.9)	141.4 (139, 143.1)	140.2 (139.4, 141.1)	0.584
GP-T1	1,885.69 ± 107.65	1,907.64 ± 95.52	1,837.69 ± 113.88	1,875.77 ± 90.19	1,951.22 ± 109.95	0.06
GP-T2	229 ± 21.93	239.14 ± 17.16	217.94 ± 24.54	230.46 ± 24.02	230.78 ± 12.66	0.061
GP-PD	119.05 ± 2.16	118.86 ± 1.13	118.83 ± 1.68	119.94 ± 2.46	118.46 ± 3.38	0.376
SC–T1	2,279 (2,044.75, 2,499.5)	2,344 (2,164.5, 2,551.25)	2,236.5 (1,992, 2,525.5)	2,207 (1,979, 2,269)	2,421 (2,048, 2,632)	0.064
SC-T2	232.17 ± 28.67	243.43 ± 27.34	228.62 ± 24.44	227.31 ± 22.03	228 ± 43.28	0.406
SC-PD	103.53 ± 4.19	104.96 ± 3.92	103.21 ± 4.3	101.73 ± 3.58	104.44 ± 4.8	0.209
MCP-T1	2,165.9 ± 217.62	2,191.21 ± 250.66	2,126.38 ± 166.23	2,155.62 ± 242.58	2,211.67 ± 230.11	0.777
MCP-T2	173 (160, 189.25)	182.5 (165.25, 190.75)	164 (159.75, 185.25)	176 (162, 180)	164 (159, 184)	0.482
MCP-PD	101.66 ± 6.07	104.62 ± 3.94	100.66 ± 6.37	100.36 ± 6.76	100.72 ± 6.62	0.207

SCC, splenium of the corpus callosum; GCC, genu of the corpus callosum; IC, inferior colliculus; LL, lateral lemniscus; PLIC, posterior limb of the internal capsule; FL, frontal lobe; CN, caudate nucleus; GP, globus pallidus; SC, semioval center; MCP, middle cerebellar peduncle.

(Table 7). However, at 3 months, WMV, WMF, MYV, and MYF displayed distinctions among the four subgroups (Table 8). Subsequent pairwise comparisons of these parameters between each pair of groups at 1.5 and 3 months revealed that in line with above findings, at 1.5-month group, there were no significant difference across these subgroups (Figures 3A–D). At 3 -month group, WMV, MYV, and MYF demonstrated no variance between the normal and mild subgroups, whereas WMF decreased in the mild subgroup. Additionally, at 3 months, WMV, WMF, MYV, and MYF decreased in the moderate and severe subgroups compared to the normal subgroup (Figures 3E–H).

Correlation of inner ear malformations and SyMRI parameters

Above findings we found NH and children with IEM were associated with SNHL. Next, we explored the correlation of

10.3389/fnins.2024.1365141

TABLE 6 T1, T2, and PD values in different regions of the brain in the 3-month group.

Variables	Total ($n = 61$)	Normal ($n = 19$)	Mild ($n = 15$)	Moderate ($n = 19$)	Severe ($n = 8$)	p-value
SCC-T1	1,739 (1,673, 1,860)	1,704(1,667.5, 1,794.5)	1,702 (1,664, 1,881)	1,747 (1,683, 1,851)	1,780.5 (1,715, 1,828)	0.608
SCC-T2	171.82 ± 14.31	165.37 ± 10.72	176.33 ± 11	173.37 ± 18.4	175 ± 13.08	0.11
SCC-PD	83.9 (83, 85.9)	83.8 (83.3, 86.3)	83 (82.3, 84.5)	85.2 (83.4, 86.8)	84.1 (83.6, 84.78)	0.173
GCC-T1	1,639 (1,596, 1,662)	1,626 (1,599, 1,653)	1,638 (1,553, 1,662.5)	1,647 (1,613, 1,677)	1,634.5 (1,599.75, 1,654.5)	0.666
GCC-T2	145 (142, 155)	145 (142, 157)	147 (143, 158)	142 (137, 155)	143.5 (142.75, 148)	0.366
GCC-PD	82.8 (81.2, 83.7)	82.7 (80.9, 83.45)	82.8 (82, 83.6)	83 (82.1, 83.95)	82.85 (81.2, 83.42)	0.903
IC-T1	1,301 (1,248, 1,386)	1,251 (1,200, 1,294)	1,301 (1,244, 1,346)	1,314 (1,273, 1,441)	1,515.5 (1,403.5, 1,572.25)	< 0.001
IC-T2	119 (116, 124)	119 (117, 121)	119 (115, 120)	121 (116.5, 126)	118.5 (112.75, 130.25)	0.887
IC-PD	80.05 ± 3.08	81.02 ± 3.05	80.99 ± 3.07	78.88 ± 2.99	78.72 ± 2.37	0.053
LL-T1	1,272.8 ± 87.74	1,230.89 ± 87.53	1,246.07 ± 46.54	1,303.37 ± 84.85	1,349.88 ± 88.39	0.001
LL-T2	118.33 ± 7.53	117.68 ± 7.52	120.33 ± 5.95	119.05 ± 9.46	114.38 ± 3.29	0.316
LL-PD	81.08 ± 3.31	81.14 ± 3.35	79.73 ± 3.29	82.01 ± 3.35	81.28 ± 2.87	0.265
PLIC -T1	950 (900, 1,024)	966 (928, 1,012.5)	965 (904, 1,005.5)	969 (918, 1,058.5)	891 (885, 900)	0.137
PLIC -T2	106.28 ± 9.05	108.42 ± 8.08	107.87 ± 7	105.64 ± 11.51	99.75 ± 5.37	0.12
PLIC -PD	71 (69, 73.2)	72.2 (69.35, 74.35)	70.3 (69.1, 72.95)	71.4 (69.4, 72.65)	68.85 (68.4, 69.55)	0.286
FL-T1	1,886 (1,772, 1,997)	1,924 (1,837, 2,008.5)	1,886 (1,723.5, 1,958)	1,887 (1,803.5, 2,012)	1,776 (1,760, 1,830.5)	0.126
FL-T2	185 (172, 207)	196 (178.5, 215)	187 (171.5, 205)	183 (179, 207)	169 (164.5, 176)	0.085
FL-PD	86.7 (85.9, 87.3)	87 (85.95, 87.35)	86.5 (85.9, 87.35)	86.5 (85.3, 87.2)	86.8 (86.62, 87.62)	0.695
CN-T1	1,523.51 ± 112.33	1,548.37 ± 100.55	1,506.93 ± 125.6	1,544.53 ± 119.27	1,445.62 ± 60.33	0.121
CN-T2	139 (128, 148)	145 (129.5, 150)	133 (130, 143.5)	139 (127, 149.5)	129 (125, 134.25)	0.372
CN-PD	83.1 (81.3, 84.1)	83.3 (82.7, 83.9)	82.5 (81, 84.3)	81.6 (81.15, 84.35)	82.2 (81.6, 83.3)	0.463
GP-T1	1,453 (1,397, 1,520)	1,450 (1,393, 1,514.5)	1,450 (1,393, 1,524)	1,502 (1,406, 1,564.5)	1,438.5 (1,413.75, 1,453.25)	0.335
GP-T2	134.48 ± 13.67	136.95 ± 11.12	128.93 ± 14.46	138.74 ± 15.03	128.88 ± 10.8	0.096
GP-PD	84.09 ± 1.57	83.85 ± 1.17	84.21 ± 1.3	84.35 ± 2.06	83.81 ± 1.67	0.735
SC–T1	1,435 (1,307, 1,643)	1,435 (1,330, 1,540)	1,456 (1,307, 1,665.5)	1,405 (1,272, 1,567.5)	1,452.5 (1,252.5, 1,829.5)	0.975
SC-T2	151 (138, 163)	150 (144, 154.5)	150 (140.5, 162.5)	153 (131.5, 168.5)	163 (150.5, 170.25)	0.632
SC-PD	81.86 ± 3.48	82.32 ± 3.04	82.01 ± 3.4	81.52 ± 4.28	81.3 ± 2.89	0.868
MCP-T1	1,361.95 ± 127.27	1,323.79 ± 135.16	1,327.93 ± 136.38	1,383.95 ± 59.91	1,464.12 ± 160.52	0.032
MCP-T2	121 (112, 134)	124 (111.5, 135.5)	116 (111, 132)	124 (115, 134.5)	113.5 (110.25, 114)	0.252
MCP-D	79.05 ± 4.28	79.77 ± 4.03	78.89 ± 5.29	78.65 ± 4.18	78.61 ± 3.49	0.854

SCC, splenium of the corpus callosum; GCC, genu of the corpus callosum; IC, inferior colliculus; LL, lateral lemniscus; PLIC, posterior limb of the internal capsule; FL, frontal lobe; CN, caudate nucleus; GP, globus pallidus; SC, semioval center; MCP, middle cerebellar peduncle.

these two risk factors with SyMRI parameters, including IC-T1, LL-T1, MCP-T1, WMV, WMF, MYV, and MYF. Children were subgrouped according to whether they have this etiology, denoted as Normal-NH, Normal-Non-NH (Non-IEM), SNHL-NH (IEM), and SNHL-Non-NH (Non-IEM). Results demonstrated that both at 1.5 and 3-month groups, there was no significant difference in these parameters between Normal-NH and Normal-Non-NH (Supplementary Figures 1, 2). Instead, we found that SNHL-IEM showed high T1 values in IC and LL, while had low values of WMV, WMF, MYV, and MYF at 3 months, as compared with SNHL-Non-IEM, although there was no difference at 1.5-month group (Figures 4, 5).

Construction and validation of the prediction model

Based on the aforementioned findings, we identified seven parameters (IC-T1, LL-T1, MCP-T1, WMV, MYV, MYF, and

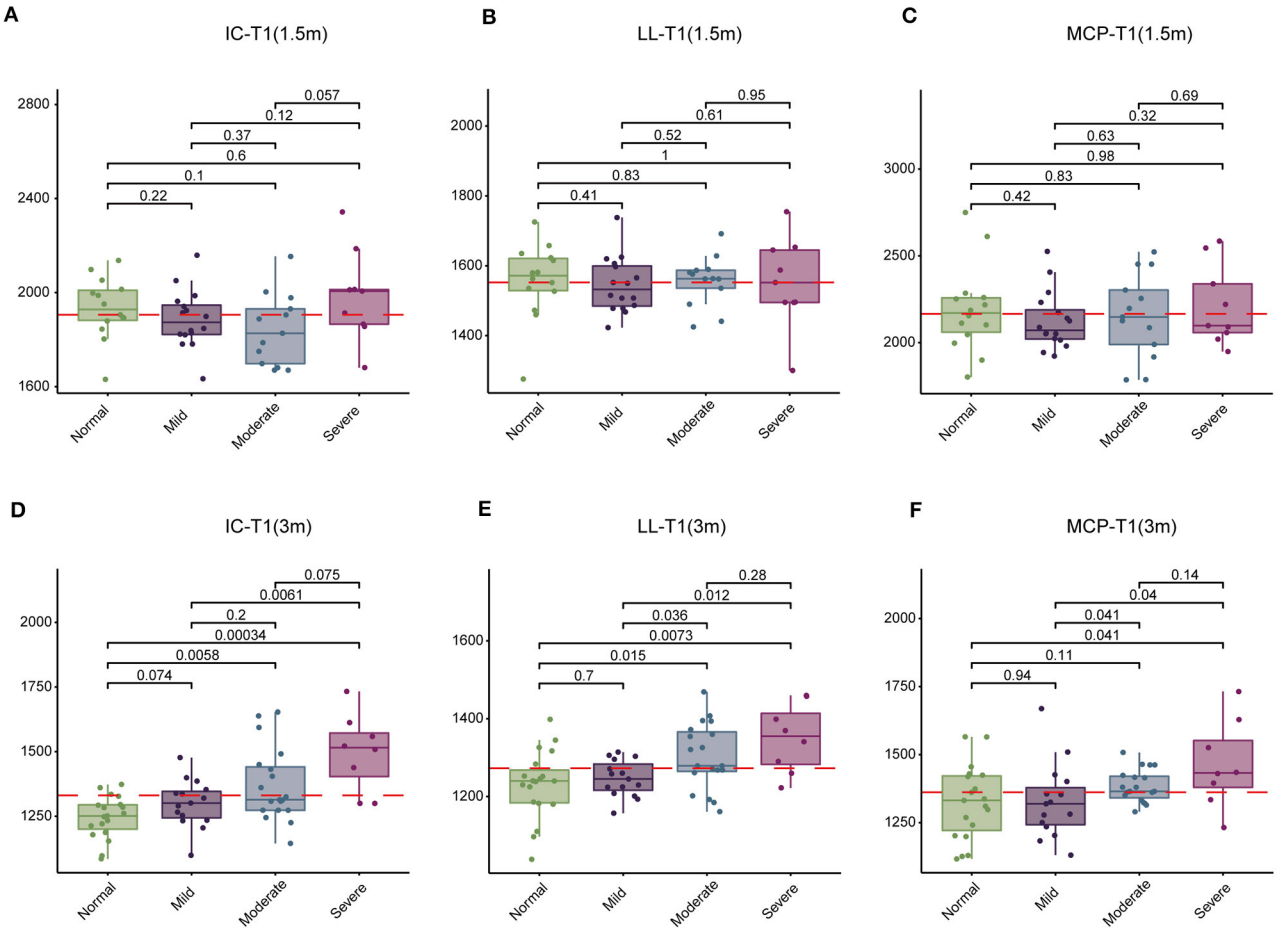


FIGURE 2
Comparison of T1 values in the inferior colliculus (IC), lateral lemniscus (LL), and middle cerebellar peduncle (MCP) among four groups of children at 1.5 and 3 months. **(A–C)** Comparison of T1 values among normal, mild, moderate, and severe subgroups at 1.5 months. **(D–F)** Comparison of T1 values among normal, mild, moderate, and severe subgroups at 3 months. T1 values are presented in milliseconds.

TABLE 7 **Summary of brain segmentation in the 1.5-month group.**

Variables	Total ($n = 52$)	Normal ($n = 14$)	Mild ($n = 16$)	Moderate ($n = 13$)	Severe ($n = 9$)	*p*-value
WMV	15.89 ± 3.95	16.63 ± 4.01	16 ± 5.08	15.62 ± 3.07	14.92 ± 2.97	0.788
GMV	507.65 (474.32, 540.88)	497.8 (481.33, 512.25)	507.65 (473.72, 521.68)	502.9 (444.3, 536.2)	548.8 (530.9, 558)	0.269
CSF	61.15 (55.88, 67.38)	61.25 (56.47, 63.08)	60.7 (56.7, 65.2)	61.2 (55.8, 68.5)	66.2 (52, 72.6)	0.986
NON	2.45 (2.08, 3)	2.6 (2.4, 2.88)	2.15 (1.8, 3.52)	2.2 (2.1, 3.3)	2.5 (2.1, 2.8)	0.62
MYV	2.86 ± 1.06	3.03 ± 1.06	2.88 ± 1.24	2.77 ± 1.16	2.68 ± 0.54	0.875
WMF	3.5 (2.8, 4.03)	3.3 (2.85, 4.18)	3.35 (2.9, 3.9)	3.7 (2.9, 4)	2.9 (2.6, 3.9)	0.734
GMF	95.5 (92.38, 96.8)	95.5 (94.15, 96.65)	95.5 (92.25, 96.8)	96.1 (92.4, 96.9)	95.2 (91.6, 96.8)	0.965
CSFF	10.6 ± 2.56	10.89 ± 1.52	10.69 ± 3.32	10.55 ± 2.78	10.03 ± 2.27	0.893
NONF	0.6 (0.5, 0.69)	0.6 (0.5, 0.69)	0.62 (0.55, 0.69)	0.6 (0.58, 0.61)	0.6 (0.5, 0.7)	0.939
MYF	0.6 (0.5, 0.7)	0.6 (0.5, 0.64)	0.6 (0.42, 0.79)	0.6 (0.4, 0.71)	0.6 (0.5, 0.63)	0.982

WMV, volume of white matter; GMV, volume of gray matter; CSF, cerebrospinal fluid; NON, non-WM/GM/CSF; MYV, myelination volume; WMF, white matter fraction; GMF, gray matter fraction; CSFF, cerebrospinal fluid fraction; NONF, NON fraction; MYF, myelin fraction.

TABLE 8 Summary of brain segmentation in the 3-month group.

Variables	Total ($n = 61$)	Normal ($n = 19$)	Mild ($n = 15$)	Moderate ($n = 19$)	Severe ($n = 8$)	p-value
WMV	32.9 (30.8, 34.6)	34.5 (32.05, 35.1)	34.2 (33.4, 34.8)	31.5 (29.2, 33.45)	29.15 (28.38, 30.35)	<0.001
GMV	607.26 ± 29.72	607.94 ± 44.2	607.96 ± 16.14	606.55 ± 24.85	606.01 ± 20.87	0.998
CSF	124.61 ± 22.52	114.48 ± 20.29	124.79 ± 14.03	128.93 ± 25.89	138.07 ± 25.14	0.056
NON	4.79 ± 1.05	4.85 ± 1.48	4.82 ± 0.65	4.81 ± 0.85	4.54 ± 0.98	0.912
MYV	8.3 (6.5, 8.9)	8.7 (6.8, 9.8)	8.6 (8.5, 9.4)	7.5 (6.6, 8.25)	5.05 (4.35, 6.25)	<0.001
WMF	5.2 (4.8, 5.9)	5.8 (5.25, 6.4)	5.2 (5.1, 5.6)	4.9 (4.7, 5.3)	4.45 (4.27, 4.65)	<0.001
GMF	93.08 ± 1.55	93.26 ± 1.78	93.07 ± 1.33	92.92 ± 1.5	93.03 ± 1.71	0.929
CSFF	13 (9.9, 14.2)	12.8 (10.5, 14)	12.1 (9.9, 14.3)	13.2 (9.55, 14.15)	12.2 (8.43, 14.93)	0.992
NONF	0.72 (0.65, 0.82)	0.8 (0.61, 1)	0.72 (0.7, 0.8)	0.76 (0.68, 0.84)	0.73 (0.68, 0.8)	0.889
MYF	1.1 (0.9, 1.4)	1.3 (1.1, 1.4)	1.21 (1.11, 1.4)	1 (0.9, 1.2)	0.8 (0.78, 0.83)	<0.001

WMV, volume of white matter; GMV, volume of gray matter; CSF, cerebrospinal fluid; NON, non-WM/GM/CSF; MYV, myelination volume; WMF, white matter fraction; GMF, gray matter fraction; CSFF, cerebrospinal fluid fraction; NONF, NON fraction; MYF, myelin fraction.

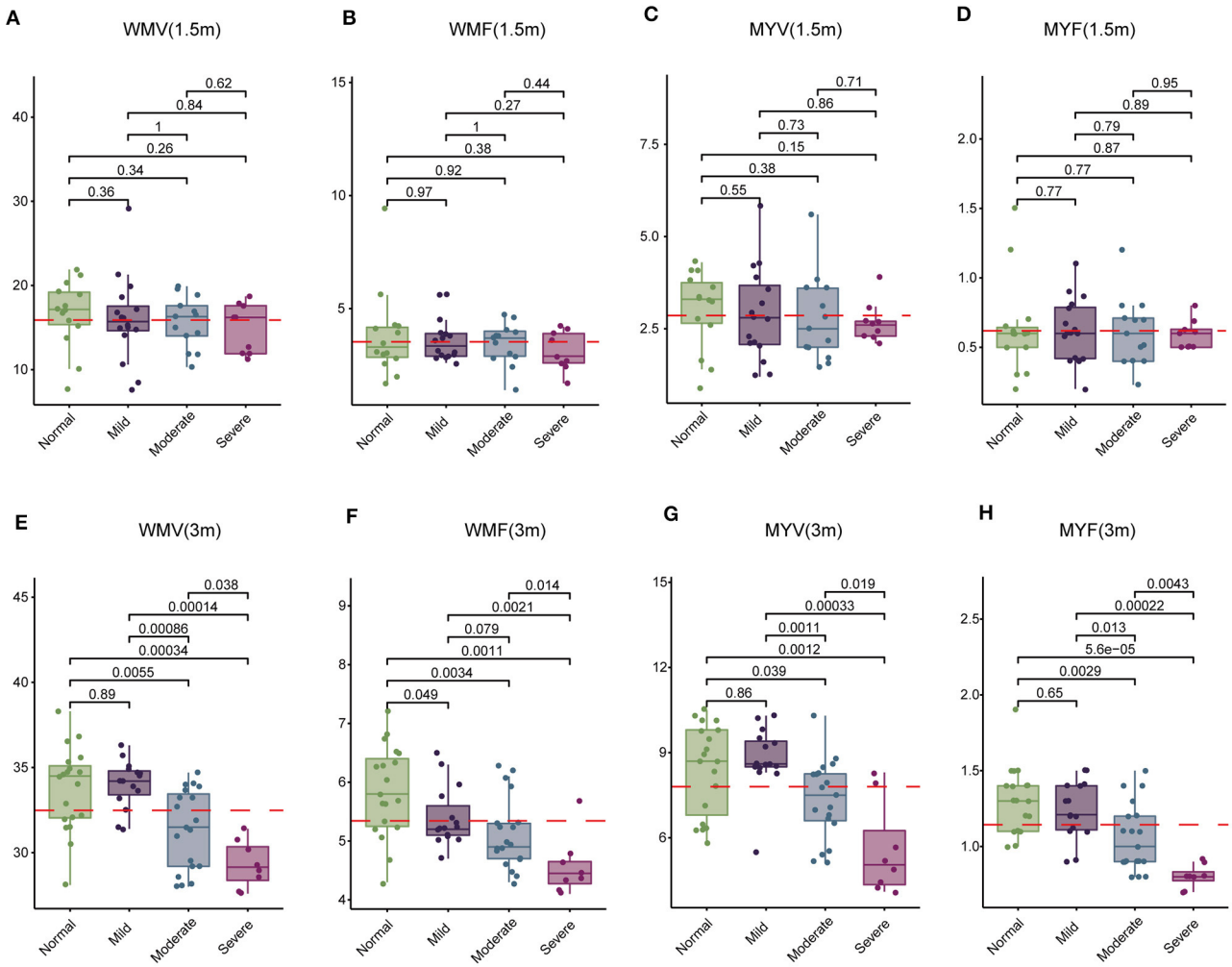


FIGURE 3
Comparison of white matter volume (WMV), white matter fraction (WMF), myelin volume (MYV), and myelin fraction (MYF) among normal, mild, moderate, and severe subgroups at 1.5 and 3 months. **(A–D)** Comparison at 1.5 months. **(E–H)** Comparison at 3 months.

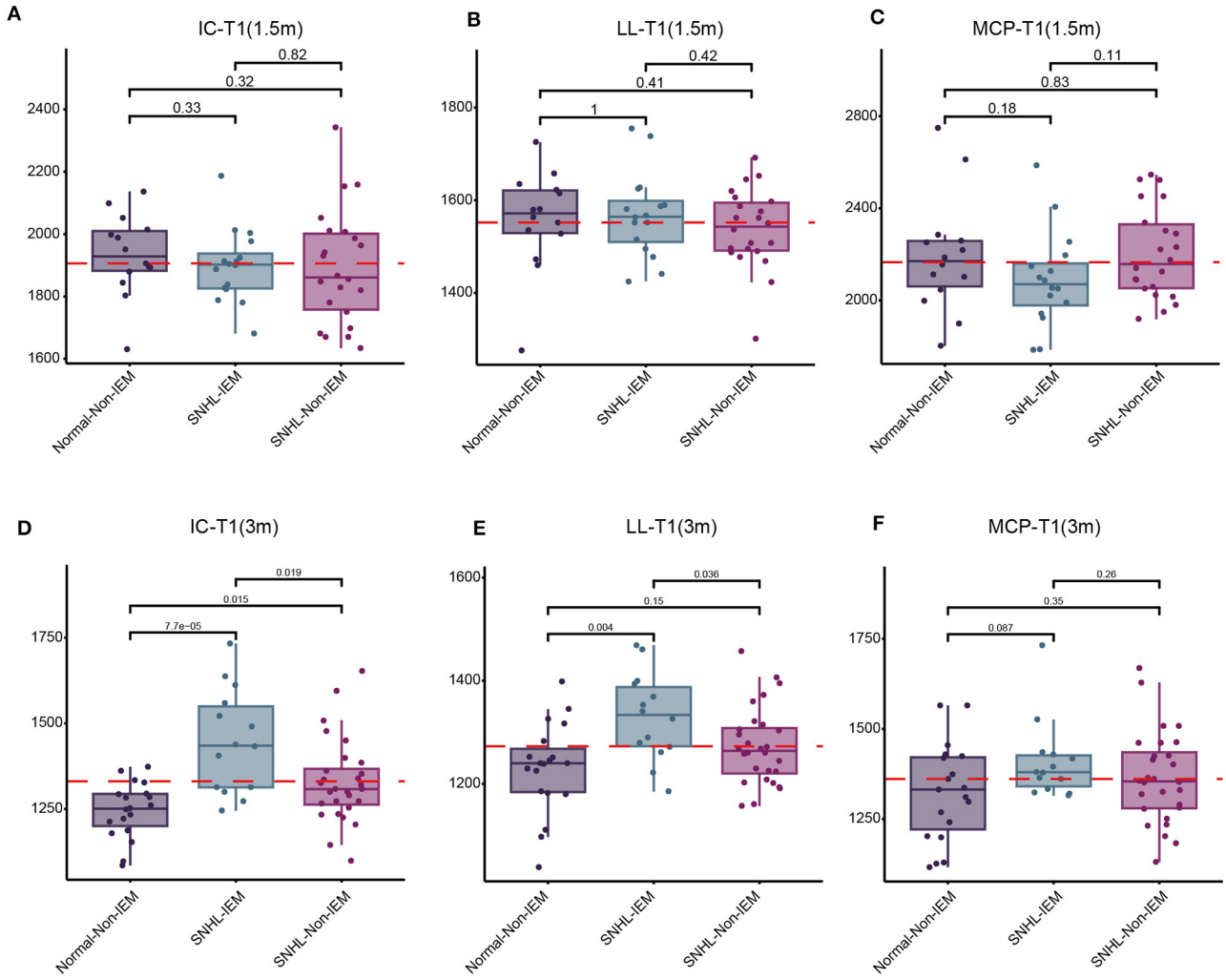


FIGURE 4
Comparison of T1 values in the inferior colliculus (IC), lateral lemniscus (LL), and middle cerebellar peduncle (MCP) among three groups of children at 1.5 and 3 months. **(A–C)** Comparison of T1 values among Normal-Non-IEM, SNHL-IEM, and SNHL-Non-IEM, subgroups at 1.5 months. **(D–F)** Comparison of T1 values among Normal-Non-IEM, SNHL-IEM, and SNHL-Non-IEM, subgroups at 3 months. T1 values are presented in milliseconds.

WMF) correlated with SNHL. Subsequently, we evaluated the predictive value of these parameters for SNHL. To achieve this, we randomly divided 61 samples at 3-month group into training and validation sets. Through univariate and multivariate analysis, we identified two independent risk factors, LL-T1 and WMF (Table 9). We then assessed the predictive performance of LL-T1 and WMF, resulting in respective AUCs of 0.620 and 0.800, respectively (Figures 6A, B). Next, we combined LL-T1 and WMF to construct a model, yielding AUCs of 0.865 and 0.806 for the training and validation sets, respectively (Figure 6C), indicating favorable performance. To further access the performance of the model, we conducted an external set. The AUC for external set was 0.736 (Figure 6D). To facilitate clinical application, we developed a nomogram for visualizing the model, enabling doctors to calculate predicted scores based on LL-T1 and WMF values and thereby predict the probability of SNHL (Figure 6E).

Discussion

SNHL manifests before language acquisition, potentially impeding linguistic development. The absence of auditory stimuli from birth in SNHL children may disrupt language learning and alter the formation of neural pathways, leading to structural changes in the brain (Chari and Chan, 2017). Late detection of hearing impairment in infants and young children with SNHL can result in profound learning and developmental challenges. Some studies have shown that high risk factors that correlated with onset of SNHL, including preterm birth, low birth weight infants, hyperbilirubinemia, cytomegalovirus infection, inner ear abnormalities, etc. (Wroblewska-Seniuk et al., 2018; Alhazmi, 2023). Our research results indicated that NH and IEM are high-risk factors for SNHL, possibly due to our analysis being not performed in the general population but in one tertiary care hospital, where there is a big neurological intensive.

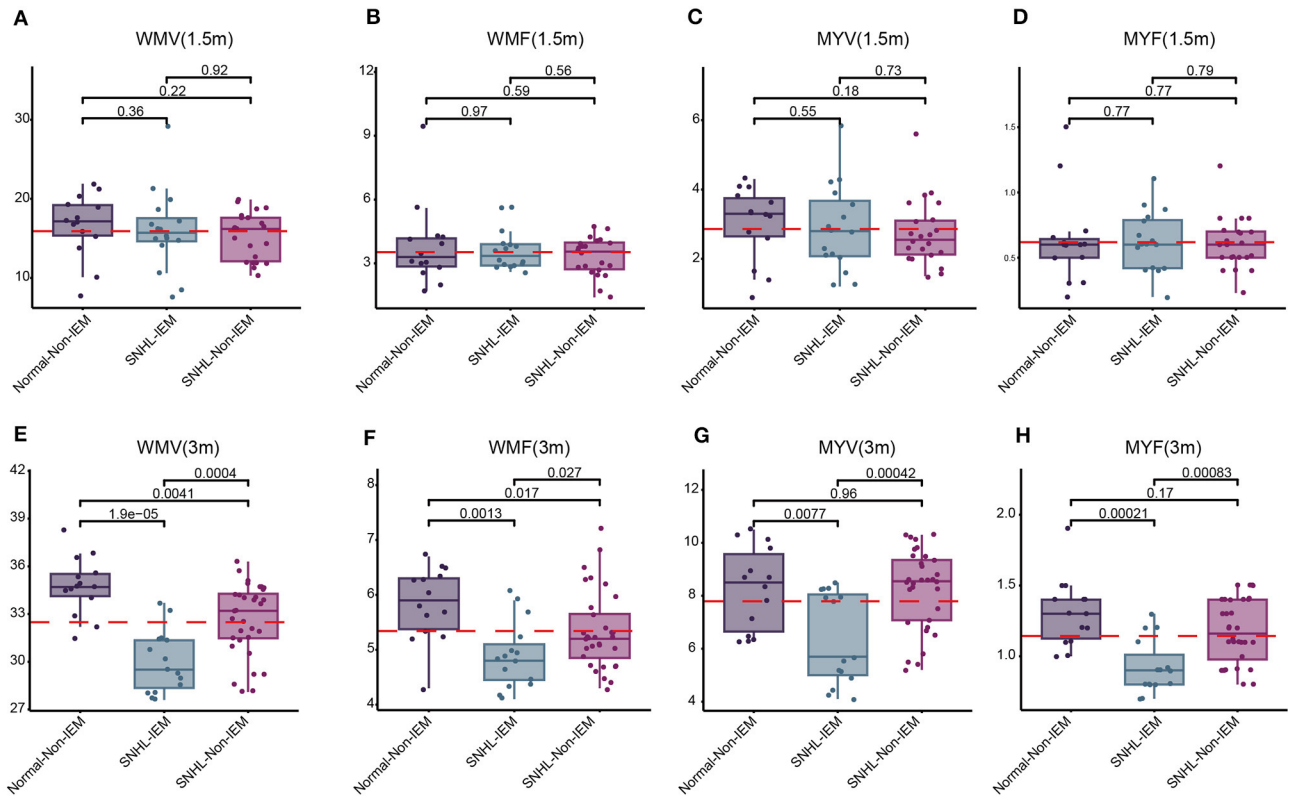


FIGURE 5
Comparison of white matter volume (WMV), white matter fraction (WMF), myelin volume (MYV), and myelin fraction (MYF) among Normal-Non-IEM, SNHL-IEM, and SNHL-Non-IEM, subgroups at 1.5 and 3 months. **(A–D)** Comparison at 1.5 months. **(E–H)** Comparison at 3 months.

TABLE 9 **Univariate and multivariate analysis of parameters correlated with SNHL.**

	Univariate analysis		Multivariate analysis	
Variables	OR (95%CI)	*P*-value	OR (95%CI)	*P*-value
WMV	0.63 (0.41–0.95)	0.029		
MYV	0.62 (0.36–1.05)	0.076		
WMF	0.17 (0.05–0.62)	0.007	0.12 (0.03–0.54)	0.006
MYF	0.02 (0.00–0.73)	0.033		
IC-T1	1.01 (1.00–1.01)	0.121		
LL-T1	1.01 (1.00–1.02)	0.129	1.01 (1.00–1.02)	0.070
MCP-T1	1.00 (1.00–1.01)	0.203		

WMV, volume of white matter; MYV, myelination volume; WMF, white matter fraction; MYF, myelin fraction; IC, inferior colliculus; LL, lateral lemniscus; MCP, middle cerebellar peduncle.

Previous investigations into macrostructural disparities between deaf individuals and those without hearing loss revealed diminished WMV but preserved gray matter volume in the auditory cortex, particularly in Heschl's gyrus (HG) and the adjacent temporal lobe. However, WMV exhibited inconsistent patterns across left and right brain hemispheres, with the most significant differences observed in the right posterior superior temporal gyrus (Hribar et al., 2014; Karns et al., 2017). Moreover, microstructural changes in white matter have been documented in individuals with hearing loss during various stages of life (Miao et al., 2013; Park et al., 2018; Kim et al., 2021). Yet, limited knowledge exists regarding white matter microstructural properties in children with SNHL. Research suggests that the gray matter volume of the right hemisphere, alongside white matter volume, is more susceptible to impairment compared to the left hemisphere (Manno et al., 2021). In our study, we found that quantitative T1 values were higher in children with SNHL than in their normally hearing counterparts. These discrepancies were observed in multiple brain regions, including the IC, LL, and MCP. Elevated quantitative T1 values signify alterations in myelin microstructure. Notably, our investigation is the first to report differences in T1 values among children with SNHL

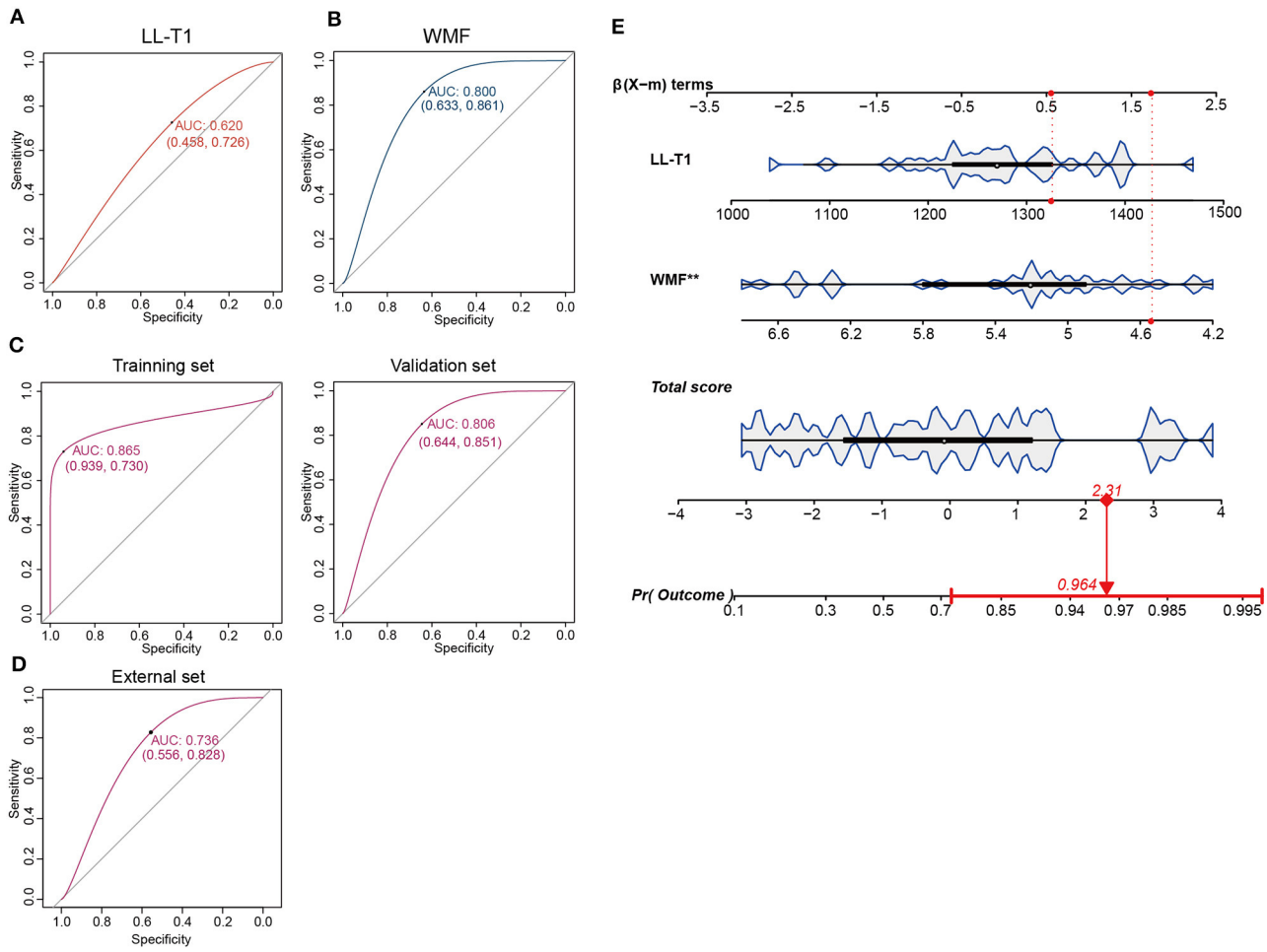


FIGURE 6
Construction and validation of the prediction model. **(A, B)** Area under the curve (AUC) for lateral lemniscus T1 value (LL-T1) and white matter fraction (WMF). **(C)** AUC for the training and validation sets. **(D)** AUC for the external set. **(E)** Nomogram illustrating the model for clinical application.

within the first 3 months of life. Analysis of whole brain volume revealed lower values of WMV, WMF, MYV, and MYF in children born with SNHL. Additionally, these parameters displayed a positive correlation with age. Compared to the control group, children with SNHL exhibited reduced WMV, WMF, MYV, and MYF.

Brain development follows a sequential pattern, with myelin sheath formation initiating around the fifth month of fetal development and progressing alongside central nervous system myelination, continuing throughout life (Mukherjee et al., 2001). White matter myelination typically commences between 6 and 8 months, with most white matter achieving maturity in myelin sheath formation by 18 months (Mukherjee et al., 2001; Barkovich, 2005; Lebel and Deoni, 2018). The observed increase in T1 values could be attributed to hearing impairment, which may impede the normal pace of development and maturation in these regions. Literature suggests that white matter development in infants follows a trajectory from dorsal to ventral, caudal to cephalic, and from central to peripheral regions (Shi et al., 2019). Early developmental activity is notable in areas such as the IC, LL, and MCP.

Assessing myelination plays a pivotal role in evaluating neurological development (Khelfaoui et al., 2024). SyMRI offers enhanced capabilities in detecting MS plaques compared to conventional MRI methods (Miller et al., 1998; Granberg et al., 2016; Hagiwara et al., 2017). Utilizing synthetic DIR and PSIR images can facilitate the identification of intra-cortical or mixed white matter-gray matter lesions (Miller et al., 1998). Vagberg et al. demonstrated the validity and reproducibility of SyMRI volumetric analysis in determining BPF in MS (Vagberg et al., 2013, 2016). In pediatric MS, BPF is notably lower compared to adult MS, primarily attributed to gray matter loss (Vagberg et al., 2013). Notably, our study found no significant differences in brain segmentation-related indices between the control and SNHL groups at 1.5 months, suggesting a relatively minor impact of SNHL on brain development at this early age. However, by 3 months, we observed no significant differences between the control and mild SNHL groups, indicating a minor effect of mild SNHL on brain development. In contrast, the moderate and severe SNHL groups exhibited significant reductions in WMV and myelination-related measures, indicating distinct structural alterations with increasing severity of SNHL. This corroborates

previous research; Smith et al. (2011) observed decreased white matter in the anterior Heschl's gyrus in individuals with hearing loss using whole-brain voxel-based morphometry. Notably, our study focused on children as young as 3 months, utilizing SyMRI to detect subtle changes in brain development associated with SNHL, underscoring the impact of early hearing abnormalities on neurological development.

Kim et al. (2017) identified that the myelin volume percentage automatically generated by SyMRI within the brain substance volume closely adhered to the established myelin maturation Gompertz model and exhibited strong correlations with R1 and R2 relaxation rates. The quantification of myelin using SyMRI presents a promising avenue for assessing brain development in children. In a study utilizing VBM, Hribar et al. (2020) observed a significant decrease in WM volume within the left medial frontal gyrus and the right suboccipital gyrus in deaf patients, with no notable difference in gray matter volume, aligning with our findings. Notably, our study unveils differences in white matter occurring before language development, particularly in subjects around 3 months old with moderate-to-severe SNHL, a phenomenon not documented in existing literature. These findings suggest potential neuroplastic changes linked to brain reorganization following early hearing deprivation in SNHL infants.

A nomogram is a graphical tool which is commonly used to estimate prognosis in oncology and medicine. With the ability to generate an individual numerical probability of a clinical event by integrating diverse prognostic and determinant variables. Rapid computation through user friendly digital interfaces, together with increased accuracy, and more easily understood prognoses, allow for seamless incorporation of nomogram derived prognosis to aid in clinical decision making. This has led to the ubiquitous appearance of nomogram in clinical use (Ohori Tatsuo et al., 2009; Balachandran et al., 2015; Gandaglia et al., 2019). This study we constructed a prediction model based on two factors key SyMRI quantitative parameters LL-T1 and WMF, for distinguishing SNHL. This is easy for clinical doctors to calculate quantitative values of LL-T1 and WMF and arrange them horizontally on the column chart with scaled line segments in their respective proportions. By calculating the total score values corresponding to each parameter and finding the corresponding predicted risk values below the total score scale, we can quickly obtain the prediction probability for SNHL.

The molecular mechanisms through which sensorineural hearing loss (SNHL) impacts brain development are still not fully understood. Traditionally, it was believed that SNHL primarily targets hair cells, with cochlear nerve loss considered secondary to hair cell degeneration. However, in cases of noise-induced hearing loss, even reversible threshold shifts (without hair cell loss) can result in permanent loss of over 50% of cochlear nerve/hair cell synapses. Similarly, in age-related hearing loss, the degeneration of cochlear synapses precedes both hair cell loss and threshold elevation (Kujawa and Liberman, 2015). There are reports indicating the possibility of spontaneous re-innervation (Puel, 1995; Pujol and Puel, 1999; Sun et al., 2001), or that some immediate synapse loss may be reversible (Liu et al., 2012; Shi et al., 2013, 2015, 2016). However, how these molecular changes manifest in SyMRI imaging remains unclear. Ongoing research aims to delve deeper into this phenomenon, elucidate its underlying mechanisms, and evaluate the potential effectiveness of therapeutic interventions.

This study is subject to certain limitations. Firstly, the sample size was relatively small, potentially impacting the statistical power and the generalizability of the research findings. Secondly, there was no follow-up conducted to assess the long-term intellectual and behavioral development of the SNHL patient group. Long-term follow-up could shed light on the impact of SNHL on various aspects such as language proficiency, motor skills, and learning abilities across different age groups, highlighting the necessity for further investigation. Thirdly, due to our hospital being a provincial key maternal and child health hospital, there may be bias in sample selection.

Conclusion

In conclusion, T1 values, coupled with measurements of WMV, MYV, WMF, and MYF, hold promise as potential indicators for early detection of brain development anomalies in children with SNHL. Quantitative assessments in areas such as the IC, LL, and MCP could assist in distinguishing patients with moderate to severe SNHL. Moreover, observed reductions in WMV and myelin levels may serve as predictive factors for the progression of moderate and severe SNHL in pediatric populations.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving humans were approved by The Third Affiliated Hospital of Zhengzhou University. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

Author contributions

PZ: Investigation, Writing – original draft, Writing – review & editing, Data curation, Formal analysis, Methodology, Project administration, Resources, Software, Validation, Visualization. JY: Data curation, Formal analysis, Methodology, Project administration, Software, Writing – original draft. YS: Data curation, Formal analysis, Investigation, Methodology, Software, Writing – original draft. MC: Formal analysis, Methodology, Software, Writing – review & editing. XZhao: Formal analysis, Methodology, Writing – review & editing. KW: Methodology, Writing – review & editing, Data curation. LL: Methodology, Writing – review & editing, Software. QX: Software, Writing – review & editing. GN: Writing – review & editing, Data curation. LM: Writing – review & editing, Methodology. XW: Methodology,

Writing – review & editing. LZ: Writing – review & editing, Formal analysis. XZhan: Writing – review & editing, Conceptualization, Funding acquisition, Investigation, Supervision, Writing – original draft.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This study received support from the National Natural Science Foundation of China (Grant No. 82371929), Zhengzhou Science and Technology Bureau Collaborative Innovation Major Project (Grant No. 18XTZX12009), and PhD Research Startup Foundation of the Third Affiliated Hospital of Zhengzhou University (Grant No. BS 20230112).

Acknowledgments

The authors express gratitude to the members of the Department of Otology Medicine, Division of Logopedics, for their valuable assistance.

Conflict of interest

KW is employed by GE Healthcare.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer CYL declared a shared affiliation with the author KW at the time of review.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2024.1365141/full#supplementary-material

SUPPLEMENTARY FIGURE 1
Comparison of T1 values in the inferior colliculus (IC), lateral lemniscus (LL), and middle cerebellar peduncle (MCP) among four groups of children at 1.5 and 3 months. **(A–C)** Comparison of T1 values among Normal-NH, Normal-Non-NH, SNHL-NH, and SNHL-Non-NH, subgroups at 1.5 months. **(D–F)** Comparison of T1 values among Normal-NH, Normal-Non-NH, SNHL-NH, and SNHL-Non-NH, subgroups at 3 months. T1 values are presented in milliseconds.

SUPPLEMENTARY FIGURE 2
Comparison of white matter volume (WMV), white matter fraction (WMF), myelin volume (MYV), and myelin fraction (MYF) among Normal-NH, Normal-Non-NH, SNHL-NH, and SNHL-Non-NH, subgroups at 1.5 and 3 months. **(A–D)** Comparison at 1.5 months. **(E–H)** Comparison at 3 months.

References

Alhazmi, W. (2023). Risk factors associated with hearing impairment in infants and children: a systematic review. *Cureus* 15:e40464. doi: 10.7759/cureus.40464

Andica, C., Hagiwara, A., Nakazawa, M., Kumamaru, K. K., Hori, M., Ikeno, M., et al. (2017). Synthetic MR imaging in the diagnosis of bacterial meningitis. *Magn. Reson. Med. Sci.* 16, 91–92. doi: 10.2463/mrms.ci.2016-0082

Andica, C., Hagiwara, A., Nakazawa, M., Tsuruta, K., Takano, N., Hori, M., et al. (2016). The advantage of synthetic MRI for the visualization of early white matter change in an infant with Sturge-Weber Syndrome. *Magn. Reson. Med. Sci.* 15, 347–348. doi: 10.2463/mrms.ci.2015-0164

Badve, C., Yu, A., Dastmalchian, S., Rogers, M., Ma, D., Jiang, Y., et al. (2017). MR fingerprinting of adult brain tumors: initial experience. *Am. J. Neuroradiol.* 38, 492–499. doi: 10.3174/ajnr.A5035

Balachandran, V. P., Gonen, M., Smith, J. J., and Dematteo, R. P. (2015). Nomograms in oncology: more than meets the eye. *Lancet Oncol.* 16, e173–e180. doi: 10.1016/S1470-2045(14)71116-7

Barkovich, A. J. (2005). Magnetic resonance techniques in the assessment of myelin and myelination. *J. Inherit. Metab. Dis.* 28, 311–343. doi: 10.1007/s10545-005-5952-z

Chari, D. A., and Chan, D. K. (2017). Diagnosis and treatment of congenital sensorineural hearing loss. *Curr. Otorhinolaryngol. Rep.* 5, 251–258. doi: 10.1007/s40136-017-0163-3

Chen, Y., Su, S., Dai, Y., Wen, Z., Qian, L., Zhang, H., et al. (2021). Brain volumetric measurements in children with attention deficit hyperactivity disorder: a comparative study between synthetic and conventional magnetic resonance imaging. *Front. Neurosci.* 15:711528. doi: 10.3389/fnins.2021.711528

Gandaglia, G., Ploussard, G., Valerio, M., Mattei, A., Fiori, C., Fossati, N., et al. (2019). A novel nomogram to identify candidates for extended pelvic lymph node dissection among patients with clinically localized prostate cancer diagnosed with magnetic resonance imaging-targeted and systematic biopsies. *Eur. Urol.* 75, 506–514. doi: 10.1016/j.eururo.2018.10.012

Goncalves, F. G., Serai, S. D., and Zuccoli, G. (2018). Synthetic brain MRI: review of current concepts and future directions. *Top. Magn. Reson. Imaging* 27, 387–393. doi: 10.1097/RMR.0000000000000189

Granberg, T., Uppman, M., Hashim, F., Cananau, C., Nordin, L. E., Shams, S., et al. (2016). Clinical feasibility of synthetic MRI in multiple sclerosis: a diagnostic and volumetric validation study. *Am. J. Neuroradiol.* 37, 1023–1029. doi: 10.3174/ajnr.A4665

Gulani, V., Schmitt, P., Griswold, M. A., Webb, A. G., and Jakob, P. M. (2004). Towards a single-sequence neurologic magnetic resonance imaging examination: multiple-contrast images from an IR TrueFISP experiment. *Invest. Radiol.* 39, 767–774. doi: 10.1097/00004424-200412000-00008

Hagiwara, A., Hori, M., Yokoyama, K., Takemura, M. Y., Andica, C., Tabata, T., et al. (2017). Synthetic MRI in the detection of multiple sclerosis plaques. *Am. J. Neuroradiol.* 38, 257–263. doi: 10.3174/ajnr.A5012

Hribar, M., Suput, D., Battelino, S., and Vovk, A. (2020). Review article: structural brain alterations in prelingually deaf. *Neuroimage* 220:117042. doi: 10.1016/j.neuroimage.2020.117042

Hribar, M., Suput, D., Carvalho, A. A., Battelino, S., and Vovk, A. (2014). Structural alterations of brain grey and white matter in early deaf adults. *Hear. Res.* 318, 1–10. doi: 10.1016/j.heares.2014.09.008

Ji, S., Yang, D., Lee, J., Choi, S. H., Kim, H., and Kang, K. M. (2022). Synthetic MRI: technologies and applications in neuroradiology. *J. Magn. Reson. Imaging* 55, 1013–1025. doi: 10.1002/jmri.27440

Johnson, J. C. S., Marshall, C. R., Weil, R. S., Bamiou, D. E., Hardy, C. J. D., and Warren, J. D. (2021). Hearing and dementia: from ears to brain. *Brain* 144, 391–401. doi: 10.1093/brain/awaa429

Karns, C. M., Stevens, C., Dow, M. W., Schorr, E. M., and Nevil,LE., H. J. (2017). Atypical white-matter microstructure in congenitally deaf adults: a region of interest and tractography study using diffusion-tensor imaging. *Hear. Res.* 343, 72–82. doi: 10.1016/j.heares.2016.07.008

Khelfaoui, H., Ibaceta-Gonzalez, C., and Angulo, M. C. (2024). Functional myelin in cognition and neurodevelopmental disorders. *Cell. Mol. Life Sci.* 81:181. doi: 10.1007/s00018-024-05222-2

Kim, E., Kang, H., Han, K. H., Lee, H. J., Suh, M. W., Song, J. J., et al. (2021). Reorganized brain white matter in early- and late-onset deafness with diffusion tensor imaging. *Ear Hear.* 42, 223–234. doi: 10.1097/AUD.0000000000000917

Kim, H. G., Moon, W. J., Han, J., and Choi, J. W. (2017). Quantification of myelin in children using multiparametric quantitative MRI: a pilot study. *Neuroradiology* 59, 1043–1051. doi: 10.1007/s00234-017-1889-9

Korver, A. M., Konings, S., Dekker, F. W., Beers, M., Wever, C. C., Frijns, J. H., et al. (2010). Newborn hearing screening vs later hearing screening and developmental outcomes in children with permanent childhood hearing impairment. *JAMA* 304, 1701–1708. doi: 10.1001/jama.2010.1501

Kujawa, S. G., and Liberman, M. C. (2015). Synaptopathy in the noise-exposed and aging cochlea: primary neural degeneration in acquired sensorineural hearing loss. *Hear. Res.* 330, 191–199. doi: 10.1016/j.heares.2015.02.009

Lebel, C., and Deoni, S. (2018). The development of brain white matter microstructure. *Neuroimage* 182, 207–218. doi: 10.1016/j.neuroimage.2017.12.097

Liu, L., Wang, H., Shi, L., Almuklass, A., He, T., Aiken, S., et al. (2012). Silent damage of noise on cochlear afferent innervation in guinea pigs and the impact on temporal processing. *PLoS ONE* 7:e49550. doi: 10.1371/journal.pone.0049550

Manno, F. A. M., Rodriguez-Cruces, R., Kumar, R., Ratnanather, J. T., and Lau, C. (2021). Hearing loss impacts gray and white matter across the lifespan: systematic review, meta-analysis and meta-regression. *Neuroimage* 231:117826. doi: 10.1016/j.neuroimage.2021.117826

Miao, W., Li, J., Tang, M., Xian, J., Li, W., Liu, Z., et al. (2013). Altered white matter integrity in adolescents with prelingual deafness: a high-resolution tract-based spatial statistics imaging study. *Am. J. Neuroradiol.* 34, 1264–1270. doi: 10.3174/ajnr.A3370

Miller, D. H., Grossman, R. I., Reingold, S. C., and Mcfarland, H. F. (1998). The role of magnetic resonance techniques in understanding and managing multiple sclerosis. *Brain* 121 (Pt 1), 3–24. doi: 10.1093/brain/121.1.3

Mukherjee, P., Miller, J. H., Shimony, J. S., Conturo, T. E., Lee, B. C., Almli, C. R., et al. (2001). Normal brain maturation during childhood: developmental trends characterized with diffusion-tensor MR imaging. *Radiology* 221, 349–358. doi: 10.1148/radiol.2212001702

Nunez-Gonzalez, L., Van Garderen, K. A., Smits, M., Jaspers, J., Romero, A. M., Poot, D. H. J., et al. (2022). Pre-contrast MAGiC in treated gliomas: a pilot study of quantitative MRI. *Sci. Rep.* 12:21820. doi: 10.1038/s41598-022-24276-5

Ohori Tatsuo, G., Riu Hamada, M., Gondo, T., and Hamada, R. (2009). Nomogram as predictive model in clinical practice. *Gan To Kagaku Ryoho* 36, 901–906.

Park, K. H., Chung, W. H., Kwon, H., and Lee, J. M. (2018). Evaluation of cerebral white matter in prelingually deaf children using diffusion tensor imaging. *Biomed Res. Int.* 2018:6795397. doi: 10.1155/2018/6795397

Puel, J. L. (1995). Chemical synaptic transmission in the cochlea. *Prog. Neurobiol.* 47, 449–476. doi: 10.1016/0301-0082(95)00028-3

Pujol, R., and Puel, J. L. (1999). Excitotoxicity, synaptic repair, and functional recovery in the mammalian cochlea: a review of recent findings. *Ann. N. Y. Acad. Sci.* 884, 249–254. doi: 10.1111/j.1749-6632.1999.tb08646.x

Shende, S. A., and Mudar, R. A. (2023). Cognitive control in age-related hearing loss: a narrative review. *Hear. Res.* 436:108814. doi: 10.1016/j.heares.2023.108814

Shi, J., Yang, S., Wang, J., Huang, S., Yao, Y., Zhang, S., et al. (2019). Detecting normal pediatric brain development with diffusional kurtosis imaging. *Eur. J. Radiol.* 120:108690. doi: 10.1016/j.ejrad.2019.108690

Shi, L., Chang, Y., Li, X., Aiken, S. J., Liu, L., and Wang, J. (2016). Coding deficits in noise-induced hidden hearing loss may stem from incomplete repair of ribbon synapses in the cochlea. *Front. Neurosci.* 10:231. doi: 10.3389/fnins.2016.00231

Shi, L., Liu, K., Wang, H., Zhang, Y., Hong, Z., Wang, M., et al. (2015). Noise induced reversible changes of cochlear ribbon synapses contribute to temporary hearing loss in mice. *Acta Otolaryngol.* 135, 1093–1102. doi: 10.3109/00016489.2015.1061699

Shi, L., Liu, L., He, T., Guo, X., Yu, Z., Yin, S., et al. (2013). Ribbon synapse plasticity in the cochleae of Guinea pigs after noise-induced silent damage. *PLoS ONE* 8:e81566. doi: 10.1371/journal.pone.0081566

Slade, K., Plack, C. J., and Nuttall, H. E. (2020). The effects of age-related hearing loss on the brain and cognitive function. *Trends Neurosci.* 43, 810–821. doi: 10.1016/j.tins.2020.07.005

Smith, K. M., Mecoli, M. D., Altaye, M., Komlos, M., Maitra, R., Eaton, K. P., et al. (2011). Morphometric differences in the Heschl's gyrus of hearing impaired and normal hearing infants. *Cereb. Cortex* 21, 991–998. doi: 10.1093/cercor/bhq164

Sun, H., Hashino, E., Ding, D. L., and Salvi, R. J. (2001). Reversible and irreversible damage to cochlear afferent neurons by kainic acid excitotoxicity. *J. Comp. Neurol.* 430, 172–181. doi: 10.1002/1096-9861(20010205)430:2<172::AID-CNE1023>3.0.CO;2-W

Surprenant, A. M., and Didonato, R. (2014). Community-dwelling older adults with hearing loss experience greater decline in cognitive function over time than those with normal hearing. *Evid. Based Nurs.* 17, 60–61. doi: 10.1136/eb-2013-101375

Vagberg, M., Ambarki, K., Lindqvist, T., Birgander, R., and Svenningsson, A. (2016). Brain parenchymal fraction in an age-stratified healthy population - determined by MRI using manual segmentation and three automated segmentation methods. *J. Neuroradiol.* 43, 384–391. doi: 10.1016/j.neurad.2016.08.002

Vagberg, M., Lindqvist, T., Ambarki, K., Warntjes, J. B., Sundstrom, P., Birgander, R., et al. (2013). Automated determination of brain parenchymal fraction in multiple sclerosis. *Am. J. Neuroradiol.* 34, 498–504. doi: 10.3174/ajnr.A3262

Van Der Weijden, C. W. J., Biondetti, E., Gutmann, I. W., Dijkstra, H., Mckerchar, R., De Paula Faria, D., et al. (2023). Quantitative myelin imaging with MRI and PET: an overview of techniques and their validation status. *Brain* 146, 1243–1266. doi: 10.1093/brain/awac436

Vanderhasselt, T., Naeyaert, M., Watte, N., Allemeersch, G. J., Raeymaeckers, S., Dudink, J., et al. (2020). Synthetic MRI of preterm infants at term-equivalent age: evaluation of diagnostic image quality and automated brain volume segmentation. *Am. J. Neuroradiol.* 41, 882–888. doi: 10.3174/ajnr.A6533

Wang, H., Liang, Y., Fan, W., Zhou, X., Huang, M., Shi, G., et al. (2019). DTI study on rehabilitation of the congenital deafness auditory pathway and speech center by cochlear implantation. *Eur. Arch. Otorhinolaryngol.* 276, 2411–2417. doi: 10.1007/s00405-019-05477-7

Wang, Y., Xiong, W., Sun, X., Lu, K., Duan, F., Wang, H., et al. (2023). Impact of environmental noise exposure as an inducing factor on the prognosis of sudden sensorineural hearing loss: a retrospective case-control study. *Front. Neurosci.* 17:1210291. doi: 10.3389/fnins.2023.1210291

Warntjes, J. B., Leinhard, O. D., West, J., and Lundberg, P. (2008). Rapid magnetic resonance quantification on the brain: optimization for clinical usage. *Magn. Reson. Med.* 60, 320–329. doi: 10.1002/mrm.21635

West, J., Warntjes, J. B., and Lundberg, P. (2012). Novel whole brain segmentation and volume estimation using quantitative MRI. *Eur. Radiol.* 22, 998–1007. doi: 10.1007/s00330-011-2336-7

Wroblewska-Seniuk, K., Dabrowski, P., Greczka, G., Szabatowska, K., Glowacka, A., Szyfter, W., et al. (2018). Sensorineural and conductive hearing loss in infants diagnosed in the program of universal newborn hearing screening. *Int. J. Pediatr. Otorhinolaryngol.* 105, 181–186. doi: 10.1016/j.ijporl.2017.12.007

Yeh, E. A., Weinstock-Guttman, B., Ramanathan, M., Ramasamy, D. P., Willis, L., Cox, J. L., et al. (2009). Magnetic resonance imaging characteristics of children and adults with paediatric-onset multiple sclerosis. *Brain* 132, 3392–3400. doi: 10.1093/brain/awp278

frontiers | Frontiers in Computational Neuroscience

TYPE Original Research
PUBLISHED 20 September 2024
DOI 10.3389/fncom.2024.1360095

OPEN ACCESS

EDITED BY
Noemi Montobbio,
University of Genoa, Italy

REVIEWED BY
Minh Hoang Gia,
Gwangju Institute of Science and Technology, Republic of Korea
Shuqiang Wang,
Chinese Academy of Sciences (CAS), China

*CORRESPONDENCE
Rosanna Turrisi
✉ rosanna.turrisi@edu.unige.it

RECEIVED 22 December 2023
ACCEPTED 03 September 2024
PUBLISHED 20 September 2024

CITATION
Turrisi R, Verri A and Barla A (2024) Deep learning-based Alzheimer's disease detection: reproducibility and the effect of modeling choices.
Front. Comput. Neurosci. 18:1360095.
doi: 10.3389/fncom.2024.1360095

Deep learning-based Alzheimer's disease detection: reproducibility and the effect of modeling choices

Rosanna Turrisi[1,2]*, Alessandro Verri[1,2] and Annalisa Barla[1,2] for the Alzheimer's Disease Neuroimaging Initiative

[1]Department of Informatics, Bioengineering, Robotics and System Engineering (DIBRIS), University of Genoa, Genoa, Italy, [2]Machine Learning Genoa (MaLGa) Center, University of Genoa, Genoa, Italy

Introduction: Machine Learning (ML) has emerged as a promising approach in healthcare, outperforming traditional statistical techniques. However, to establish ML as a reliable tool in clinical practice, adherence to best practices in *data handling*, and *modeling design and assessment* is crucial. In this work, we summarize and strictly adhere to such practices to ensure reproducible and reliable ML. Specifically, we focus on Alzheimer's Disease (AD) detection, a challenging problem in healthcare. Additionally, we investigate the impact of modeling choices, including different data augmentation techniques and model complexity, on overall performance.

Methods: We utilize Magnetic Resonance Imaging (MRI) data from the ADNI corpus to address a binary classification problem using 3D Convolutional Neural Networks (CNNs). Data processing and modeling are specifically tailored to address data scarcity and minimize computational overhead. Within this framework, we train 15 predictive models, considering three different data augmentation strategies and five distinct 3D CNN architectures with varying convolutional layers counts. The augmentation strategies involve affine transformations, such as *zoom*, *shift*, and *rotation*, applied either concurrently or separately.

Results: The combined effect of data augmentation and model complexity results in up to 10% variation in prediction accuracy. Notably, when affine transformation are applied separately, the model achieves higher accuracy, regardless the chosen architecture. Across all strategies, the model accuracy exhibits a concave behavior as the number of convolutional layers increases, peaking at an intermediate value. The best model reaches excellent performance both on the internal and additional external testing set.

Discussions: Our work underscores the critical importance of adhering to rigorous experimental practices in the field of ML applied to healthcare. The results clearly demonstrate how data augmentation and model depth—often overlooked factors– can dramatically impact final performance if not thoroughly investigated. This highlights both the necessity of exploring neglected modeling aspects and the need to comprehensively report all modeling choices to ensure reproducibility and facilitate meaningful comparisons across studies.

KEYWORDS

deep learning, Alzheimer's disease, data augmentation, model depth, reproducibility

1 Introduction

Advanced Machine Learning (ML) techniques have proven to be highly effective in healthcare applications, such as cancer detection and prognosis (Cruz and Wishart, 2006; Sajda, 2006; Kourou et al., 2015; Shen et al., 2019; Chaunzwa et al., 2021), heart diseases prediction (Mohan et al., 2019; Palaniappan and Awang, 2008), and neurodegenerative diseases' diagnosis (Pereira et al., 2016; Montolío et al., 2021). However, it is still premature to assert that ML is ready to be employed as a standard in clinical practice. For instance, in Roberts et al. (2021), the authors reviewed thousands of papers on the use of ML to detect COVID-19 and found that none achieved the robustness and reproducibility required for medical use. This issue is not specific to ML methods for COVID-19 detection but involves the entire ML community (Ioannidis, 2005; Pineau et al., 2021), particularly the field of ML in healthcare (Stupple et al., 2019; Beam et al., 2020; Heil et al., 2021). To address this issue, Luo et al. (2016) asked 11 researchers with expertise in biomedical ML to produce a set of rules ensuring that ML models within clinical settings are sufficiently reported. These rules mainly relate to paper writing, providing a checklist for each article section. Although Luo et al. (2016) offers a useful tool for checking final manuscripts, it does not identify specific practices for developing ML methods in healthcare and is often very general when it comes to report ML model details (e.g., identifying if the study is retrospective/prospective and if the prediction task is regression/classification).

In our manuscript, we identify an essential set of practical guidelines, and we highlight the importance of fully adhering to them. To demonstrate this, we present a practical application of ML in healthcare by following these guidelines and demonstrating the impact of modeling choices on the final performance. Specifically, we focus on Deep Learning (DL) for Alzheimer's Disease (AD) diagnosis. AD is the most common type of dementia, impacting over 30 million individuals globally. It is characterized by (i) a pre-symptomatic stage where pathological molecular changes and neuronal dysfunctions occur at brain level, (ii) a prodromal stage identified as mild cognitive impairment (MCI) syndrome; (iii) an early-stage where cognitive symptoms of AD become more evident; (iv) a late stage with overt dementia. This progressive neurodegenerative disorder leads to cognitive and functional decline, impairing daily activities and eventually resulting in death. Hence, timely and accurate diagnosis of AD is crucial for effective treatments. Structural Magnetic Resonance Imaging (MRI) has proven to be a powerful tool for predicting AD due to its ability to visualize detailed brain structures and identify changes associated with the disease, such as hippocampal atrophy (Jack et al., 2000; Van De Pol et al., 2006), cortical thinning (Du et al., 2007), and brain volume loss (Pini et al., 2016).

In this study, we leverage low-resolution MRI scans and address the challenge of discriminating patients with AD from Cognitively Normal (CN) subjects using a 3D-Convolutional Neural Network (CNN) (LeCun et al., 1995). We combine different data augmentation strategies and CNN depths, creating a total of 15 DL models. We show that these modeling choices can lead to significant variations in prediction accuracy, up to 10%. The best model demonstrates excellent accuracy on the testing set and good properties of generalization to an external dataset. It is worth noting that the proposed approach can be readily extended to other modeling choices and healthcare applications.

The paper is structured as follows. The Materials and Methods section includes the guidelines for ML reliability and reproducibility, and introduces state-of-the-art studies in the AD field. Then, it details data handling and the experimental setup, including modeling challenges and choices made. The Results section evaluates the effect of the modeling choices, comparing augmentation strategies and architectures. The Discussion section relates findings to state-of-the-art studies and illustrates future perspectives.

2 Materials and methods

2.1 Guidelines

To begin, we summarize the general guidelines for reliable and reproducible ML pertaining to two key aspects: *data handling*, and *model design and assessment*.

Data handling (D)

1. Data collection/selection should align with the scientific problem at hand (e.g., utilizing cross-sectional data for diagnostic confirmation or longitudinal data for prognostic purposes), avoiding bias and information leakage (Saravanan et al., 2018).
2. Data quality should be assessed by identifying missing values and inconsistencies, and improved by applying appropriate imputation and cleaning methods (Lin and Tsai, 2020).
3. Data harmonization can be used to compensate for heterogeneous data from different acquisition techniques (Kourou et al., 2018).
4. Data augmentation can be employed as a solution for small sample size or unbalanced samples per class, a common case in the biomedical field.
5. The whole data handling process should be described in details in order to ensure reproducibility.

Model design and assessment (M)

1. The versioned code used for conducting the experiments should be publicly shared to ensure transparency and reproducibility.
2. Every decision in the design of the predictive model should be justified, with recognition of uncontrollable factors (Haibe-Kains et al., 2020).
3. Details about the samples used in the training/testing split should be disclosed to guarantee benchmarking.
4. A well-designed experiment should avoid assessing results on a non-representative testing set. To this aim, resampling strategies (Batista et al., 2004) such as k-fold cross-validation or boosting can be utilized to comprehensively assess the model's performance. Further, models based on random weights

Abbreviations: ML, Machine Learning; DL, Deep Learning; CNN, Convolutional Neural Network; CL, Convolutional Layers; AD, Alzheimer's Disease; MCI, Mild Cognitive Impairment; CN, Cognitive Normal; ADNI, Alzheimer's Disease NEuroimaging Initiative; MRI, Magnetic Resonance Imaging; D, Data handling; M, Model design and assessment.

initialization should be repeated for different trials in order to assess their stability.
5. The performance metrics should be chosen according to the specific scientific objectives of the study (Sokolova and Lapalme, 2009; Chicco and Jurman, 2020).
6. Testing the model on external datasets is ideal to evaluate its generalization properties (Basaia et al., 2019).

These guidelines are followed throughout the rest of the paper and referenced within the text whenever a rule is applied in the experiments.

2.2 State of the art

AD is a neurodegenerative disease and the most common form of dementia globally, characterized by progressive neurodegeneration, leading to cognitive and functional decline, impaired daily activities, and eventually, death (Wu et al., 2017; Dubois et al., 2016). Brain imaging, particularly MRI scans, plays a crucial role in diagnosing AD by providing detailed insights into the structural brain changes associated with the disease. In recent years, ML models have shown significant potential in utilizing imaging data to improve automated AD diagnosis (Yu et al., 2022) and predict AD-related brain abnormality (Zong et al., 2024). For instance, Zuo et al. (2024) use multiple brain image modalities with an adversarial learning strategy for AD progression prediction and to identify abnormal brain connections. Similarly, Pan et al. (2024) proposes a generative adversarial network with a decoupling module to detect abnormal neural circuits.

As reported in Arya et al. (2023), the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset (Mueller et al., 2005) is the most frequently employed dataset in AD studies based on ML and DL approaches. ADNI comprises heterogeneous datasets collected during different temporal phases (ADNI1, ADNI/GO, ADNI2, and ADNI3), each characterized by varying MRI acquisition protocols. ADNI1 includes longitudinal acquisitions on 1.5T and 3T scanners with T1- and T2-weighted sequences; ADNI-GO/ADNI2 contains imaging data acquired at 3T with similar T1-weighted parameters to ADNI1; ADNI3 exclusively utilizes MRI obtained from 3T scanners. Further, within a temporal phase, multiple acquisitions are done at different time steps (e.g., baseline, screening, follow up).

The heterogeneity of ADNI allowed for many experimental setups in the literature, with varying results depending on sample size [ranging from hundreds (Liu et al., 2014; Alinsaif and Lang, 2021; Long et al., 2017; Korolev et al., 2017) to thousands (Salehi et al., 2020; Basaia et al., 2019)], images resolution, or sequence type. However, this variability and the lack of a universally recognized benchmark have hindered fair comparisons of published models. Another consequence is that AD studies are more susceptible to information leakage. In Wen et al. (2020), the authors reviewed 32 studies using CNN models for AD diagnosis and found that about 50% of them reported biased results due to data leakage. These factors underscore the essential need for carefully selecting the dataset (D1), reporting details on data processing (D5, M3), taking into account the dataset size (D4, M3, M4) and choosing the model (M2) and the evaluation metrics accordingly (M5). In the rest of the section, we discuss state of the art (SOTA) studies on MRI-based AD classification using ADNI and describe their experimental approaches in relation to the criteria D and M. We emphasize that a systematic review is behind the purpose of this work, which has the scope of highlighting good and bad practices in ML for healthcare.

We considered the studies reported in a recent PRISMA-based review (Arya et al., 2023), selecting 8 articles that used solely MRI scans from ADNI dataset (Mehmood et al., 2021; Li and Yang, 2021; Pan et al., 2020; Alickovic et al., 2020; Korolev et al., 2017; Yue et al., 2019; Xiao et al., 2017; Tong et al., 2014). To increase the sample of DL-based articles, we further considered three SOTA articles (Salehi et al., 2020; Basaia et al., 2019; Ghaffari et al., 2022), for a total of 11 articles. We found that none of them fully adhered to the guidelines listed in the previous section. In particular:

- D1: 73% of studies did not report the ADNI phase, and 91% did not specify the time step (e.g., baseline, follow-up). This information is crucial to ensure that baseline and follow-up data are not mixed, thereby preventing data leakage. Additionally, 27% of studies did not provide information about MRI resolution (i.e., 1.5T or 3T).
- D4: Data augmentation is applied in only 4 papers (Mehmood et al., 2021; Pan et al., 2020; Basaia et al., 2019; Ghaffari et al., 2022). These papers lack important details, such as transformation parameters and the size of the final training set.
- M1: Only the authors in Korolev et al. (2017) provided the code used for data processing and modeling.
- M2: Only 27% of the works considered different model architectures. Additionally, none of the DL approaches explored model depth as a hyperparameter.
- M3: Three articles split the dataset into training/testing following previous work, whereas the remaining ones did not detail the samples in the splits, preventing benchmarking.
- M4: Resampling strategies were not used in 45% of experiments. Furthermore, no DL-based methods tested model robustness to random weight initialization.
- M5: 91% of studies adopted multiple evaluation metrics. However, standard deviation for resampling strategies was reported in only three papers.
- M6: Generalization across datasets was tested and reported in only two articles.

Note that D2 and D3 are not evaluated here as data quality is ensured by ADNI experts and none of the considered studies rely on different acquisition techniques.

The literature review reveals that none of the considered SOTA studies are fully reproducible due to the absence of available validated code, insufficient details about data processing and augmentation, and lack of information about dataset splits and experimental specifics. Furthermore, the reliability of these works is sometimes limited by unrepresentative testing sets and the lack of evaluation on external datasets. It is also interesting to note that the number of employed samples varies from 170 to 1,662, with a median of 433, a mean of 653, and a standard deviation of 495. This, along with the variability in MRI resolution, makes model comparisons unfeasible. Finally, we noted that model depth and data augmentation strategy (in terms of the number of augmented samples and types of transformations) were completely neglected factors. This led us to investigate whether and to what extent these two modeling choices impact the classification task.

TABLE 1 ADNI1 demographic description.

1.5T	CN	AD
Subjects	307	243
Age	75.2 ± 7.6	75.9 ± 5.0
Sex (M/F)	159/148	130/113
3T	**CN**	**AD**
Subjects	47	33
Age	75.1 ± 3.9	74.0 ± 8.1
Sex (M/F)	18/29	11/22

1.5 and 3T datasets.

2.3 Data

For our experiments, we adopted the ADNI dataset (Mueller et al., 2005) considering T1-weighted 1.5T MRI scans from the ADNI1 data collected during screening, which is the baseline exam. This includes 550 MRI exams from 307 CN subjects and 243 AD patients. Additionally, we used an ADNI1 subset of 80 3T MRI exams as an external testing set, to evaluate the best model in a *domain shift* setting (Buchanan et al., 2021). Table 1 reports demographic details about the two datasets (D1). We recall that MRI exams are three-dimensional data describing the structure of the brain. Figure 1 displays a 2D projection of brain images captured from a CN subject (first row) and an AD patient (second row) on the *sagittal*, *coronal*, and *axial* planes. All data were preprocessed by ADNI experts, ensuring data quality and harmonization (D2, D3; more information in Supplementary Section 1).

2.3.1 Data augmentation (D4)

Data augmentation is a common procedure that simultaneously addresses data scarcity and creates a model invariant to a given set of transformations (Shorten and Khoshgoftaar, 2019). Different augmentation strategies can result in varied training sets, affecting model performance and computational cost. In this study, the original set is augmented by applying, separately or simultaneously, *zoom*, *shift*, and *rotation* transformations, as shown in Figure 2 (see Supplementary Section 1.3 for details on the transformation parameters). To study the effect of different transformations and sample sizes on model performance, we compared the following three data augmentation strategies:

- **Strategy (A)**. To each image, we simultaneously apply all the transformations (i.e., a zoom by a random factor, a random shift, and a rotation by a random angle). The size of the augmented data will match the number of training samples N.
- **Strategy (B)**. To each image, we separately apply each transformation, generating three different distorted images. The size of the augmented data will be three times the number of training samples, $3N$.
- **Strategy (C)**. To each image, we simultaneously apply all the transformations, as in strategy A. We repeat the process three times so that the number of augmented samples matches the one of strategy B ($3N$).

Therefore, strategies **(A)** and **(C)** rely on the same procedure, while strategies **(B)** and **(C)** generate the same number of samples. Although other augmentation techniques (e.g., color transformation, adding noise, and random erasing) may be beneficial, a comprehensive study of data augmentation is beyond the scope of this work. Instead, our goal is to investigate whether and how slight variations in data augmentation choices, often underestimated, impact model performance. In order to avoid data leakage (Wen et al., 2020), data augmentation is performed only on the training set after dataset split, leaving validation and testing sets at the original sample size.

2.3.2 Data processing (D5)

As already noted, ADNI images were collected with different protocols and scanning systems, hence they are very heterogeneous in size, see Table 2. To enable the use of ML methods, it is necessary to select a common volume size. This choice, often left unexplained in literature, defines fundamental characteristics of the pipeline, such as the amount of information contained in the image and the input space dimension, on which model choice and computational burden depend.

In our experiments, images are downsized to 96×96×73. The principle guiding this choice derives from computational issues. We first reduced the image dimension, rescaling the image by 50% along all dimensions, and we then resized images to match the smallest one. An alternative strategy may be zero-padding to match the biggest image, but this would increase memory requirements. Finally, intensity normalization was applied omitting the zero intensity voxels from the calculation of the mean. This procedure allows having homogeneous data with a fixed size. Note that we do not select any Region Of Interest (ROI) (Long et al., 2017) within the images. Although this setup challenges the classification task, it eliminates the typically laborious and time-consuming feature engineering process.

2.4 Experimental setup

2.4.1 Guide to the model choice (M2)

Choosing the optimal DL model is not straightforward, as the vast numbers of network and training parameters makes a "brute-force" model selection approach unfeasible. Here, we illustrate the model choices made a priori based on the issues posed by the examined task.

2.4.1.1 Type of data

Working with 3D images presents computational and memory challenges. As a solution, several studies in the literature adopt three 2D projections of the MRI. Nevertheless, this approach requires three separate models, leading to increased overall wall-clock time. Moreover, extracting features from the 2D projections may result in the loss of crucial volumetric information and a simplified representation of the studied phenomenon. In this work, we adopted a 3D CNN that directly extracts volumetric features.

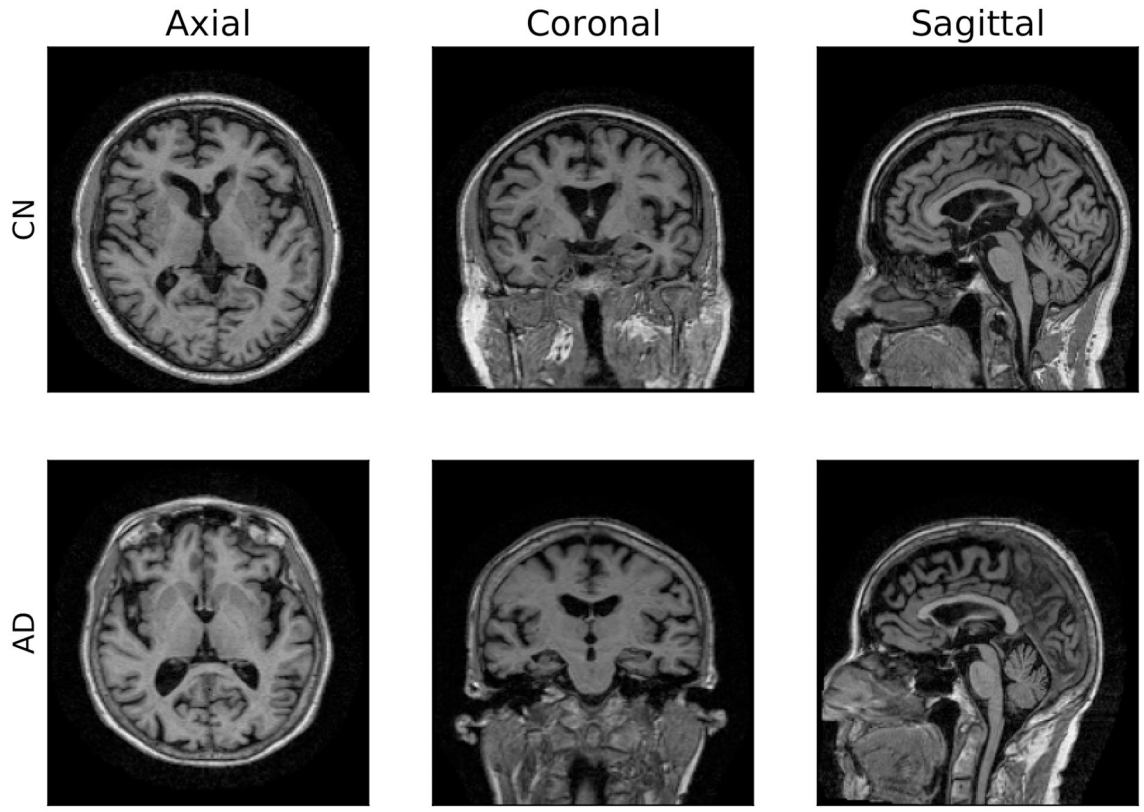


FIGURE 1
2D visualization of 3D MRI scans. Axial, coronal and sagittal planes of two brain images from ADNI dataset.

2.4.1.2 Limited amount of data

To overcome the limited dataset size, we implemented the following strategies aimed at controlling model complexity and preventing overfitting: data augmentation; adding an ℓ_2 penalty; and limiting the number of filters per layer. The latter method resulted in a substantial parameter reduction across the network. For instance, in a 2-layer CNN with $3\times3\times3$ filters, reducing the number of filters to 32 to 8 in the first layer and from 64 to 16 in the second layer (25% of the initial values) leads to a considerable reduction of 93% in the number of learnable parameters (from 56,256 to 3,696).

2.4.1.3 Memory capacity

3D models usually require a huge amount of memory capacity, that depends both on the input dimension and the model size. To reduce the required memory: i) we re-scaled the images to halve the data dimension; ii) we used stochastic gradient descent with a batch size that balances the memory cost while retaining a representative subset; iii) we balanced the number of filters and the batch size to reduce the computational burden of the activation layer.

2.4.2 Model details

We report experiments on the CN/AD binary classification. A preliminary analysis, performed on 1.5T MRI data with a standard training/validation/test split (75%/15%/10%), denoted a very high variance due to the limited sample size of the testing set. For this reason, to guarantee a correct assessment of model performance and stability, we set up a stratified-K-fold cross-validation loop. We set K = 7, from Fold 0 to Fold 6 (training/validation/test, with a proportion of 70%/15%/15%), that ensures having enough data for the learning phase (M4). All folds were fully balanced, except for Fold 6 which had an unbalanced ratio between AD and CN samples as the total amount of samples per class do not match exactly. We further tested our model on the external dataset of 3T MRI scans (M6). Note that this task is particularly challenging because: i) the evaluation is subject to the domain shift problem, and ii) the training MRI scans have half the resolution of the external MRI exams.

We adopted as baseline network an architecture with 4 Convolutional Layers (CL) followed by a fully-connected layer, as depicted in Figure 3. We will refer to this architecture as **4 CL** model. To investigate the optimal CNN depth, we inserted additional convolutional layers without pooling operations so that the number of layers is the only factor impacting in the model. Specifically, we added 2, 4, 6 and 8 convolutional layers in correspondence to the arrows of Figure 3. We refer to these models as **6 CL**, **8 CL**, **10 CL**, and **12 CL**. For instance, in the **10 CL** architecture 6 convolutional layers are added to the **4 CL** baseline: two layers are inserted in correspondence of the first and second arrows, and one layer in correspondence of the third and fourth arrows. Additional details on network and training parameters can be found in the Supplementary Section 2. In order to test model stability to initial random weights, each model was run 10 times

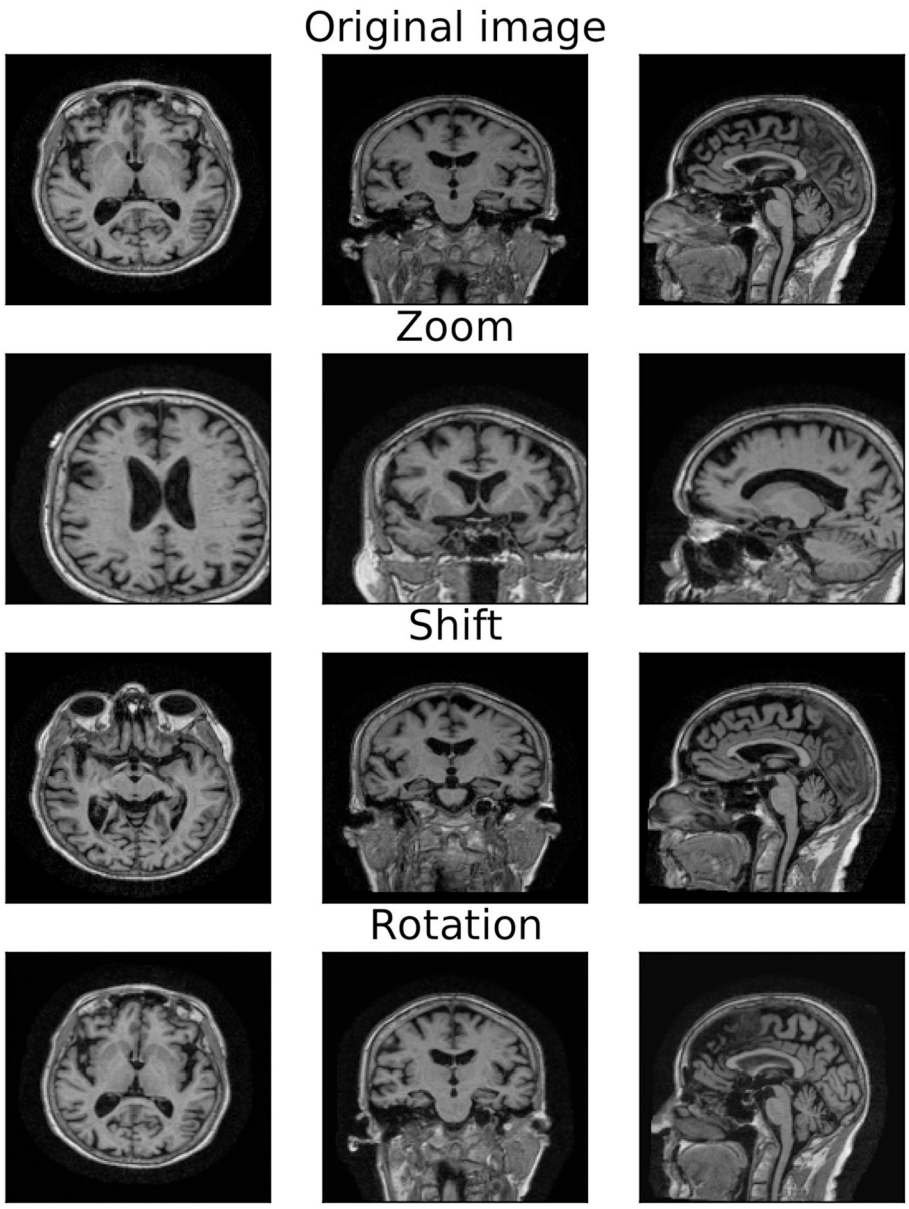


FIGURE 2
Original and transformed MRI image. 2D projections of the original MRI image **(first row)** and the augmented image obtained by applying *zoom* **(second row)**, *shift* **(third row)**, and *rotation* **(last row)** transformations.

(M4). Model selection was performed based on accuracy. The best one is further analyzed based on Confusion Matrix, Precision, Recall, F1-score, AUC and AUCPRC (M5).

All the experiments were conducted using Python version 3.8 and PyTorch 1.12.1, running on a Tesla K40c GPU. Samples identifiers and the Python code necessary to reproduce the experiments are available on GitHub (M1, M3).

3 Results

In the following, we compare 15 models obtained by combining different augmentation strategies with varying network depths, then we illustrate in detail the results of the best model. Results based on not-augmented data are not reported, as they were substantially worse than the ones obtained by using augmentation.

3.1 Architecture and augmentation choice

We assessed the optimal architecture and augmentation strategy based on the accuracy on the validation set, which is shown in Figure 4. To verify the impact of these factors on the classification task, we performed a statistical analysis of the results obtained by the different models. Initially, we used the Shapiro-Wilk test (Shapiro and Wilk, 1965) to assess the normality of our data, which revealed that the data were not normally distributed. Consequently, we adopted a non-parametric approach

to determine significant differences in models' performance. Specifically, we applied the Kruskal-Wallis test (Kruskal and Wallis, 1952) to compare performance across the 15 models. This analysis yielded a statistically significant difference (p-value = 7.45e-07), indicating that the classification task varies significantly among models with different augmentation strategies and network depth.

3.1.1 Data augmentation

Strategy (A) (in yellow) considerably underperforms Strategy (B) (in green), regardless of the CNN architecture used. This can be attributed to the lower number of samples in the augmented data. Surprisingly, Strategies (A) and (C) (in fuchsia) achieve very similar accuracy for a higher number of layers. Finally, although Strategies (B) and (C) generate the same amount of data, Strategy (B) outperforms Strategy (C) across all network depths. To validate these findings, we repeated the Kruskal-Wallis test comparing models using strategy (A), (B), and (C), for each architecture. All tests resulted in p-values less than 0.05, confirming significant differences in performance across different augmentation strategies. Furthermore, as Strategy (B) resulted in the most effective data augmentation approach, we conducted additional statistical analysis on it. Specifically, we used the Conover-Iman test (Conover and Iman, 1979) for pairwise comparison between models based on strategy (B) and those employing different data augmentation strategies. Results revealed a significant difference between strategy (B) and strategy (A) for all network depth, and between strategy (B) and strategy (C) for the **8 CL**, **10 CL**, and **12 CL** architectures. These outcomes underscore the superiority of strategy (B) across all tested architectures, and demonstrate that applying affine transformations separately is more effective than applying them simultaneously.

TABLE 2 1.5 T1-weighted MRI scans.

MRI size	CN	AD	Total
256 × 256 × 184	8	8	16
256 × 256 × 170	40	34	74
256 × 256 × 160	4	0	4
256 × 256 × 166	97	82	179
256 × 256 × 162	0	1	1
192 × 192 × 160	117	86	203
256 × 256 × 146	1	0	1
256 × 256 × 161	2	0	2
256 × 256 × 180	38	32	70

Number of CN and AD MRI scans grouped by size.

3.1.2 Network depth

The accuracy curves for all augmentation methods show a similar pattern: the best results are obtained for intermediate amounts of layers, while accuracy decreases for higher numbers of convolutional layers. The same behavior can be observed in Figure 5 where we report for each cross-validation fold the distribution of accuracy in the 10 trials. Using the Kruskal-Wallis test, we found that these differences across architectures were significant when using strategies (A) and (B).

The **8 CL** model with strategy (B) emerges as the best-performing combination, exhibiting greater stability within

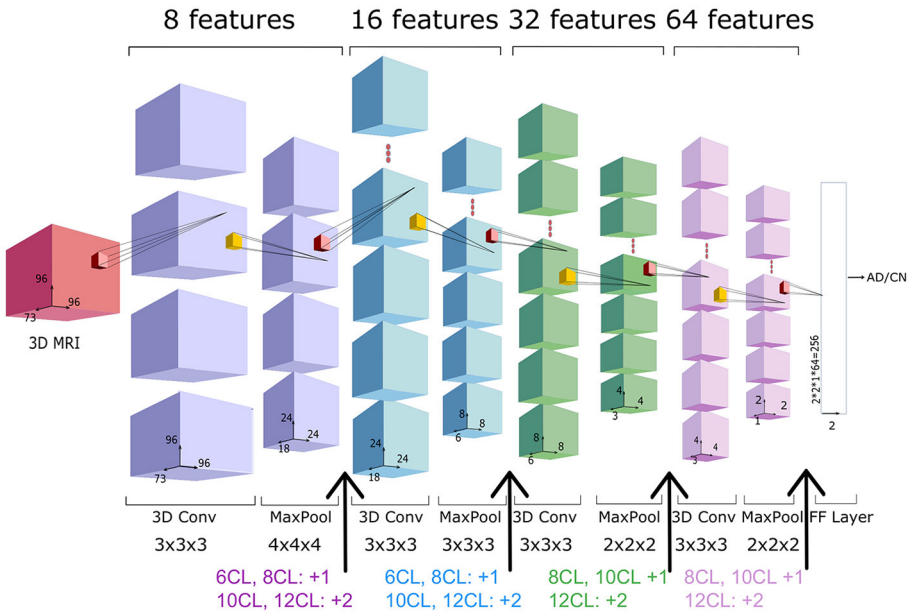


FIGURE 3
3D-CNN architecture. Architecture of the **4 CL** baseline network, composed by four blocks of a convolutional and pooling layers, followed by a fully connected (FC) layer. The total number of features ($8 * i$) in the i-th convolutional layer is reported above each layer, whereas the filter dimension is reported below. In the experiments, we consider other four extended versions of the baseline architecture (6CL, 8CL, 10CL, 12CL) duplicating once or twice the convolutional layer preceding the arrows.

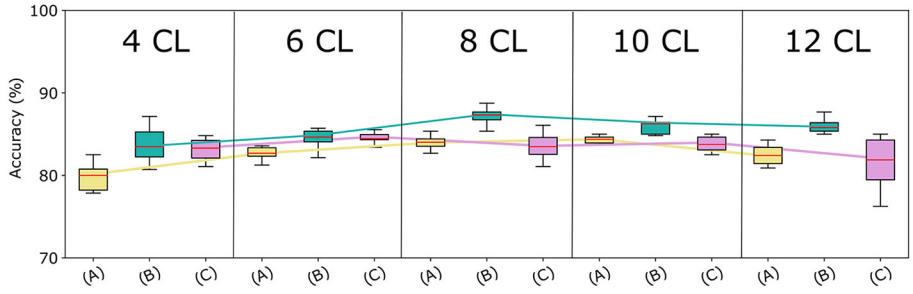


FIGURE 4
Models accuracy at varying of architecture depth and augmentation strategies. Comparison among the proposed CNN-based architectures with the three augmentation strategies, in terms of median accuracy on the validation set. The *y*-axis reports the model accuracy distribution on the 10 trials (%) and the *x*-axis presents varying augmentation strategies (A), (B), and (C) in 5 blocks—one for each CNN architecture.

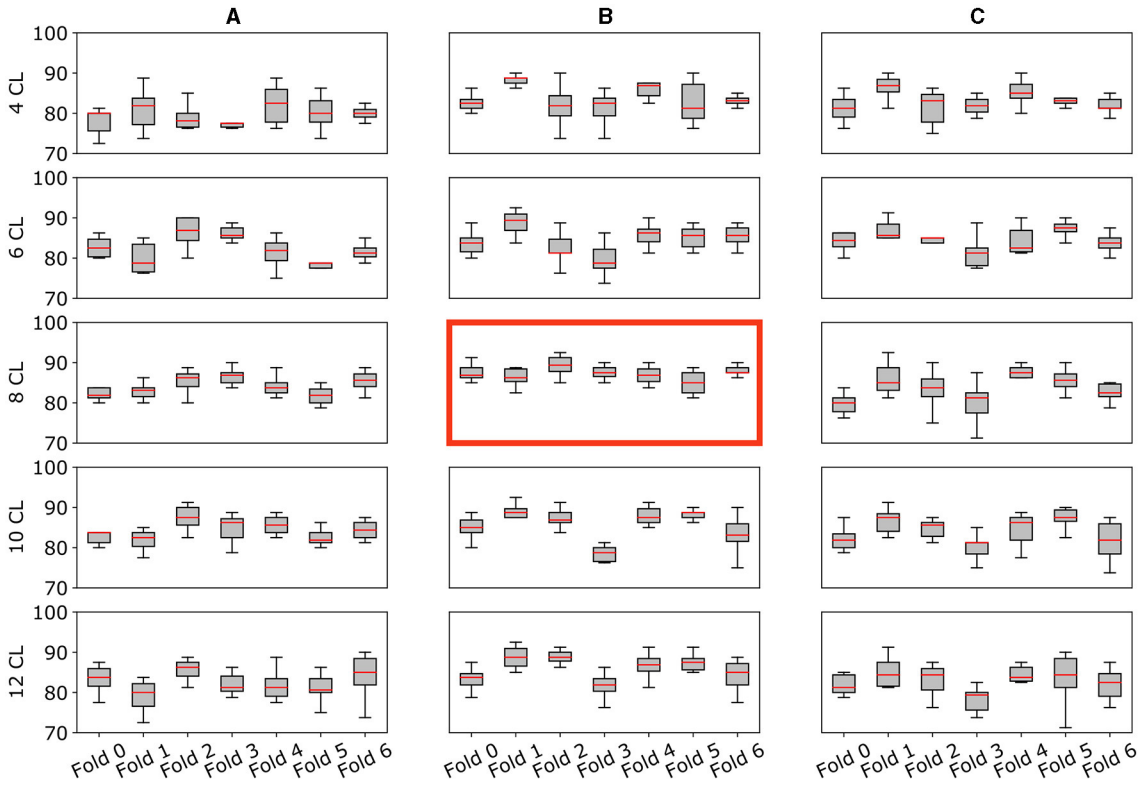


FIGURE 5
Model's performance and stability across folds. Multiple plots for the comparison of the validation accuracy for all architectures **(A–C)** and augmentation strategies (4CL, 6CL, 8CL, 10CL, 12CL). Each subplot reports the model accuracy on all 7-fold splits. Specifically, the y-axis reports the accuracy distribution on the 10 trials (%) for each fold (x-axis). The best model [8 CL, **(B)**] is highlighted with a red border.

and across folds compared to the other combinations. Further details and specific results of the statistical analysis are available in the Supplementary material.

3.2 Best model performance and insight

The combination of a CNN with 8 convolutional layers and the (B) augmentation strategy [**8 CL**, (B)] turned out to be the best model, reaching an accuracy of 87.21 ± 0.88% on the validation set and 81.95 ± 1.26% on the testing set.

A complete evaluation of this model is reported in Figure 6: left panel reports mean and standard deviation for Precision, Recall, F1-score, AUC and AUCPRC of CN and AD classes over the 7 folds; right panel shows the Confusion matrix obtained by counting True Positive, True Negative, False Positive, and False Negative scores over the 7 folds. Figure 7 gives an insight on the layers behavior and how they are learning the optimal model. The Left Panel displays the learned filters of every convolutional layer for one AD patient on the three considered median planes, i.e., *sagittal*, *coronal* and *axial*. It is clear that the filters capture more abstract features at increasing depth values. Panel (b) presents, for each convolutional

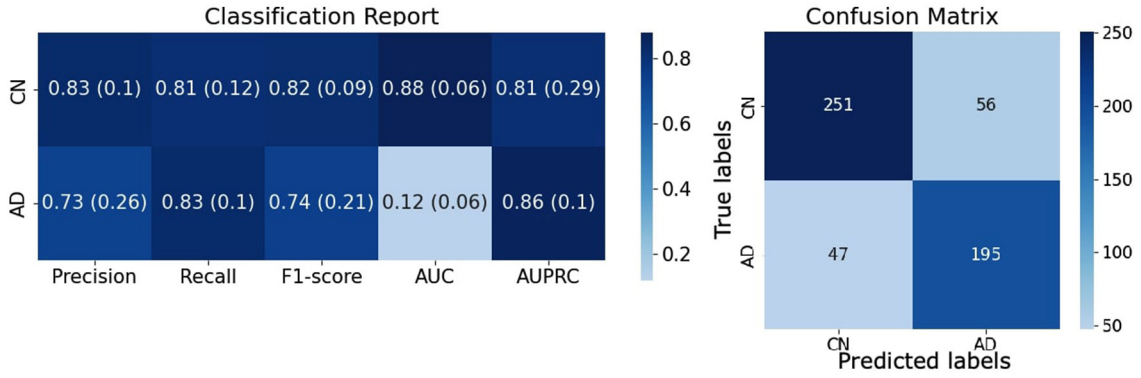


FIGURE 6
Evaluation of the [**8 CL**, (B)] model on the testing set. **(Left)** Complete evaluation of the model on CN and AD classes averaged over the 7 folds. **(Right)** Confusion matrix of the classification results counted over the 7 folds.

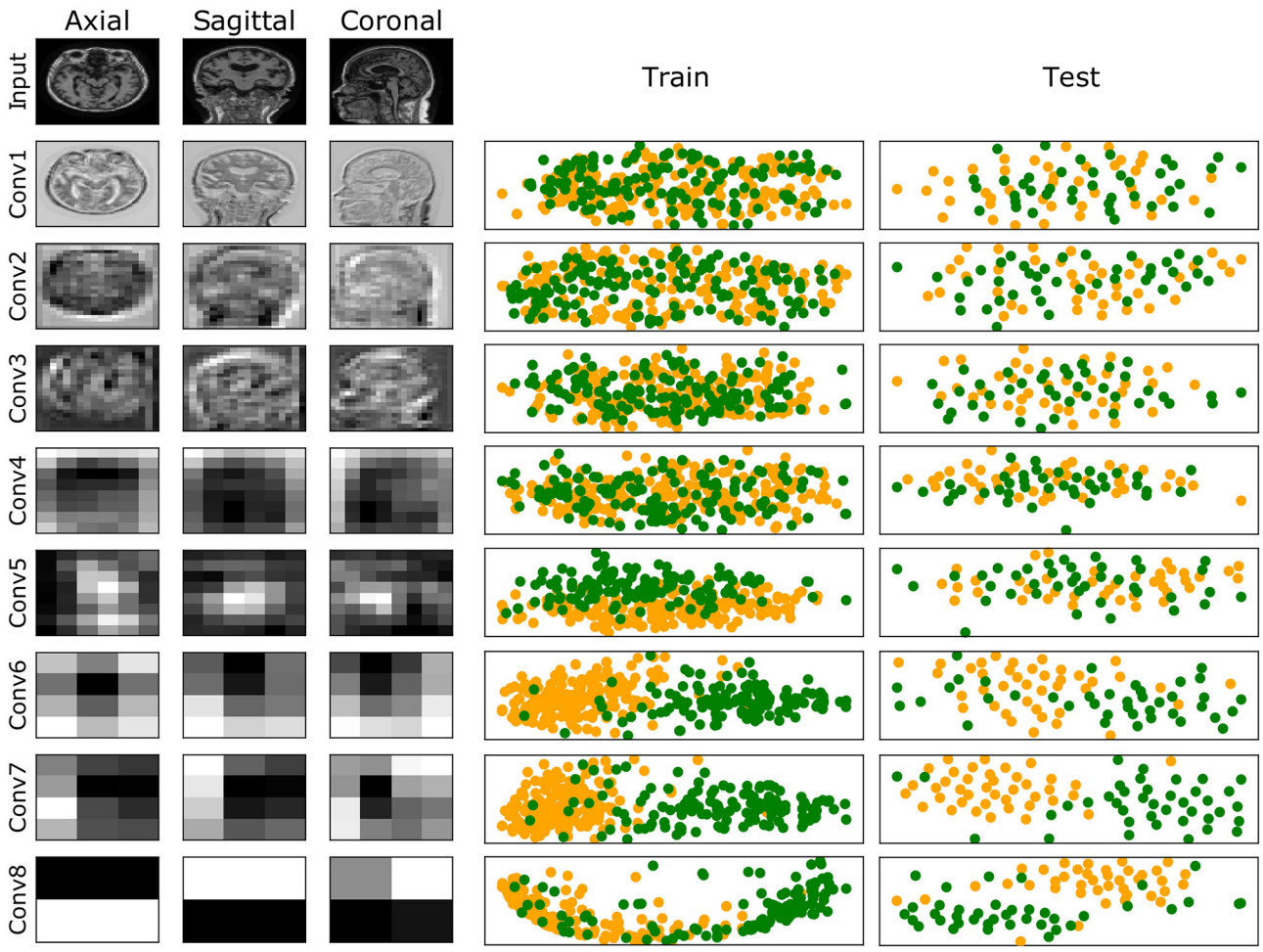


FIGURE 7
(Left) Illustration of the learned filters by the best model for one of the AD samples. Columns show filters for the three median planes, and rows show the filters for the input (raw data) and the convolutional layers at increasing depth. **(Right)** Training and test embeddings for each convolutional layer of the [**8 CL**, (B)] model projected by t-SNE. For increasing depth, AD (green) and CN (yellow) samples are better clustered.

layer, the layer outputs (*embeddings*) of training and test samples projected on a two-dimensional plane through t-distributed Stochastic Neighbor Embedding (t-SNE) (Van der Maaten and Hinton, 2008). Both projections show that the embeddings are more evidently clustered as the number of layers increases.

To further understand the properties and limits of the (**8 CL**, (B)) model, we assessed the effect of dropout, finding that it does not improve its performance (details in Supplementary Section 3.2). Also, we tested the model on an external dataset of 3T MRI scans, obtaining an accuracy of 71% and an AUC of 0.76 (a complete evaluation can be found in Supplementary Section 3.3).

4 Discussion

In this paper, we summarized a list of 5 items concerning *data handling* (D) and 6 items on *model design and assessment* (M), outlining the criteria that should be adhered to in order to ensure reliability, robustness, and reproducibility in ML for healthcare. Based on these criteria, we constructed an experimental pipeline for MRI-based binary classification of AD vs. CN subjects. Specifically, the experiments were conducted on a pre-processed subset of the ADNI dataset, consisting of 1.5T MRI scans collected during the screening ADNI1 phase (D1). This subset, previously pre-processed by ADNI experts, ensures high data quality (D2) and harmonization (D3). Although the dataset is balanced, its size is limited. To address potential overfitting and ensure reliable results, data augmentation (D4), model complexity reduction (M2), and resampling (M4) strategies were employed. All these aspects are thoroughly discussed (D5). The list of selected samples was made publicly available to enable benchmarking in further studies (M3), along with the Python code (M1).

Additionally, we thoroughly investigated the combined impact of data augmentation strategies (by varying the number of augmented data and the application of transformations) and architecture depth (M2), resulting in a total of 15 models. As reported in Section 2.2, these factors are often neglected in the literature, which typically aims to generate the largest possible number of augmented data and use state-of-the-art architectures (even when very large). Our findings demonstrate that improper settings for these experimental aspects can drastically hamper model performance, reducing accuracy by up to 10 points. Results showed that, independently of the adopted architecture, Strategy (B) always outperformed the others. As strategies (B) and (C) leverage the same amount of training samples, these results suggest that applying the affine transformations separately may help the model build invariance to each of them. Interestingly, strategies (A) and (C) show similar performances for intermediate-to-large models, even though strategy (A) relies on only one-third of the samples generated by strategy (C). We recall that Strategy (A) adopts the same combination of transformations as Strategy (B). This may indicate that the way transformations are combined and applied to the original data has a greater impact than the augmented dataset size itself. Future work will extend this investigation to other data augmentation strategies, including different types of transformation (e.g., color space transformations, Kernel filters, random erasing).

For all augmentation approaches, we found that the curve of the model accuracy at increasing depths tends to be a concave function, reaching the maximum for an intermediate depth value. Although the widespread notion for which deeper neural networks better generalize in a general framework, this result is in line with other studies (Zhang et al., 2021; Vento and Fanfarillo, 2019) in which authors showed that smaller models perform better when only a limited amount of data is available, as they are less subject to overfitting. Although we did not test them, this observation may extend to other SOTA architectures. Indeed, our **8 CL** CNN has 220k trainable parameters, while SOTA architectures are typically much larger. For example, ResNet18, ResNet50, and ResNet101 (He et al., 2016) consist of 11.7M, 25.6M, and 44.5M parameters, respectively. The smallest Vision Transformer model (ViT-Base) (Dosovitskiy et al., 2020) includes 86M parameters. EfficientNet-B1 (Tan and Le, 2019) and MobileNetV2 (Sandler et al., 2018), considered among the smallest SOTA architectures, have 7.8M and 3.5M parameters, respectively. Using larger SOTA models may be more effective when pre-trained to leverage transfer learning. However, it is important to note that the vast majority of pre-trained models have been trained on natural 2D images, and they are not immediately usable in the context of medical 3D scans. Future work will delve into these aspects.

The best model we identified is the combination of a CNN with 8 convolutional layers and the (B) augmentation strategy [**8 CL**, (B)]. The model accuracy in validation and testing is $87.21 \pm 0.88\%$ and $81.95 \pm 1.26\%$, respectively, which is 4.2% increase in accuracy with respect to [**4 CL**, (B)] model. Also, Figure 5 shows how [**8 CL**, (B)] is more stable than all other models with respect to both cross-validation folds and training trials. These results appear in line with current SOTA studies relying on similar datasets. For instance, Pan et al. (2020) reach 84% of accuracy by using 499 1.5T MRI scans, and Xiao et al. (2017) obtain 85.7% using a dataset of 654 1.5T MRI images. Similarly to our work, Korolev et al. (2017) train a 3D-CNN model on 231 samples, showing 79% of accuracy. Nonetheless, we argue that a true comparison is not completely feasible as other works employ different datasets and data types, the number of samples varies both in training and testing sets, experimental designs are very heterogeneous and, most importantly, performance is always assessed on one trial, without any variability estimation. As an additional evaluation, we tested the best model in a *domain shift* context (M6), i.e., on 3T MRI data, reaching 71% of accuracy. We remark that this is a very challenging task as the image resolution deeply differs from the one in the training set.

To the best of our knowledge, this is the first work in the AD domain to delve into these modeling aspects and quantify their impact on performance estimation. Future work will extend this analysis to other architectures, different data augmentation transformations, and to a multi-class classification setting that includes MCI subjects.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://adni.loni.usc.edu/data-samples/access-data/.

Ethics statement

The studies involving humans were approved by Albany Medical Center Committee on Research Involving Human Subjects

Institutional Review Board, Boston University Medical Campus and Boston Medical Center Institutional Review Board, Butler Hospital Institutional Review Board, Cleveland Clinic Institutional Review Board, Columbia University Medical Center Institutional Review Board, Duke University Health System Institutional Review Board, Emory Institutional Review Board, Georgetown University Institutional Review Board, Health Sciences Institutional Review Board, Houston Methodist Institutional Review Board, Howard University Office of Regulatory Research Compliance, Icahn School of Medicine at Mount Sinai Program for the Protection of Human Subjects, Indiana University Institutional Review Board, Institutional Review Board of Baylor College of Medicine, Jewish General Hospital Research Ethics Board, Johns Hopkins Medicine Institutional Review Board, Lifespan - Rhode Island Hospital Institutional Review Board, Mayo Clinic Institutional Review Board, Mount Sinai Medical Center Institutional Review Board, Nathan Kline Institute for Psychiatric Research and Rockland Psychiatric Center Institutional Review Board, New York University Langone Medical Center School of Medicine Institutional Review Board, Northwestern University Institutional Review Board, Oregon Health and Science University Institutional Review Board, Partners Human Research Committee Research Ethics, Board Sunnybrook Health Sciences Centre, Roper St. Francis Healthcare Institutional Review Board, Rush University Medical Center Institutional Review Board, St. Joseph's Phoenix Institutional Review Board, Stanford Institutional Review Board, The Ohio State University Institutional Review Board, University Hospitals Cleveland Medical Center Institutional Review Board, University of Alabama Office of the IRB, University of British Columbia Research Ethics Board, University of California Davis Institutional Review Board Administration, University of California Los Angeles Office of the Human Research Protection Program, University of California San Diego Human Research Protections Program, University of California San Francisco Human Research Protection Program, University of Iowa Institutional Review Board, University of Kansas Medical Center Human Subjects Committee, University of Kentucky Medical Institutional Review Board, University of Michigan Medical School Institutional Review Board, University of Pennsylvania Institutional Review Board, University of Pittsburgh Institutional Review Board, University of Rochester Research Subjects Review Board, University of South Florida Institutional Review Board, University of Southern, California Institutional Review Board, UT Southwestern Institution Review Board, VA Long Beach Healthcare System Institutional Review Board, Vanderbilt University Medical Center Institutional Review Board, Wake Forest School of Medicine Institutional Review Board, Washington University School of Medicine Institutional Review Board, Western Institutional Review Board, Western University Health Sciences Research Ethics Board, and Yale University Institutional Review Board. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

RT: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis. AV: Writing – review & editing, Writing – original draft. AB: Writing – review & editing, Writing – original draft, Supervision.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. Rosanna Turrisi was supported by a research fellowship funded by the DECIPHER-ASL–Bando PRIN 2017 grant (2017SNW5MB–Ministry of University and Research, Italy).

Acknowledgments

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: ADNI Acknowledgment List.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling editor NM declared a shared parent affiliation with the authors at the time of review.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fncom.2024.1360095/full#supplementary-material

References

Alickovic, E., Subasi, A., and Initiative, A. D. N. (2020). "Automatic detection of alzheimer disease based on histogram and random forest," in *CMBEBIH 2019: Proceedings of the International Conference on Medical and Biological Engineering, 16-18 May 2019, Banja Luka, Bosnia and Herzegovina* (Cham: Springer), 91–96.

Alinsaif, S., and Lang, J. (2021). 3d shearlet-based descriptors combined with deep features for the classification of alzheimer's disease based on MRI data. *Comput. Biol. Med*. 138:104879. doi: 10.1016/j.compbiomed.2021.104879

Arya, A. D., Verma, S. S., Chakarabarti, P., Chakrabarti, T., Elngar, A. A., Kamali, A.-M., et al. (2023). A systematic review on machine learning and deep learning techniques in the effective diagnosis of alzheimer's disease. *Brain Informat*. 10:17. doi: 10.1186/s40708-023-00195-7

Basaia, S., Agosta, F., Wagner, L., Canu, E., Magnani, G., Santangelo, R., et al. (2019). Automated classification of alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks. *NeuroImage: Clini*. 21:101645. doi: 10.1016/j.nicl.2018.101645

Batista, G. E., Prati, R. C., and Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorat. Newslett*. 6, 20–29. doi: 10.1145/1007730.1007735

Beam, A. L., Manrai, A. K., and Ghassemi, M. (2020). Challenges to the reproducibility of machine learning models in health care. *JAMA* 323, 305–306. doi: 10.1001/jama.2019.20866

Buchanan, C. R., Mu noz Maniega, S., Valdés Hernández, M. C., Ballerini, L., Barclay, G., Taylor, A. M., et al. (2021). Comparison of structural mri brain measures between 1.5 and 3 t: Data from the lothian birth cohort 1936. *Hum. Brain Mapp*. 42, 3905–3921. doi: 10.1002/hbm.25473

Chaunzwa, T. L., Hosny, A., Xu, Y., Shafer, A., Diao, N., Lanuti, M., et al. (2021). Deep learning classification of lung cancer histology using ct images. *Sci. Rep*. 11, 1–12. doi: 10.1038/s41598-021-84630-x

Chicco, D., and Jurman, G. (2020). The advantages of the matthews correlation coefficient (MCC) over f1 score and accuracy in binary classification evaluation. *BMC Genom*. 21, 1–13. doi: 10.1186/s12864-019-6413-7

Conover, W. J., and Iman, R. L. (1979). *Multiple-Comparisons Procedures*. Los Alamos, NM: Los Alamos National Lab. (LANL).

Cruz, J. A., and Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. *Cancer Inform*. 2:117693510600200030. doi: 10.1177/117693510600200030

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: transformers for image recognition at scale. *arXiv [Preprint]*. arXiv:2010.11929. doi: 10.48550/arXiv.2010.11929

Du, A.-T., Schuff, N., Kramer, J. H., Rosen, H. J., Gorno-Tempini, M. L., Rankin, K., et al. (2007). Different regional patterns of cortical thinning in Alzheimer's disease and frontotemporal dementia. *Brain* 130, 1159–1166. doi: 10.1093/brain/awm016

Dubois, B., Hampel, H., Feldman, H. H., Scheltens, P., Aisen, P., Andrieu, S., et al. (2016). Preclinical Alzheimer's disease: definition, natural history, and diagnostic criteria. *Alzheimer's & Dement*. 12, 292–323. doi: 10.1016/j.jalz.2016.02.002

Ghaffari, H., Tavakoli, H., and Pirzad Jahromi, G. (2022). Deep transfer learning-based fully automated detection and classification of Alzheimer's disease on brain mri. *Br. J. Radiol*. 95:20211253. doi: 10.1259/bjr.20211253

Haibe-Kains, B., Adam, G. A., Hosny, A., Khodakarami, F., Massive Analysis Quality Control (MAQC) Society Board of Directors, Waldron, L., et al. (2020). Transparency and reproducibility in artificial intelligence. *Nature* 586, E14–E16. doi: 10.1038/s41586-020-2766-y

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 770–778. doi: 10.1109/CVPR.2016.90

Heil, B. J., Hoffman, M. M., Markowetz, F., Lee, S.-I., Greene, C. S., and Hicks, S. C. (2021). Reproducibility standards for machine learning in the life sciences. *Nat. Methods* 18, 1132–1135. doi: 10.1038/s41592-021-01256-7

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Med*. 2:e124. doi: 10.1371/journal.pmed.0020124

Jack Jr, C. R., Petersen, R. C., Xu, Y., O'brien, P., Smith, G. E., Ivnik, R. J., et al. (2000). Rates of hippocampal atrophy correlate with change in clinical status in aging and ad. *Neurology* 55, 484–490. doi: 10.1212/WNL.55.4.484

Korolev, S., Safiullin, A., Belyaev, M., and Dodonova, Y. (2017). "Residual and plain convolutional neural networks for 3d brain MRI classification," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)* (Melbourne, VIC: IEEE), 835–838.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., and Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J*. 13, 8–17. doi: 10.1016/j.csbj.2014.11.005

Kourou, K. D., Pezoulas, V. C., Georga, E. I., Exarchos, T. P., Tsanakas, P., Tsiknakis, M., et al. (2018). Cohort harmonization and integrative analysis from a biomedical engineering perspective. *IEEE Rev. Biomed. Eng*. 12, 303–318. doi: 10.1109/RBME.2018.2855055

Kruskal, W. H., and Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *J. Am. Stat. Assoc*. 47, 583–621. doi: 10.1080/01621459.1952.10483441

LeCun, Y., and Bengio, Y. (1995). "Convolutional networks for images, speech, and time series," in *The Handbook of Brain Theory and Neural Networks*, ed. M. A. Arbib (MIT press), 3361.

Li, Q., and Yang, M. Q. (2021). Comparison of machine learning approaches for enhancing Alzheimer's disease classification. *PeerJ* 9:e10549. doi: 10.7717/peerj.10549

Lin, W.-C., and Tsai, C.-F. (2020). Missing value imputation: a review and analysis of the literature (2006-2017). *Artif. Intellig. Rev*. 53, 1487–1509. doi: 10.1007/s10462-019-09709-4

Liu, S., Liu, S., Cai, W., Pujol, S., Kikinis, R., and Feng, D. (2014). "Early diagnosis of alzheimer's disease with deep learning," in *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)* (Beijing: IEEE), 1015–1018.

Long, X., Chen, L., Jiang, C., Zhang, L., and Initiative, A. D. N. (2017). Prediction and classification of alzheimer disease based on quantification of mri deformation. *PLoS ONE* 12:e0173372. doi: 10.1371/journal.pone.0173372

Luo, W., Phung, D., Tran, T., Gupta, S., Rana, S., Karmakar, C., et al. (2016). Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *J. Med. Internet Res*. 18:e323. doi: 10.2196/jmir.5870

Mehmood, A., Yang, S., Feng, Z., Wang, M., Ahmad, A. S., Khan, R., et al. (2021). A transfer learning approach for early diagnosis of alzheimer's disease on mri images. *Neuroscience* 460, 43–52. doi: 10.1016/j.neuroscience.2021.01.002

Mohan, S., Thirumalai, C., and Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access* 7, 81542–81554. doi: 10.1109/ACCESS.2019.2923707

Montolío, A., Martín-Gallego, A., Cegoñino, J., Orduna, E., Vilades, E., Garcia-Martin, E., et al. (2021). Machine learning in diagnosis and disability prediction of multiple sclerosis using optical coherence tomography. *Comput. Biol. Med*. 133:104416. doi: 10.1016/j.compbiomed.2021.104416

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C., Jagust, W., et al. (2005). The Alzheimer's disease neuroimaging initiative. *Neuroimaging Clin. N. Am*. 15:869. doi: 10.1016/j.nic.2005.09.008

Palaniappan, S., and Awang, R. (2008). "Intelligent heart disease prediction system using data mining techniques," in *2008 IEEE/ACS International Conference on Computer Systems and Applications* (Doha: IEEE), 108–115.

Pan, D., Zeng, A., Jia, L., Huang, Y., Frizzell, T., and Song, X. (2020). Early detection of alzheimer's disease using magnetic resonance imaging: a novel approach combining convolutional neural networks and ensemble learning. *Front. Neurosci*. 14:501050. doi: 10.3389/fnins.2020.00259

Pan, J., Zuo, Q., Wang, B., Chen, C. P., Lei, B., and Wang, S. (2024). Decgan: Decoupling generative adversarial network for detecting abnormal neural circuits in Alzheimer's disease. *IEEE Trans. Artif. Intellig*. doi: 10.1109/TAI.2024.3416420

Pereira, C. R., Weber, S. A., Hook, C., Rosa, G. H., and Papa, J. P. (2016). "Deep learning-aided parkinson's disease diagnosis from handwritten dynamics," in *2016 29th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)* (Sao Paulo: IEEE), 340–346.

Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché Buc, F., et al. (2021). Improving reproducibility in machine learning research: a report from the neurips 2019 reproducibility program. *J. Mach. Learn. Res*. 22, 7459–7478.

Pini, L., Pievani, M., Bocchetta, M., Altomare, D., Bosco, P., Cavedo, E., et al. (2016). Brain atrophy in alzheimer's disease and aging. *Ageing Res. Rev*. 30:25–48. doi: 10.1016/j.arr.2016.01.002

Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., et al. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and ct scans. *Nat. Mach. Intellig*. 3, 199–217. doi: 10.1038/s42256-021-00307-0

Sajda, P. (2006). Machine learning for detection and diagnosis of disease. *Annu. Rev. Biomed. Eng*. 8, 537–565. doi: 10.1146/annurev.bioeng.8.061505.095802

Salehi, A. W., Baglat, P., Sharma, B. B., Gupta, G., and Upadhya, A. (2020). "A CNN model: earlier diagnosis and classification of alzheimer disease using MRI," in *2020 International Conference on Smart Electronics and Communication (ICOSEC)* (Trichy: IEEE), 156–161.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE*

Conference on Computer Vision and Pattern Recognition (Salt Lake City, UT: IEEE), 4510–4520.

Saravanan, N., Sathish, G., and Balajee, J. M. (2018). Data wrangling and data leakage in machine learning for healthcare. *JETIR*. 5, 553–557.

Shapiro, S. S., and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika* 52, 591–611. doi: 10.1093/biomet/52.3-4.591

Shen, L., Margolies, L. R., Rothstein, J. H., Fluder, E., McBride, R., and Sieh, W. (2019). Deep learning to improve breast cancer detection on screening mammography. *Sci. Rep*. 9, 1–12. doi: 10.1038/s41598-019-48995-4

Shorten, C., and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *J. Big Data* 6, 1–48. doi: 10.1186/s40537-019-0197-0

Sokolova, M., and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Inform. Proc. Manage*. 45, 427–437. doi: 10.1016/j.ipm.2009.03.002

Stupple, A., Singerman, D., and Celi, L. A. (2019). The reproducibility crisis in the age of digital medicine. *NPJ Digit. Med*. 2, 1–3. doi: 10.1038/s41746-019-0079-z

Tan, M., and Le, Q. (2019). "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning* (New York: PMLR), 6105–6114.

Tong, T., Wolz, R., Gao, Q., Guerrero, R., Hajnal, J. V., Rueckert, D., et al. (2014). Multiple instance learning for classification of dementia in brain mri. *Med. Image Anal*. 18, 808–818. doi: 10.1016/j.media.2014.04.006

Van De Pol, L. A., Hensel, A., van der Flier, W. M., Visser, P. J., Pijnenburg, Y. A., Barkhof, F., et al. (2006). Hippocampal atrophy on mri in frontotemporal lobar degeneration and alzheimer's disease. *J. Neurol. Neurosurg. Psychiat*. 77, 439–442. doi: 10.1136/jnnp.2005.075341

Van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res*. 9:11.

Vento, D. D., and Fanfarillo, A. (2019). "Traps, pitfalls and misconceptions of machine learning applied to scientific disciplines," in *Proceedings of the Practice and Experience in Advanced Research Computing on Rise of the Machines* (*Learning)* (New York; NY: Association for Computing Machinery), 1–8.

Wen, J., Thibeau-Sutre, E., Diaz-Melo, M., Samper-González, J., Routier, A., Bottani, S., et al. (2020). Convolutional neural networks for classification of alzheimer's disease: Overview and reproducible evaluation. *Med. Image Anal*. 63:101694. doi: 10.1016/j.media.2020.101694

Wu, Y.-T., Beiser, A. S., Breteler, M. M., Fratiglioni, L., Helmer, C., Hendrie, H. C., et al. (2017). The changing prevalence and incidence of dementia over time–current evidence. *Nat. Rev. Neurol*. 13, 327–339. doi: 10.1038/nrneurol.2017.63

Xiao, Z., Ding, Y., Lan, T., Zhang, C., Luo, C., Qin, Z., et al. (2017). Brain mr image classification for Alzheimer's disease diagnosis based on multifeature fusion. *Comput. Math. Methods Med*. 2017:1952373. doi: 10.1155/2017/1952373

Yu, W., Lei, B., Wang, S., Liu, Y., Feng, Z., Hu, Y., et al. (2022). Morphological feature visualization of alzheimer's disease via multidirectional perception gan. *IEEE Trans. Neural Netw. Learn. Syst*. 34, 4401–4415. doi: 10.1109/TNNLS.2021.3118369

Yue, L., Gong, X., Li, J., Ji, H., Li, M., and Nandi, A. K. (2019). Hierarchical feature extraction for early alzheimer's disease diagnosis. *IEEE Access* 7, 93752–93760. doi: 10.1109/ACCESS.2019.2926288

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Commun. ACM* 64, 107–115. doi: 10.1145/3446776

Zong, Y., Zuo, Q., Ng, M. K.-P., Lei, B., and Wang, S. (2024). A new brain network construction paradigm for brain disorder via diffusion-based graph contrastive learning. *IEEE Trans. Pattern Analy. Mach. Intellig*. doi: 10.1109/TPAMI.2024.3442811

Zuo, Q., Wu, H., Chen, C. P., Lei, B., and Wang, S. (2024). Prior-guided adversarial learning with hypergraph for predicting abnormal connections in Alzheimer's disease. *IEEE Trans. Cybernet*. doi: 10.1109/TCYB.2023.3344641

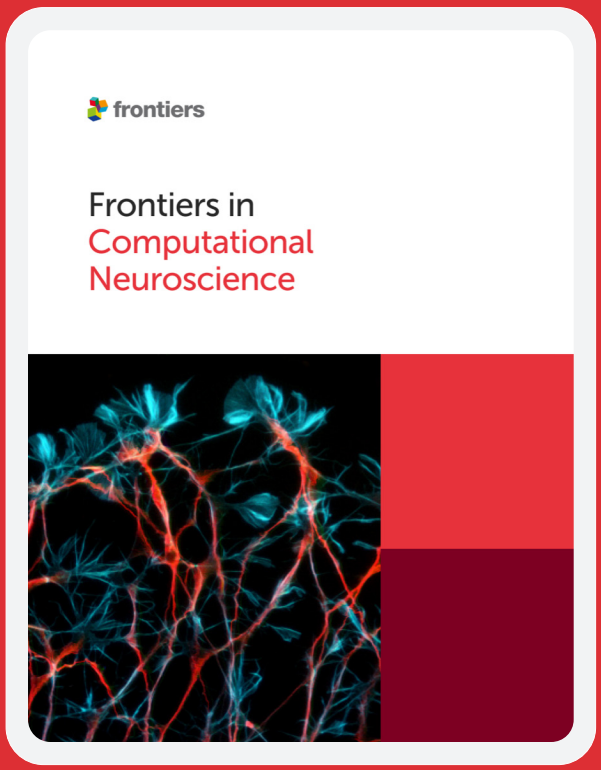
Frontiers in
Computational Neuroscience

Fosters interaction between theoretical and experimental neuroscience

Part of the world's most cited neuroscience series, this journal promotes theoretical modeling of brain function, building key communication between theoretical and experimental neuroscience.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact