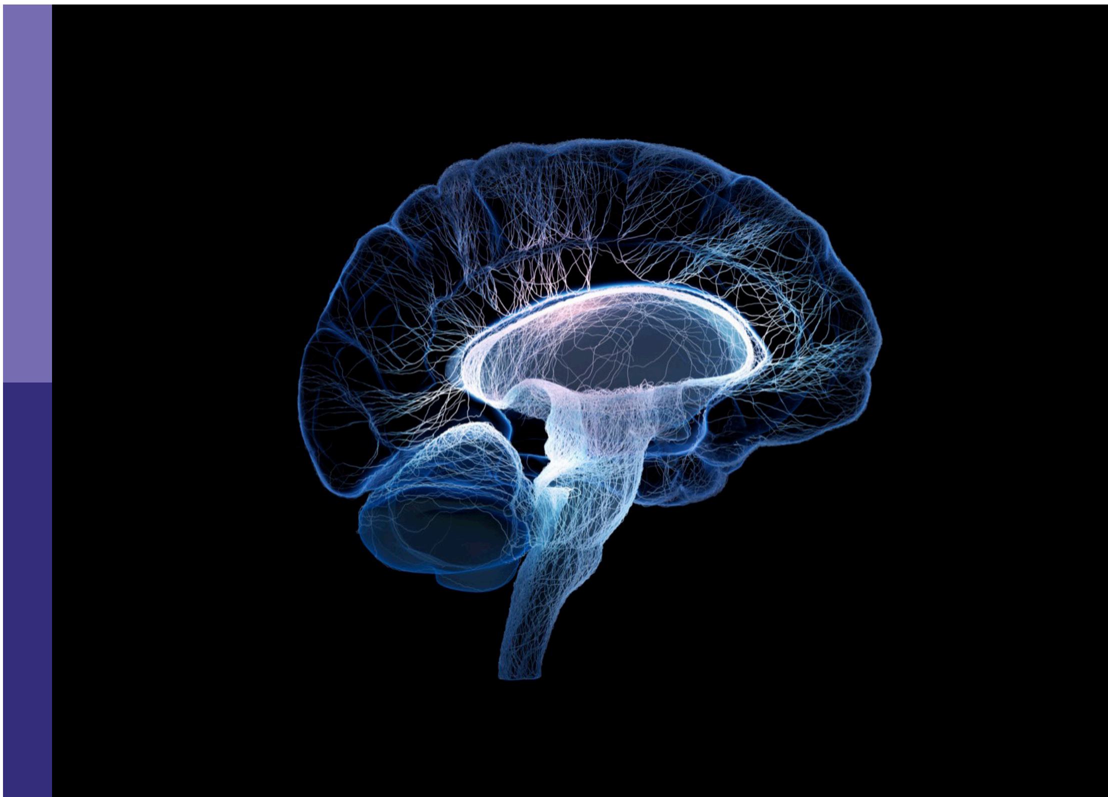
frontiers | Research Topics

Brain functional analysis and brain-like intelligence

Edited by
Shihui Ying, Zhiqiang Tian and Zhengwang Wu

Published in
Frontiers in Neuroscience



ISSN 1664-8714
ISBN 978-2-8325-4615-4
DOI 10.3389/978-2-8325-4615-4

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

Brain functional analysis and brain-like intelligence

Topic editors

Shihui Ying — Shanghai University, China
Zhiqiang Tian — Xi'an Jiaotong University, China
Zhengwang Wu — University of North Carolina at Chapel Hill, United States

Citation

Ying, S., Tian, Z., Wu, Z., eds. (2024). *Brain functional analysis and brain-like intelligence*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-4615-4

Table of contents

frontiers | Frontiers in Neuroscience

TYPE Editorial
PUBLISHED 06 March 2024
DOI 10.3389/fnins.2024.1383481

OPEN ACCESS

EDITED AND REVIEWED BY
Michela Chiappalone,
University of Genoa, Italy

*CORRESPONDENCE
Shihui Ying
✉ shying@shu.edu.cn

RECEIVED 07 February 2024
ACCEPTED 16 February 2024
PUBLISHED 06 March 2024

CITATION
Tian Z, Wu Z and Ying S (2024) Editorial: Brain functional analysis and brain-like intelligence. *Front. Neurosci.* 18:1383481. doi: 10.3389/fnins.2024.1383481

Editorial: Brain functional analysis and brain-like intelligence

Zhiqiang Tian[1], Zhengwang Wu[2] and Shihui Ying[3]*

[1]School of Software Engineering, Xi'an Jiaotong University, Xi'an, China, [2]Department of Radiology, UNC-Chapel Hill, Chapel Hill, NC, United States, [3]Department of Mathematics, School of Science, Shanghai University, Shanghai, China

KEYWORDS

brain functional analysis, brain image understanding, EEG signal-based epileptic seizure prediction, brain-machine fusion, complex network, neuroscience-inspired machine learning

Editorial on the Research Topic
Brain functional analysis and brain-like intelligence

1 Introduction

The Research Topic "*Brain functional analysis and brain-like intelligence*" belongs to the journals, Frontiers in Neuroscience. The aim of this Research Topic is to establish a bridge between brain functional analysis and brain-like machine intelligence, which will promote the basic theory of AI, as well as the mechanism of brain function.

With the improvement in data collection and computing power, artificial intelligence (AI), represented by deep learning, has been developing rapidly. However, there exist huge gaps between natural data and brain data. It therefore suffers from performance degradation if directly applying a traditional deep learning method to the brain data including image, voxel, and electroencephalography (EEG)-based signal. This brings considerable challenges to the brain-related application. Toward these Research Topics, we strive to offer a thorough understanding of the most recent advancements drawing from all the manuscripts that have been published. The key aspects of this topic can be summarized under the following categories: brain image understanding: registration, recognition, and segmentation; EEG signal-based epileptic seizure prediction; AI for brain science.

2 Published papers

2.1 Brain image understanding: registration, recognition, and segmentation

With the continuous development and improvement of computer vision technology, AI-based brain image understanding technology including brain image registration, recognition, and segmentation, has increasingly important value in improving the accuracy and efficiency of clinical diagnosis. As an important upstream task of brain image understanding, brain image registration plays an important role on significantly affecting the subsequent downstream process. However, it typically suffers from high model complexity due to the ill-conditioned inverse problem of brain image registration (Fu et al., 2020). Toward this Research Topic, Fang et al. in the article "*Decoupled learning for brain* *image registration*" decomposed this problem into two simpler sub-problems and adopted two light neural networks to approximate their solutions to reduce the complexity.

With the well registered brain images, its recognition is qualified to be conducted. However, existing brain image recognition methods typically suffer from weak learning ability on shape features. Toward this Research Topic, the article "*STNet: shape and texture joint learning through two-stream network for knowledge guided image recognition*" by Wang et al. proposed a shape and texture joint learning mechanism. In this work, the pyramid-grouped convolution and the deformable convolution are adopted to enhance the shape features.

For medical visual segmentation, the article "*Dual consistent pseudo label generation for multi-source domain adaptation without source data for brain image segmentation*" by Cai et al. proposed a pseudo label generation mechanism for multi-source domain adaptation for brain image segmentation. In this method, a dual consistency constraint including the inter-domain and the intra-domain is presented to guide the generation of the pseudo labels. Beside 2-d segmentation, the 3-d voxels identification is also deserved attention. The article "*Groupwise structural sparsity for discriminative voxels identification*" by Ji, Zhang et al. tackled the absence information of sufficient sample sizes for psychological experiments by proposing a stable hierarchical voting (SHV) mechanism. SHV is enabled to evaluate the quality of spatial random sampling and minimizes the risk of false and missed detection.

2.2 EEG signal-based epileptic seizure prediction

In clinical settings, automatic epileptic seizure prediction is crucial to reducing the heavy burden for patients with intractable epilepsy (Zhao et al., 2021). Electroencephalography (EEG) signals record brain activity and provide valuable information about brain dysfunction. But visually evaluating these signals, which is a non-invasive and affordable way to detect seizures, can be time-consuming and subjective. Thus, there's room for improvement. The article "*Epileptic seizure detection with deep EEG features by convolutional neural network and shallow classifiers*" by Zeng et al. exploits the usage of deep learning to achieve automatically detecting seizures as an urgent problem in clinical application. To effectively detect seizures, EEG signals are adopted as input and the combination of deep feature extractor and shallow classifier is proved to be the most effective. Different from this work, in the article "*An effective fusion model for seizure prediction: GAMRNN*", Ji, Xu, et al. explored the effectiveness of the convolutional attention module toward electroencephalography-based epileptic seizure recognition. In this work, the effectiveness of Lion optimizer is also demonstrated in terms of convergence and the ability to facilitate the recognition performance.

2.3 AI for brain science

Brain-machine interfaces (BMI) have developed rapidly in recent years, but still face critical issues such as accuracy and stability (Liu et al., 2020). By mimicking the architecture and functionality of biological nervous systems, neuromorphic computing models emerge as a potential avenue for creating advanced neuroprosthesis with exceptional performance. Qi et al. demonstrates that neuromorphic computing could be a promising method to realize BMI in the article "*Neuromorphic computing facilitates deep brain-machine fusion for high-performance neuroprosthesis*." The article demonstrates utilizing neuro-morphological computational models to simulate the characteristics of biological neural systems contributes to realizing brain-machine integration and bring new breakthroughs for high-performance and long-term-usable BMI systems.

In recent years, the dynamic behavior of complex networks, especially neural networks, has attracted extensive attention because it can help us understand how the brain processes information, stores memories, and makes decisions (Shine, 2021). The article "*Learning based sliding mode synchronization for fractional order Hindmarsh-Rose neuronal models with deterministic learning*" by Chen et al. proposed a learning based sliding mode control algorithm is proposed by using the deterministic learning (DL) mechanism. With DL mechanism, the synchronization process can be started quickly by recalling the empirical dynamics of neurons. Therefore, fast synchronization effect is achieved by reducing the online computing time.

Inspired by neuroscience, some interpretable machine learning algorithms have been proposed (Lindsay, 2020), such as reinforcement learning mechanisms that simulate brain function. Zhao et al. proposed a neuroscience-inspired reinforcement learning mechanism in the article "*A semi-independent policies training method with shared representation for heterogeneous multi-agent reinforcement learning*." It is claimed to be the first work to adopt a hard-parameter-sharing scheme to multi-agent reinforcement learning for balancing the conflicting requirements of agents' specialization and fast network convergence.

Moreover, collision prediction algorithms based on the neural model of lobule giant motion detectors (LGMD) is also deserved attention (Zhang et al., 2022). Zheng et al. proposed a LGMD-based model with a binocular structure in the article "*Enhancing LGMD based model for collision prediction via binocular structure*" to address the issue that existing LGMD-based methods are not qualified to learn the valuable depth distance feature. In this work, The depth distance of the moving object is extracted by calculating the binocular disparity facilitating a clear differentiation of the motion patterns.

3 Conclusion

Toward the huge gaps between the traditional artificial intelligence and brain science, this Research Topic has gathered considerable original research articles, which made an attempt to establish a bridge between brain functional analysis and brain-like machine intelligence. Among these articles, many brain data-oriented deep learning methods are proposed, toward various downstream tasks such as registration, recognition, segmentation, and detection, which have a vital significance to the clinical application. Moreover, AI for brain science is also discussed, such as brain-machine interfaces and complex neural network construction of brain. These works all contribute positively to reducing the gaps.

Author contributions

ZT: Writing – original draft, Writing – review & editing. ZW: Writing – review & editing. SY: 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 NSFC under Grant No. 62173269.

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

Fu, Y., Lei, Y., Wang, T., Curran, W. J., Liu, T., Yang, X., et al. (2020). Deep learning in medical image registration: a review. *Phys. Med. Biol.* 65:20TR01. doi: 10.1088/1361-6560/ab843e

Lindsay, G. W. (2020). Attention in psychology, neuroscience, and machine learning. *Front. Comput. Neurosci.* 14:29. doi: 10.3389/fncom.2020.00029

Liu, Z., Tang, J., Gao, B., Yao, P., Li, X., Liu, D., et al. (2020). Neural signal analysis with memristor arrays towards high-efficiency brain-machine interfaces. *Nat. Commun.* 11:4234. doi: 10.1038/s41467-020-18105-4

Shine, J. M. (2021). The thalamus integrates the macrosystems of the brain to facilitate complex, adaptive brain network dynamics. *Prog. Neurobiol.* 199:101951. doi: 10.1016/j.pneurobio.2020.101951

Zhang, Y., Zhao, J., Hua, M., Luan, H., Liu, M., Lei, F., et al. (2022). "O-lgmd: an opponent colour lgmd-based model for collision detection with thermal images at night," in *International Conference on Artificial Neural Networks* (Cham: Springer), 249–260. doi: 10.1007/978-3-031-15934-3_21

Zhao, S., Yang, J., and Sawan, M. (2021). Energy-efficient neural network for epileptic seizure prediction. *IEEE Trans. Biomed. Eng.* 69:401. doi: 10.1109/TBME.2021.3095848

TYPE Perspective
PUBLISHED 12 May 2023
DOI 10.3389/fnins.2023.1153985

Neuromorphic computing facilitates deep brain-machine fusion for high-performance neuroprosthesis

Yu Qi[1]*, Jiajun Chen[1] and Yueming Wang[2]*

[1]Affiliated Mental Health Center & Hangzhou Seventh People's Hospital, MOE Frontier Science Center for Brain Science and Brain-Machine Integration, Zhejiang University School of Medicine, Hangzhou, China, [2]Qiushi Academy for Advanced Studies, Zhejiang University, Hangzhou, China

Check for updates

OPEN ACCESS

EDITED BY
Zhiqiang Tian,
Xi'an Jiaotong University, China

REVIEWED BY
Shoeb Shaikh,
Siemens, Singapore

*CORRESPONDENCE
Yu Qi
✉ qiyu@zju.edu.cn
Yueming Wang
✉ ymingwang@zju.edu.cn

RECEIVED 30 January 2023
ACCEPTED 10 April 2023
PUBLISHED 12 May 2023

CITATION
Qi Y, Chen J and Wang Y (2023) Neuromorphic computing facilitates deep brain-machine fusion for high-performance neuroprosthesis.
Front. Neurosci. 17:1153985.
doi: 10.3389/fnins.2023.1153985

Brain-machine interfaces (BMI) have developed rapidly in recent years, but still face critical issues such as accuracy and stability. Ideally, a BMI system would be an implantable neuroprosthesis that would be tightly connected and integrated into the brain. However, the heterogeneity of brains and machines hinders deep fusion between the two. Neuromorphic computing models, which mimic the structure and mechanism of biological nervous systems, present a promising approach to developing high-performance neuroprosthesis. The biologically plausible property of neuromorphic models enables homogeneous information representation and computation in the form of discrete spikes between the brain and the machine, promoting deep brain-machine fusion and bringing new breakthroughs for high-performance and long-term usable BMI systems. Furthermore, neuromorphic models can be computed at ultra-low energy costs and thus are suitable for brain-implantable neuroprosthesis devices. The intersection of neuromorphic computing and BMI has great potential to lead the development of reliable, low-power implantable BMI devices and advance the development and application of BMI.

KEYWORDS

brain-machine interface, brain-computer interface, neuromorphic model, brain-like computing, neuroprosthesis, brain-machine fusion

1. Introduction

Brain-machine interface (BMI) is a technology that enables direct interaction between the brain and external devices such as cursors, robotic arms, and prosthetic limbs, which has demonstrated great potential in various applications, including gaming, smart homes, and neural or motor rehabilitation (Hochberg et al., 2012).

Most recently, intracortical brain-machine interfaces (iBMI), which decode information from single-neuron-level neural signals, have seen rapid progress and enabled new forms of neuroprosthesis, such as brain-to-handwriting (Willett et al., 2021), BMI-based speech synthesis (Moses et al., 2021), and implantable neural therapies for epilepsy (Berényi et al., 2012) and depression (Scangos et al., 2021). The emergence of BMI technology companies, represented by Neuralink, has sparked a wave of rapid development of brain-implantable hardware and devices, boosting the clinical application of BMIs.

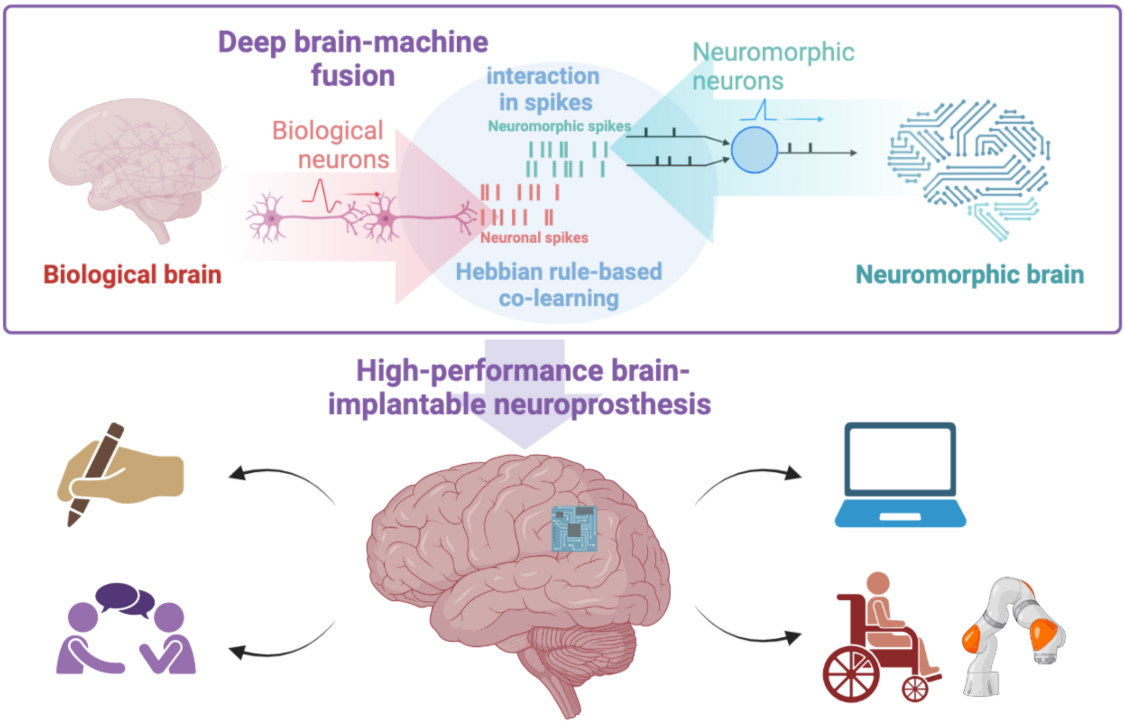


FIGURE 1
Neuromorphic computing facilitates deep brain-machine fusion.

2. Challenges for high-performance BMIs

Ideally, an iBMI system would take the form of brain-implantable neuroprosthesis and would work collaboratively with the brain, like an extension of the brain (Wu et al., 2016). The brain and the iBMI-based neuroprosthesis should be closely connected and integrated, with both sides adapting to, learning with, and compensating for each other as one. However, such a deep connection is difficult to achieve, given the fundamental difference between the brain and the machine. Specifically, from the side of the biological brain, information is encoded in spike trains of neurons. While from the side of computing machines, the basic unit for computation is vectors in real values. The gap between representation and computing lays a barrier to deep fusion between brain and machine, degrading the performance of iBMI systems. Lacking the deep connection between the brain and the machine, the existing iBMI systems still face critical challenges that have seriously hindered clinical application, including:

2.1. Degree of freedom and accuracy

Most motor iBMIs can only control 2–3 degrees of freedom at the same time, typical applications include 2D cursors and 3D robotic arms. The accuracy of the online control process is around 60–90% with full brain control with a path efficiency of 0.4–0.8 (Collinger et al., 2013; Wodlinger et al., 2014), which still cannot meet the clinical use requirements.

2.2. Adaptation

Most existing BMI systems lack the ability to adapt over time and exhibit limited cross-day or long-term performance (Qi et al., 2019; Degenhart et al., 2020). Since brain signals change dynamically over time, a BMI system usually has to be recalibrated every day to maintain its performance, which seriously affects the user experience (Brandman et al., 2018).

2.3. Low-cost computing

In particular, brain signals are high-throughput data, and neural decoding approaches are commonly energy-intensive, leading to issues such as low battery life. Thus, most existing brain-implantable devices only contain a limited number of channels (usually below 50 recording channels) (Rosenthal and Reynolds, 2019; Shupe et al., 2021). Especially, for brain-implantable devices, existing wireless devices usually cannot continuously work for more than 1–2 days (Shaikh et al., 2019).

3. Neuromorphic computing facilitates deep brain-machine fusion

Neuromorphic computing models, which mimic the structure and mechanism of biological neural circuits, provide a promising new option for building high-performance neuroprosthesis (**Figure 1**).

Neuromorphic computing technologies, such as spiking neural networks (SNNs), simulate neuron models of the brain and compute in the form of discrete spikes (Maass, 1997). Computational neuron models, such as the Hodgkin–Huxley model, spike response model, and leaky integrate-and-fire (LIF), mimicking the behavior of biological neurons, are the basic unit for information representation and computation. And the learning process is based on discrete spikes generated by the neuromorphic neurons, following the Hebbian rules and spike timing-dependent plasticity (STDP) rule that resemble biological nervous systems. Additionally, supervised algorithms like tempotron and resume (Gütig and Sompolinsky, 2006), which are derived from artificial neural network technologies, can also be utilized in the learning process.

Another advantage of neuromorphic computing is the ability of ultra-low-cost computing. The spike-based computing is an event-driven asynchronous process, which greatly saves computing energy consumption and realizes ultra-low power consumption computing deployed on neuromorphic chips. Taking partial integro-differential equations solving task as an example, the neuromorphic computing systems simulate the brain's neural processes, the neuromorphic computing chip TrueNorth (Merolla et al., 2014) demonstrates a much lower power consumption (10^{-3} to 10^{-1} W) than commodity server-class computing chips (such as the Intel Xeon E5-2662, which consumes around 10^2 W) (Smith et al., 2022), while still achieving comparable performance.

These features make the neuromorphic computing model a suitable option for developing such high-performance neuroprosthesis.

3.1. Providing a deep and precise connection between brain and machine

With the natural biological plausibility, neuromorphic models enable homogeneous information representation and computation between brain and machine, by direct information transfer in the form of spike trains, which can potentially enclose the connection between both sides. Traditionally, neuronal spike trains are transformed into continuous values in temporal bins to be fed into decoders (Hochberg et al., 2012; Willett et al., 2021), where the precise timing and spike order between neurons are inevitably lost. The direct spike-based interaction between brain and machine enables more precise information transfer, thus can boost the accuracy and stability of BMI systems.

3.2. Facilitating brain-machine co-adaptation

With the Hebbian learning rule that is shared between biological neurons and neuromorphic neurons, BMI systems can learn and develop adaptively with the brain in an online process, which is able to bring new breakthroughs for long-term BMI systems. Besides, neuromorphic models are also expected to overcome the issue of "catastrophic forgetting," which is prevalent in current machine learning models (Imam and Cleland, 2020). They thus are able to perform continuous learning, and facilitate long-term and stable BMIs.

3.3. Enabling fully-implantable BMI devices

With the assistance of neuromorphic chips, neuromorphic models can compute with ultra-low energy cost (Basu et al., 2018), providing an ideal solution for wireless fully brain-implantable neuroprosthesis devices (Shaikh et al., 2019).

Currently, although there are only a few studies on the intersection of neuromorphic computing and BMI, they demonstrate the potential advantages of neuromorphic-model-based neural decoding. Imam and Cleland (2020) proposed a neuromorphic olfactory circuit for online learning of odor recognition and demonstrated the superiority of neuromorphic models in online one-shot learning and continuous learning. Li et al. (2019) proposed a "bioelectronic nose" using SNN decoder to decode odor information from neural activities recorded from the olfactory bulb of rats, demonstrating that neuromorphic models have improved performance and sensitivity (quicker response) compared to traditional machine learning approaches. Kasabov (2014) proposed a special neuromorphic model called NeuCube, which has demonstrated superior performance in brain signal processing tasks. Dethier et al. (2013) implemented a Kalman filter with spike computing and constructed a real-time cursor control BMI system, and found that a neuromorphic network with 2,000 neurons can achieve a success rate of over 94%, and the performance is stably maintained for at least 1 h in a pinball task. These studies demonstrated the advantages of neuromorphic model-based BMIs to some extent, while the deep fusion between the brain and machine, and the close intersection between neuromorphic computing and BMI is to be studied. Especially, with the advantages of neuromorphic computing models, the performance of BMI can be improved in both accuracy and stability, and BMI devices can hopefully meet the requirements of being small, energy-efficient, and fully brain-implantable, which could greatly benefit the clinical use and commercialization of BMIs.

4. Discussion

The field of BMI is currently in a period of rapid development. Neuromorphic computing, with its advantages of biological plausibility, continuous learning, and ultra-low energy consumption, perfectly aligns with the core challenges that BMI faces. The intersection of neuromorphic computing and BMI holds immense promise for the development of reliable and low-power implantable BMI devices and would significantly improve the long-term stability and usability of BMIs.

Data availability statement

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

Author contributions

YQ proposed the perspective and wrote the manuscript. JC contributed to the manuscript writing. YW provided funding and contributed to the paper manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was partly supported by grants from the National Key R&D Program of China (2018YFA0701400), the Key R&D Program of Zhejiang (2022C03011), the Lingang Laboratory (LG-QS-202202-04), the National Natural Science Foundation of China (62276228), the Fundamental Research Funds for the Central Universities, and the Starry Night Science Fund of Zhejiang University Shanghai Institute for Advanced Study (SNZJU-SIAS-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.

References

Basu, A., Acharya, J., Karnik, T., Liu, H., Li, H., and Seo, J. -S. (2018). Low-power, adaptive neuromorphic systems: Recent progress and future directions. *IEEE J. Emerg. Sel. Topics Circuits Syst.* 8, 6–27. doi: 10.1109/JETCAS.2018.2816339

Berényi, A., Belluscio, M., Mao, D., and Buzsáki, G. (2012). Closed-loop control of epilepsy by transcranial electrical stimulation. *Science* 337, 735–737. doi: 10.1126/science.1223154

Brandman, D. M., Hosman, T., Saab, J., Burkhart, M., Shanahan, B., Ciancibello, J., et al. (2018). Rapid calibration of an intracortical brain–computer interface for people with tetraplegia. *J. Neural Eng.* 15:026007. doi: 10.1088/1741-2552/aa9ee7

Collinger, J. L., Wodlinger, B., Downey, J. E., Wang, W., Tyler-Kabara, E., Weber, D., et al. (2013). High-performance neuroprosthetic control by an individual with tetraplegia. *Lancet* 381, 557–564. doi: 10.1016/S0140-6736(12)61816-9

Degenhart, A. D., Bishop, W. E., Oby, E. R., Tyler-Kabara, E., Chase, S., Batista, A., et al. (2020). Stabilization of a brain–computer interface via the alignment of low-dimensional spaces of neural activity. *Nat. Biomed. Eng.* 4, 672–685. doi: 10.1038/s41551-020-0542-9

Dethier, J., Nuyujukian, P., Ryu, S. I., Shenoy, K., and Boahen, K. (2013). Design and validation of a real-time spiking-neural-network decoder for brain–machine interfaces. *J. Neural Eng.* 10:036008. doi: 10.1088/1741-2560/10/3/036008

Gütig, R., and Sompolinsky, H. (2006). The tempotron: a neuron that learns spike timing–based decisions. *Nat. Neurosci.* 9, 420–428. doi: 10.1038/nn1643

Hochberg, L. R., Bacher, D., Jarosiewicz, B., Masse, N., Simeral, J., Vogel, J., et al. (2012). Reach and grasp by people with tetraplegia using a neurally controlled robotic arm. *Nature* 485, 372–375. doi: 10.1038/nature11076

Imam, N., and Cleland, T. A. (2020). Rapid online learning and robust recall in a neuromorphic olfactory circuit. *Nat. Mach. Intell.* 2, 181–191.

Kasabov, N. K. (2014). NeuCube: A spiking neural network architecture for mapping, learning and understanding of spatio-temporal brain data. *Neural Netw.* 52, 62–76. doi: 10.1016/j.neunet.2014.01.006

Li, M., Ruan, H., Qi, Y., Guo, T., Wang, P., and Pan, G. (2019). Odor recognition with a spiking neural network for bioelectronic nose. *Sensors* 19:993. doi: 10.3390/s19050993

Maass, W. (1997). Networks of spiking neurons: the third generation of neural network models. *Neural Netw.* 10, 1659–1671. doi: 10.1016/S0893-6080(97)00011-7

Merolla, P. A., Arthur, J. V., Alvarez-Icaza, R., Cassidy, A., Sawada, J., Akopyan, F., et al. (2014). Artificial brains. A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science* 345, 668–673.

Moses, D. A., Metzger, S. L., Liu, J. R., Anumanchipalli, G., Makin, J., Sun, P., et al. (2021). Neuroprosthesis for decoding speech in a paralyzed person with anarthria. *N. Engl. J. Med.* 385, 217–227.

Qi, Y., Liu, B., Wang, Y., and Pan, G. (2019). Dynamic ensemble modeling approach to nonstationary neural decoding in brain-computer interfaces. *Adv. Neural Inf. Process Syst.* 32, 6089–6098.

Rosenthal, J., and Reynolds, M. S. (2019). A 1.0-Mb/s 198-pJ/bit Bluetooth Low-Energy compatible single sideband backscatter uplink for the NeuroDisc brain–computer interface. *IEEE Trans. Microw. Theory Techn.* 67, 4015–4022.

Scangos, K. W., Khambhati, A. N., Daly, P. M., Makhoul, G., Sugrue, L., Zamanian, H., et al. (2021). Closed-loop neuromodulation in an individual with treatment-resistant depression. *Nat. Med.* 27, 1696–1700.

Shaikh, S., So, R., Sibindi, T., Libedinsky, C., and Basu, A. (2019). Towards intelligent intracortical BMI (i ^{2}BMI): Low-power neuromorphic decoders that outperform Kalman filters. *IEEE Trans. Biomed. Circuits Syst.* 13, 1615–1624.

Shupe, L. E., Miles, F. P., Jones, G., Yun, R., Mishler, J., Rembado, I., et al. (2021). Neurochip3: An autonomous multichannel bidirectional brain-computer interface for closed-loop activity-dependent stimulation. *Front. Neurosci.* 15:718465. doi: 10.3389/fnins.2021.718465

Smith, J. D., Hill, A. J., Reeder, L. E., Franke, B., Lehoucq, R., Parekh, O., et al. (2022). Neuromorphic scaling advantages for energy-efficient random walk computations. *Nat. Electron.* 5, 102–112.

Willett, F. R., Avansino, D. T., Hochberg, L. R., Henderson, J., and Shenoy, K. (2021). High-performance brain-to-text communication via handwriting. *Nature* 593, 249–254.

Wodlinger, B., Downey, J. E., Tyler-Kabara, E. C., Schwartz, A., Boninger, M., and Collinger, J. (2014). Ten-dimensional anthropomorphic arm control in a human brain- machine interface: difficulties, solutions, and limitations. *J. Neural eng.* 12:016011.

Wu, Z., Zhou, Y., Shi, Z., and Zhang, C. (2016). Cyborg intelligence: recent progress and future directions. *IEEE Intell. Syst.* 31, 44–50.

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 22 May 2023
DOI 10.3389/fnins.2023.1145526

Check for updates

OPEN ACCESS

EDITED BY
Zhengwang Wu,
University of North Carolina at Chapel Hill,
United States

REVIEWED BY
Longwang Huang,
Chongqing University of Posts and
Telecommunications, China
Qingbo Meng,
University College London, United Kingdom

*CORRESPONDENCE
Wei Zeng
✉ zengwei@lyun.edu.cn

RECEIVED 16 January 2023
ACCEPTED 02 May 2023
PUBLISHED 22 May 2023

CITATION
Zeng W, Shan L, Su B and Du S (2023) Epileptic seizure detection with deep EEG features by convolutional neural network and shallow classifiers. *Front. Neurosci.* 17:1145526.
doi: 10.3389/fnins.2023.1145526

Epileptic seizure detection with deep EEG features by convolutional neural network and shallow classifiers

Wei Zeng[1,2]*, Liangmin Shan[1,2], Bo Su[1,2] and Shaoyi Du[3]

[1]School of Physics and Mechanical and Electrical Engineering, Longyan University, Longyan, China, [2]School of Mechanical Engineering and Automation, Fuzhou University, Fuzhou, China, [3]Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, China

Introduction: In the clinical setting, it becomes increasingly important to detect epileptic seizures automatically since it could significantly reduce the burden for the care of patients suffering from intractable epilepsy. Electroencephalography (EEG) signals record the brain's electrical activity and contain rich information about brain dysfunction. As a non-invasive and inexpensive tool for detecting epileptic seizures, visual evaluation of EEG recordings is labor-intensive and subjective and requires significant improvement.

Methods: This study aims to develop a new approach to recognize seizures automatically using EEG recordings. During feature extraction of EEG input from raw data, we construct a new deep neural network (DNN) model. Deep feature maps derived from layers placed hierarchically in a convolution neural network are put into different kinds of shallow classifiers to detect the anomaly. Feature maps are reduced in dimensionality using Principal Component Analysis (PCA).

Results: By analyzing the EEG Epilepsy dataset and the Bonn dataset for epilepsy, we conclude that our proposed method is both effective and robust. These datasets vary significantly in the acquisition of data, the formulation of clinical protocols, and the storage of digital information, making processing and analysis challenging. On both datasets, extensive experiments are performed using a cross-validation by 10 folds strategy to demonstrate approximately 100% accuracy for binary and multi-category classification.

Discussion: In addition to demonstrating that our methodology outperforms other up-to-date approaches, the results of this study also suggest that it can be applied in clinical practice as well.

KEYWORDS

electroencephalogram (EEG), epileptic seizure detection, deep features, shallow classifiers, deep neural network (DNN), convolution neural network

1. Introduction

Epileptic seizures are brain's electrical activities that occurs suddenly and unexpectedly (Arab et al., 2010). It affects the daily life of more than 50 million individuals in the world due to the brain dysfunction (Solaija et al., 2018). The recurrent epileptic seizure usually occurs without any obvious external symptoms (Zhou et al., 2020). Currently, using metal electrodes fixed to the brain scalp in a standard configuration, electroencephalogram (EEG) signals record neural activity. Physiologically, they offer deep insight into the brain's state and can be used to detect seizure onsets non-invasively and economically. Traditionally, clinical diagnosis relies on the visual screening and inspection of pronged EEG recordings by board-certified physicians, which is cumbersome, subjective and error-prone (Martis et al., 2015). A reliable, efficient, and accurate EEG analysis and classification system is therefore urgently needed to detect seizures in a timely manner. To handle this problem, different tools have been developed and applied rapidly in recent years, including signal processing and artificial intelligence (Gupta et al., 2018; Li et al., 2019; Subasi et al., 2019; Shoeibi et al., 2021; Tuncer et al., 2021a).

Detection of seizures using EEG generally involves two phases: separating features and classifying them. In the first phase, numerous features generated from four domains, including time, frequency, time-frequency, and non-linear, are incorporated. To analyse time-domain characteristics, morphological parameters, including duration, amplitude, kurtosis, and peak are representative (Wang et al., 2020). There is widespread use of fast Fourier transform (Li and Chen, 2021), as well as power spectral density in frequency domain analysis, provided the EEG signal is static (Al Ghayab et al., 2018). The EEG signal, however, does not display stationarity. Hence, methods of time-frequency domain analysis are usually used for the analysis o time-varying properties of the EEG signal (Sharma et al., 2020), such as time-frequency distribution (Wu et al., 2021) and wavelet transform (Tuncer et al., 2021b). In wavelet transforms, relative frequency information, which is present at low frequencies as well as relative time information, is captured at high frequencies via multiresolution analysis (Sharmila and Geethanjali, 2019). In addition to the wavelet transform, other variations have been proposed, such as the empirical wavelet transform, wavelet packet transform, and wavelet packet entropy. Another popular approach to extracting features is the empirical mode decomposition (EMD) in combination with its variants (Li et al., 2021). Intrinsic mode functions (IMFs) are created when the EEG signal is broken into subsignals. Nonetheless, EMD cannot handle multi-channel signals. Cura and Akan (2021) proposed a single- and multi-channel EEG-based dynamic pattern decomposition (DMD) method to analyze epileptic signals. They extracted high-order spectral moments and subband powers to detect seizure. In non-linear domain, complexity metrics are proposed to depict chaotic properties of the EEG signal, like Hurst exponent, Lyapunov exponent, and various entropies. Other kinds of non-linear metrics, such as Lempel-Ziv complexity, have also been widely used. Rout et al. (2021) used variational mode decomposition (VMD) to identify three band-limited eigenmode functions (BLIMFs) in EEG raw data. In order to derive information-rich spectral and temporal features from BLIMFs, the Hilbert Transform was applied. In addition, the most discriminatory compressed form of privileged information was analyzed based on approximate entropy (ApEn). Anuragi et al. (2022) employed EWT to break down the EEG recordings into Fourier Bessel Series Expansion (FBSE) based subbands. These subbands were then reconstructed as a three-dimensional (3D) phase space representation (PSR). An Euclidean distance of the 3D PSR was used in order to calculate features like line length, log energy entropy, and norm energy entropy. Shankar et al. (2021) used a recurrence plot (RP) technique to analyze brain rhythms with two-dimensional images generated from the EEG signal, which could preserve the non-linear characteristics of EEG. As an additional assessment of image quality, RP entropy and root mean square skewness were used along with RP image criteria.

In the second phase, a variety of machine learning algorithms were proposed to extract EEG signal features, such as artificial neural networks and logistic regression (Abbasi and Goldenholz, 2019; Beniczky et al., 2021). EEG signals during seizures were differentiated using DWT and arithmetic coding by Amin et al. (2020). Various classifiers were then used to detect seizure activity, including Naïve Bayes (NB), multi-layer perceptron (MLP), k nearest neighbors (KNN), and support vector machine (SVM). Anter et al. (2022) utilize a NB based hybrid genetic whale optimization algorithm for feature selection. Afterwards, the ictal and non-ictal EEG signals were classified using an adaptive ELM based on a differential evolution algorithm. To separate EEG signals into distinct bands, Shoeibi et al. (2022) used TQWT. Then, 13 different types of fuzzy entropies were calculated as features from different subbands. Afterwards, EEG recordings were separated using an adaptive neuro-fuzzy inference system.

Due to rapid development in deep learning (DL) over the past few years, several emerging algorithms have been utilized to handle seizure detection problems. While building a multi-layer neural network, DL approaches can minimize the impact of irrelevant features and alleviate computation costs. Acharya et al. (2018) developed a multi-layer deep convolutional neural network (CNN) to determine whether a patient was in a normal, preictal, or seizure state. At present, the generalization and classification abilities of existing DL models may be limited by the use of inter-layer static connection weights. To overcome such problems, A new network architecture called Variable Weight Convolutional Neural Networks (VWCNN) was proposed by Jia et al. (2022). In its convolutional and fully-connected layers, dynamic weights were used instead of static weights to adapt to different EEG characteristics. This model could handle a variety of situations. Sahani et al. (2021) used modified particle swarm optimization based on log energy entropy maxima to calculate optimized values. Then, epileptic seizures were detected using a combination of multiple complex deep neural networks.

Among machine learning systems, representative features have often been hand-designed and empirically chosen. Such systems are more likely to produce false positives and are prone to misdiagnosis. By contrast, DL automatically generates features instead of using any hard-crafted features, and have the potential to provide superior classification performance (Murat et al., 2021). These techniques automate feature extraction and no manual feature extraction is required due to the end-to-end structure of DL models. In this work, we build an efficient and reliable deep neural network (DNN) to recognize epilepsy, utilizing features from CNN layers without any preprocessing of input EEG signals. This study makes a major contribution to the identification of presence and developing stages of seizure using information from deep feature maps of CNN together with shallow classifiers. An effective way for reduction of the dimensionality of deep feature maps is the employment of Principal Component Analysis (PCA) (Jolliffe and Cadima, 2016).

Throughout the article, the following structures are followed. The proposed method is described in detail in Section 2, which includes description of EEG data, extraction of deep feature, and EEG classification for seizure detection. Section 3 designs comprehensive experiments and provides corresponding results.

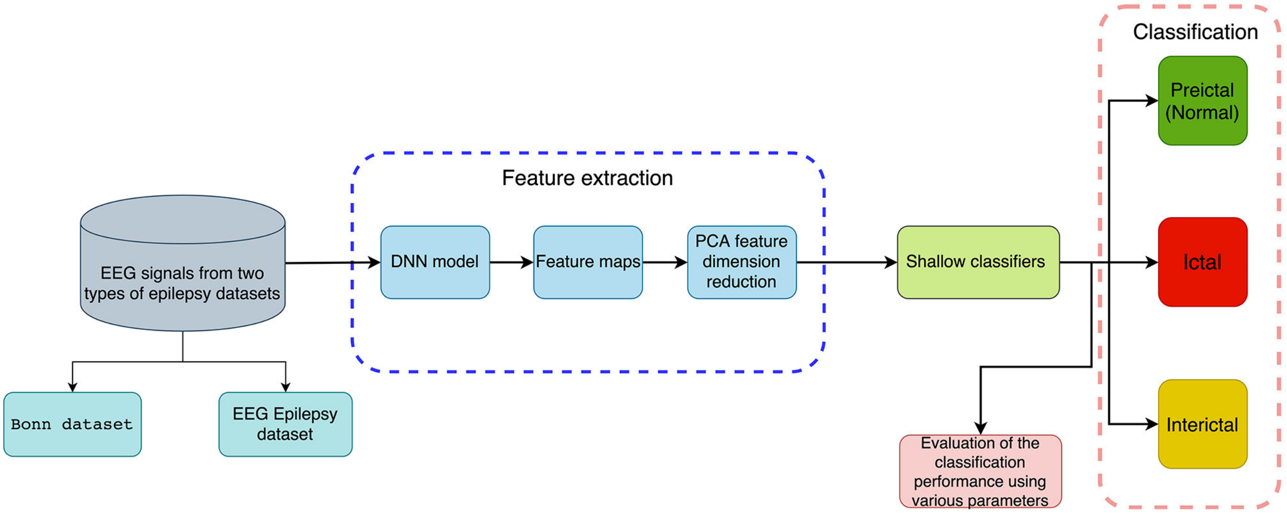

Classification
Preictal (Normal)
Feature extraction
EEG signals from two types of epilepsy datasets
DNN model
Feature maps
PCA feature dimension reduction
Shallow classifiers
Ictal
Bonn dataset
EEG Epilepsy dataset
Evaluation of the classification performance using various parameters
Interictal

FIGURE 1
An illustration of the proposed method for classifying EEG recordings with deep features and shallow classifiers for the detection of epileptic seizures (binary and multi-class classification).

Section 4 presents a comprehensive discussion about the results and contribution. Section 5 gives a brief conclusion.

2. Materials and methods

This section briefly introduces a method for distinguishing normal and abnormal EEG signals with information extracted through deep features for detecting epileptic seizures. It consists of a feature extraction phase and a classification phase, which includes several steps. Firstly, EEG recordings are subjected to DNN-based feature extraction without any preprocessing, followed by PCA reduction of feature dimension. Secondly, features are put into five traditional machine learning classifiers to detect epileptic seizures. It includes binary classification (seizure vs. seizure-free or preictal vs. interictal) and multi-class classification (preictal vs. interictal vs. ictal). A flowchart showing our method is available in Figure 1.

2.1. EEG database

2.1.1. Dataset-1

A part of the experimental data for this study comes from the Bonn dataset, which is publicly available (Andrzejak et al., 2001). Each subset of the dataset contains 100 artifact-free, single-channel intracranial EEG clips of 23.6 s each, labeled A, B, C, D, and E (also Z, O, N, F, and S, accordingly). An amplifier system with 128 channels and a band-pass filter between 0.53 and 40 Hz was used to record the EEG signals at 173.61 Hz. Therefore, each signal contains 4,097 records, that is, each signal has a data length of 4,097. These data are demonstrated in Figure 2. Table 1 summarizes details about this dataset.

2.1.2. Dataset-2

In Dateset-2, segmented EEG recordings were obtained from 10 epilepsy patients (Swami et al., 2016). With a GrassTelefactor Comet AS40 amplifier system and a 200 Hz sampling rate, all EEG recordings were acquired. The duration of each EEG recording is approximately 5.12 s (1,024 samples). These data are demonstrated in Figure 2. The scalp electrodes for EEG recordings were gold-plated and adhered to the 10-20 standard in compliance with the recording procedure. First, an EEG signal was filtered with a bandpass filter having a cutoff frequency of 0.5 and 70 Hz. Afterwards, It was divided by clinical experts into ictal (group F), interictal (group G), and preictal (group H) phases. Table 2 summarizes details about this dataset.

2.2. Deep feature extraction

DL techniques learn a set of empirical features at multiple abstraction levels, capable of learning complex functions through input data independent of hand-crafted features. It undergoes a learning process by progressively extracting multiple features from low layers to high layers (Murat et al., 2021). Therefore, we use the DNN-based model to automatically generate features. Figure 3 demonstrates this DNN-based model.

Our DNN model outputs feature maps after we have connected the convolutional layer. PCA is used to remove useless features and reduce redundancy, which can alleviate the computational cost and enhance the performance and generalization. Figure 3 demonstrates the feature extraction steps and details.

Table 3 summarizes a detailed parameter representation of the DNN model. We add a Batch Normalization (BatchNorm) layer after each convolutional layer, with axis 2 and momentum 0.9 to speed up training. An activation function for rectified linear unit (ReLU) follows each BatchNorm layer. We use L_2 regularization to alleviate overfitting with a dropout of 0.4 upon reaching the

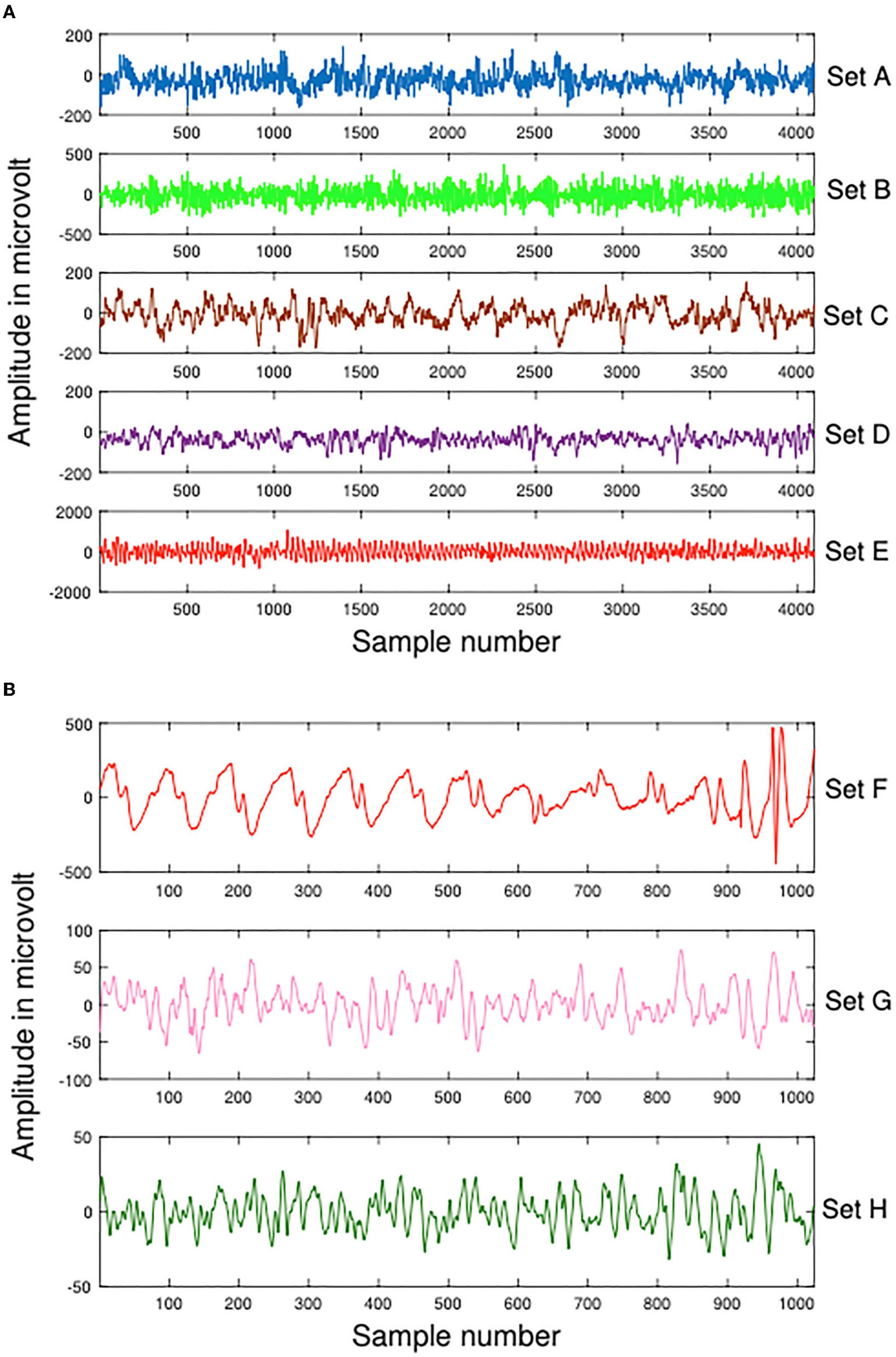


FIGURE 2
Samples of Dataset-1 Bonn dataset and Dataset-2 EEG Epilepsy dataset. **(A)** Dataset-1 Bonn dataset A, B, C, D, and E. **(B)** Dataset-2 EEG Epilepsy dataset F, G, and H.

TABLE 1 Overview of Dataset-1.

Items	Set A	Set B	Set C	Set D	Set E
Participants	5 healthy controls	5 healthy controls	5 epileptic patients	5 epileptic patients	5 epileptic patients
Electrode types	Scalp	Scalp	Intracranial	Intracranial	Intracranial
Participants' states	Awake with opened eyes	Awake with closed eyes	Interictal	Interical	Ictal
Total number of epochs	100	100	100	100	100
Sampling rate (Hz)	173.61	173.61	173.61	173.61	173.61
Duration of each epoch (second)	23.6	23.6	23.6	23.6	23.6

TABLE 2 Overview of Dataset-2.

Items	Set F	Set G	Set H
Participants	10 epilepsy patients	10 epilepsy patients	10 epilepsy patients
Electrode types	Scalp	Scalp	Scalp
Participants' states	Ictal	Interictal	Preictal (normal)
Total number of epochs	50	50	50
Sampling rate (Hz)	200	200	200
Duration of each epoch (second)	5.12	5.12	5.12

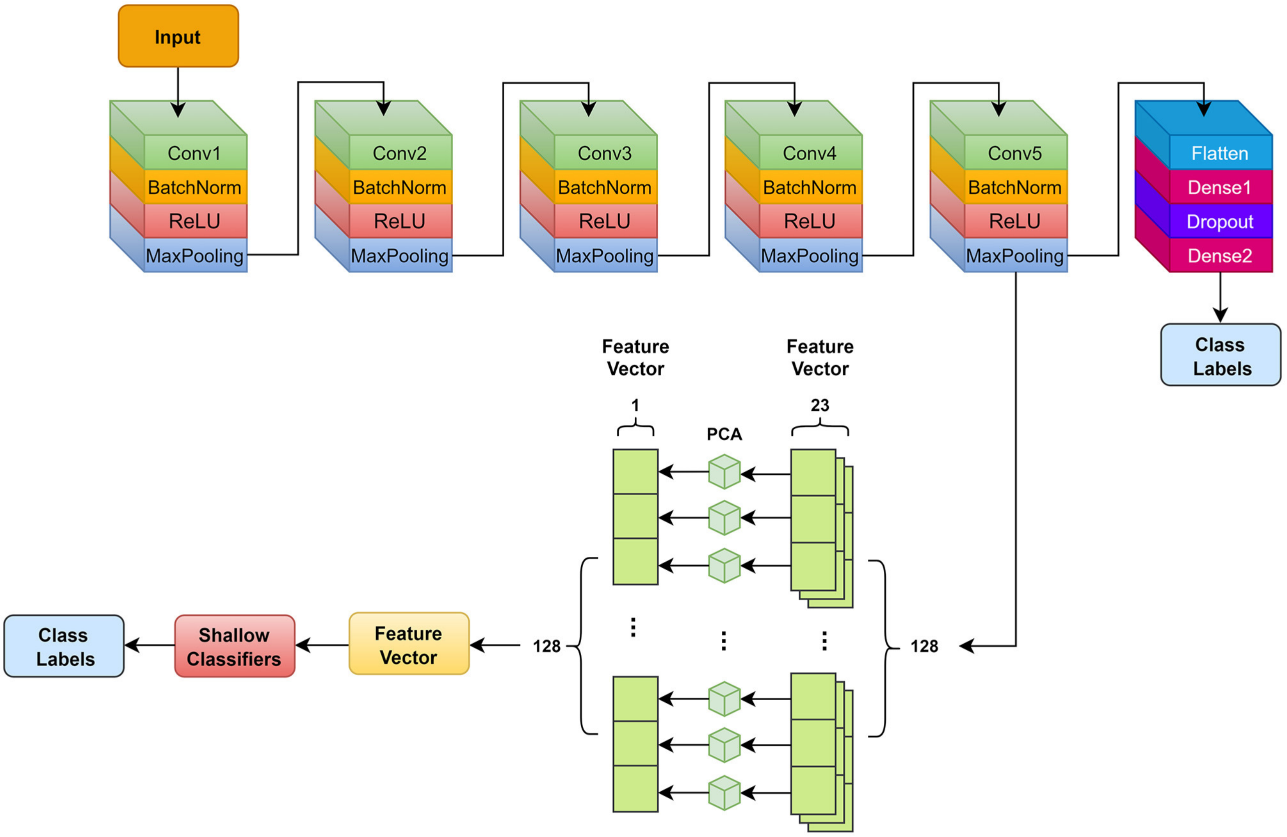


FIGURE 3
Deep neural network model and feature extraction used in this study. Conv, convolution.

first fully connected layer. Aggregate data are used for subject-level assessments. Our neural network weights are updated by using the cross-entropy loss function and Adam optimization. There are three settings: 0.0001, 50, and 300, which are the learning rate, batch size, and epochs. A 0.001 learning rate is applied to the data, a batch size of 50, and 300 epochs are used when

TABLE 3 Model summary of DNN.

No	Layer name	Layer parameters	Output shape	Number of params
1	1D Convolution	Filters = 32, kernel_size = 3, input_shape = (4097,1), stride = 1, padding = “valid”	(4095,32)	128
2	BatchNorm	Axis = 2, momentum = 0.9	(4095,32)	128
3	Activation	ReLU	(4095,32)	0
4	1D MaxPooling	Pool_size = 2 stride = 2 padding = “valid”	(2047,32)	0
5	1D Convolution	Filters = 64, kernel_size = 5, stride = 1, padding = “valid”	(2043,64)	10,304
6	BatchNorm	Axis = 2, momentum = 0.9	(2043,64)	256
7	Activation	ReLU	(2043,64)	0
8	1D MaxPooling	Pool_size = 4, stride = 4, padding = “valid”	(510,64)	0
9	1D Convolution	Filters = 128, kernel_size=13, stride = 1, padding = “valid”	(498,128)	106,624
10	BatchNorm	Axis = 2, momentum = 0.9	(498,128)	512
11	Activation	ReLU	(498,128)	0
12	1D MaxPooling	Pool_size = 4, stride = 4, padding = “valid”	(124,128)	0
13	1D Convolution	Filters = 256, kernel_size = 17, stride = 1, padding = “valid”	(108,256)	557,312
14	BatchNorm	Axis = 2, momentum = 0.9	(108,256)	1,024
15	Activation	ReLU	(108,256)	0
16	1D MaxPooling	Pool_size = 2, stride = 2, padding = “valid”	(54,256)	0
17	1D Convolution	Filters = 128, kernel_size = 9, stride = 1, padding = “valid”	(46,128)	295,040
18	BatchNorm	Axis = 2, momentum = 0.9	(46,128)	512
19	Activation	ReLU	(46,128)	0
20	1D MaxPooling	Pool_size = 2, stride = 2, padding = “valid”	(23,128)	0
21	Flatten	-	2,944	0
22	Dense	Unit = 64, activation = “ReLU”, kernel_regularizer = L_2 (0.03)	64	188,480
23	Dropout	Rate = 0.4	64	0
24	Dense	Unit = 5, activation = “softmax”	5	325

training the model. A feature map sized 23 × 128 is exported from the MaxPooling layer ahead of the flatten layer. We split the eigenvectors into 128 small eigenvectors of shape size 23 × 1. PCA is then used to perform dimensionality reduction on each of the small eigenvectors, resulting in 128 eigenvectors of shape size 1 × 1. These feature vectors are concatenated with 1 × 128 shape size and fed into shallow classifiers below for classification.

2.3. Machine learning classifiers

For epileptic seizure detection, in addition to support vector classifier (SVC) (Lau and Wu, 2003), several classical machine learning classifiers are employed, including k-nearest neighbors (KNN) (Kramer, 2013), gradient boosting (GB) (Natekin and Knoll, 2013), random forest (RF) (Lau and Scornet, 2016), Gaussian Naïve Bayes (GNB) (Griffis et al., 2016), decision tree (DT) (Safavian and Landgrebe, 1991), and multi-layer perception (MLP) (Murtagh, 1991). Shallow classifiers are still the classifier of choice despite deep learning approaches becoming increasingly overwhelming. To solve supervised classification problems, discriminant analysis is utilized to reduce the distance between each class and increase the variability between different classes (Ye et al., 2004; Murat et al., 2021).

3. Results

We design comprehensive experiments on two databases and illustrate the results of classifying EEG categories into binary and multi-class. The DNN model is implemented in TensorFlow backend using a 10-core Intel Core i9 CPU and RTX3090 GPU on a high-performance computer.

Cross-validation using a K-fold ($K = 10$) method verifies the effectiveness of the classification. Each iteration will use $K - 1$ folds to train and the remaining folds to test. In addition to accuracy (ACC), specificity (SPF), and sensitivity (SEN), we use another four classic performance indicators: negative predictive value (NPV), positive predictive value (PPV), F1 score, and Matthews correlation coefficient (MCC). Here is the calculation: a True Positive is equal to TP, a False Negative is equal to FN, a True Negative is equal to

TABLE 4 Different experimental cases for Bonn and EEG Epilepsy datasets.

Dataset	Case	Groups	Description				
			Type 1	Type 2	Type 3	Type 4	Type 5
Bonn	1	A vs. E	Normal (eyes open)	Ictal	-	-	-
	2	B vs. E	Normal (eyes closed)	Ictal	-	-	-
	3	AB vs. E	Normal	Ictal	-	-	-
	4	C vs. E	Interictal	Ictal	-	-	-
	5	D vs. E	Interictal	Ictal	-	-	-
	6	CD vs. E	Interictal	Ictal	-	-	-
	7	A vs. D	Normal	Interictal	-	-	-
	8	ABCD vs. E	Non-seizure	Seizure	-	-	-
	9	AB vs. CDE	Normal	Epileptic	-	-	-
	10	A vs. C vs. E	Normal	Interitcal	Ictal	-	-
	11	AB vs. CD vs. E	Normal	Interitcal	Ictal	-	-
	12	A vs. B vs. C vs. D vs. E	Normal (eyes open)	Normal (eyes closed)	Interictal	Interictal	Ictal
EEG Epilepsy	I	F vs. G	Ictal	Interictal	-		
	II	F vs. H	Ictal	Preictal	-		
	III	G vs. H	Interictal	Preictal	-		
	IV	F vs. GH	Seizure	Seizure-free	-		
	V	F vs. G vs. H	Ictal	Interictal	Preictal		

TN, and a False Positive is equal to FP. For larger MCC value, the classifier performs better.

$$\mathrm{SEN} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \times 100(\%) \tag{1}$$

$$\mathrm{SPF} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \times 100(\%) \tag{2}$$

$$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FN}+\mathrm{FP}} \times 100(\%) \tag{3}$$

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \times 100(\%) \tag{4}$$

$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FN}} \times 100(\%) \tag{5}$$

$$\mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FN}\times\mathrm{FP}}{\sqrt{(\mathrm{TP}+\mathrm{FN})(\mathrm{TP}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})}} \times 100(\%) \tag{6}$$

$$\mathrm{F1\ score} = \frac{2\times\mathrm{TP}}{2\times\mathrm{TP}+\mathrm{FN}+\mathrm{FP}} \times 100(\%) \tag{7}$$

Table 4 shows the comprehensive experiments setting. Using the Bonn and EEG Epilepsy datasets, twelve and five different classification problems are proposed, respectively. They focus on differentiating between preictal (normal), interictal, and ictal EEG signals, including binary and multi-class classification.

Figures 4, 5 show the overall accuracy and loss curves for the model trained on two datasets. It is obvious that after almost 300 epochs the network converges.

Tables 5, 6 illustrate the classification results for different cases on two datasets, respectively. To further illustrate the performance of each shallow classifier, Figures 6, 7 show the ROC curves and associated AUC for 12 cases of the Bonn dataset and 5 cases of the EEG epilepsy dataset. Our study demonstrates improved accuracy in discriminating between preictal, interictal, and ictal EEG signals. As a whole, the proposed method performs well and yields good results, demonstrating that it can distinguish various classes of EEG signals effectively.

4. Discussion

The seizure detection literature shows that several methods are currently available to handle binary and multi-category classification issues. Experimental results for 17 epilepsy detection cases have been presented and discussed in detail. A comparison of our algorithm with other up-to-date solutions is provided in Table 7.

For Bonn dataset, the first three cases handle binary classification. Regarding Case 1 (A-E), when using the EEG spectrum as input, Cetin et al. (2015) calculated autoregressive coefficients, which were then fed into back propagation (BP) and Elman neural networks. A 98.3% accuracy rate was reported as the best. Jiang et al. (2020) used a symplectic geometric decomposition method to derive features from EEG signals and put them into an SVM for EEG classification. It was reported that the accuracy was

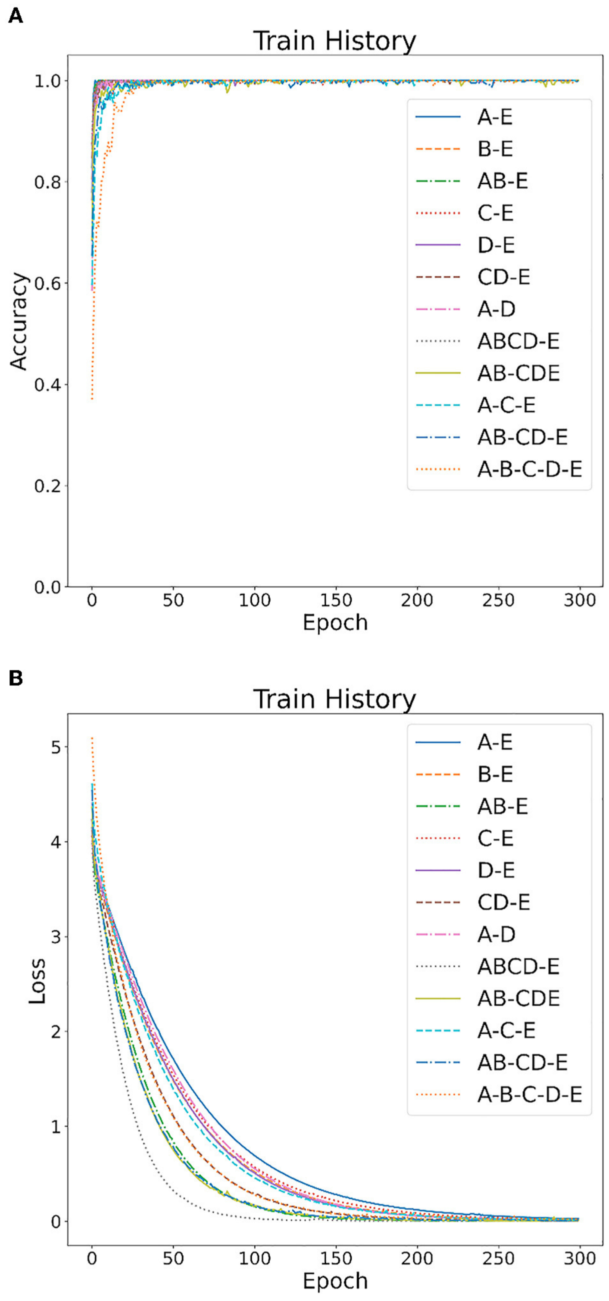


FIGURE 4
CNN model training on Bonn dataset: **(A)** accuracy curve, **(B)** loss curve.

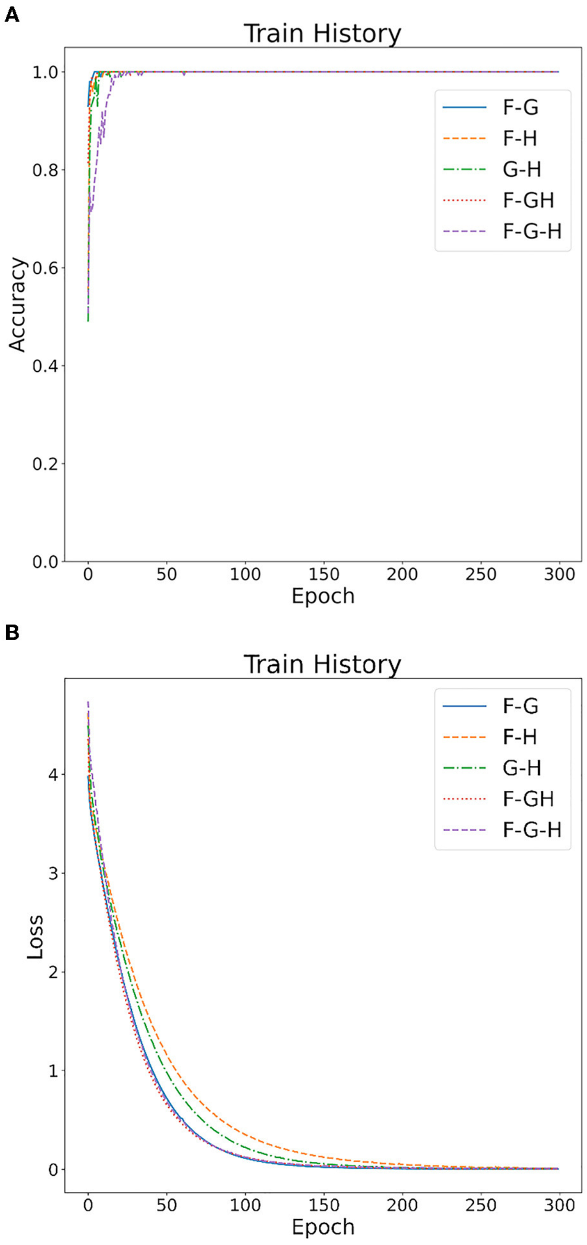


FIGURE 5
CNN model training on EEG Epilepsy dataset: **(A)** accuracy curve, **(B)** loss curve.

100%. In an attempt to find the optimal parameters of an SVM to classify epileptic EEG, a mixture model was constructed using genetic algorithms (GA), as well as particle swarm optimization (PSO) by Subasi et al. (2019). A 99.38% accuracy rate was reported as the best. Using New Weighted Complex Networks (NWCNs), Supriya et al. (2021) extracted three features from EEG data: Modular Gain (MG), Average Weighted Degree (AWD), and Edge Weight Fluctuation (EWF). Three features' separation performance was examined using an SVM model with three different kernels. They obtained 100% classification accuracy. Prabhakar and Lee (2022) employed K-singular value decomposition (K-SVD) to derive sparse descriptions from EEG signals and extracted features using self-organizing maps (SOMs). The data was then fed into ELM, deep learning, and transfer learning models for classification, with an accuracy rate of 98.35%. Unlike previous methods, ours is 100% accurate.

According to Swami et al. (2016), a dual-tree complex wavelet transform (DT-CWT) was employed to divide EEG recordings into multiple subbands on a six-level scale in Case 2 (B-E). These subbands acted as features and classified EEG signals with a general regressive neural network (GRNN). A 98.9% accuracy rate was reported as the best. Ahmedt-Aristizabal et al. (2018)

TABLE 5 The evaluation of the performance of the proposed approach using 10-fold cross-validation style on Bonn dataset with 12 cases.

Case	Classifier	ACC (%)	SPF (%)	SEN (%)	PPV (%)	NPV (%)	MCC (%)	F1 (%)
Case 1	SVC	100	100	100	100	100	100	100
	KNN	100	100	100	100	100	100	100
	RF	100	100	100	100	100	100	100
	GNB	99.50	100	99.00	100	99.00	99.00	99.49
	GB	100	100	100	100	100	100	100
	DT	100	100	100	100	100	100	100
	MLP	100	100	100	100	100	100	100
Case 2	SVC	100	100	100	100	100	100	100
	KNN	100	100	100	100	100	100	100
	RF	100	100	100	100	100	100	100
	GNB	100	100	100	100	100	100	100
	GB	100	100	100	100	100	100	100
	DT	99.50	99.00	100	99.00	100	99.00	99.50
	MLP	100	100	100	100	100	100	100
Case 3	SVC	100	100	100	100	100	100	100
	KNN	100	100	100	100	100	100	100
	RF	100	100	100	100	100	100	100
	GNB	100	100	100	100	100	100	100
	GB	100	100	100	100	100	100	100
	DT	100	100	100	100	100	100	100
	MLP	100	100	100	100	100	100	100
Case 4	SVC	100	100	100	100	100	100	100
	KNN	100	100	100	100	100	100	100
	RF	100	100	100	100	100	100	100
	GNB	99.50	100	99.00	100	99.00	99.00	99.49
	GB	99.50	99.00	100	99.00	100	99.00	99.50
	DT	100	100	100	100	100	100	100
	MLP	100	100	100	100	100	100	100
Case 5	SVC	99.50	100	99.00	100	99.00	99.00	99.49
	KNN	100	100	100	100	100	100	100
	RF	100	100	100	100	100	100	100
	GNB	99.00	100	98.00	100	98.03	98.01	98.98
	GB	99.50	99.00	100	99.00	100	99.00	99.50
	DT	100	100	100	100	100	100	100
	MLP	100	100	100	100	100	100	100
Case 6	SVC	99.66	100	99.50	100	99.00	99.25	99.74
	KNN	100	100	100	100	100	100	100
	RF	99.66	99.00	100	99.50	100	99.25	99.75
	GNB	99.33	99.00	99.50	99.50	99.00	98.49	99.50
	GB	99.33	98.00	100	99.00	100	98.50	99.50
	DT	99.33	99.00	99.50	99.50	99.00	98.49	99.50
	MLP	100	100	100	100	100	100	100

(Continued)

TABLE 5 (Continued)

Case	Classifier	ACC (%)	SPF (%)	SEN (%)	PPV (%)	NPV (%)	MCC (%)	F1 (%)
Case 7	SVC	100	100	100	100	100	100	100
	KNN	100	100	100	100	100	100	100
	RF	100	100	100	100	100	100	100
	GNB	100	100	100	100	100	100	100
	GB	99.50	99.00	100	99.00	100	99.00	99.50
	DT	99.00	99.00	99.00	99.00	99.00	98.00	99.00
	MLP	100	100	100	100	100	100	100
Case 8	SVC	99.80	100	99.75	100	99.00	99.37	99.87
	KNN	100	100	100	100	100	100	100
	RF	100	100	100	100	100	100	100
	GNB	99.80	100	99.75	100	99.00	99.37	99.87
	GB	99.80	99.00	100	99.75	100	99.37	99.87
	DT	99.80	100	99.75	100	99.00	99.37	99.87
	MLP	99.80	100	99.75	100	99.00	99.37	99.87
Case 9	SVC	99.60	99.66	99.50	99.50	99.66	99.16	99.50
	KNN	99.60	99.66	99.50	99.50	99.66	99.16	99.50
	RF	99.80	99.66	100	99.50	100	99.58	99.75
	GNB	99.60	100	99.00	100	99.33	99.16	99.49
	GB	99.80	99.66	100	99.50	100	99.58	99.75
	DT	99.20	99.00	99.50	98.51	99.66	98.33	99.00
	MLP	99.60	99.66	99.50	99.50	99.66	99.16	99.50
Case 10	SVC	99.66	99.83	99.66	99.66	99.83	99.49	99.66
	KNN	99.66	99.83	99.66	99.66	99.83	99.49	99.66
	RF	100	100	100	100	100	100	100
	GNB	99.33	99.66	99.33	99.33	99.66	98.99	99.33
	GB	100	100	100	100	100	100	100
	DT	99.00	99.50	99.00	99.00	99.50	98.49	99.00
	MLP	99.66	99.83	99.66	99.66	99.83	99.49	99.66
Case 11	SVC	99.80	99.90	99.80	99.80	99.90	99.70	99.80
	KNN	99.80	99.90	99.80	99.80	99.90	99.70	99.80
	RF	99.60	99.80	99.60	99.60	99.80	99.40	99.60
	GNB	99.20	99.60	99.20	99.20	99.60	98.80	99.20
	GB	99.20	99.60	99.20	99.20	99.60	98.80	99.20
	DT	99.00	99.50	99.00	99.00	99.50	98.50	99.00
	MLP	99.80	99.90	99.80	99.80	99.90	99.70	99.80
Case 12	SVC	99.80	99.95	99.80	99.80	99.95	99.74	99.80
	KNN	98.80	99.70	98.80	98.80	99.70	98.49	98.80
	RF	98.80	99.70	98.80	98.80	99.70	98.49	98.80
	GNB	93.60	98.40	93.60	93.60	98.40	91.99	93.60
	GB	99.60	99.90	99.60	99.60	99.90	99.49	99.60
	DT	97.60	99.40	97.60	97.60	99.40	96.99	97.60
	MLP	99.80	99.95	99.80	99.80	99.95	99.74	99.80

TABLE 6 An evaluation of the performance of the proposed approach using 10-fold cross-validation with 5 cases of EEG epilepsy dataset.

Case	Classifier	ACC (%)	SPF (%)	SEN (%)	PPV (%)	NPV (%)	MCC (%)	F1 (%)
Case I	SVC	100	100	100	100	100	100	100
	KNN	100	100	100	100	100	100	100
	RF	100	100	100	100	100	100	100
	GNB	100	100	100	100	100	100	100
	GB	100	100	100	100	100	100	100
	DT	100	100	100	100	100	100	100
	MLP	100	100	100	100	100	100	100
Case II	SVC	100	100	100	100	100	100	100
	KNN	100	100	100	100	100	100	100
	RF	100	100	100	100	100	100	100
	GNB	100	100	100	100	100	100	100
	GB	100	100	100	100	100	100	100
	DT	100	100	100	100	100	100	100
	MLP	100	100	100	100	100	100	100
Case III	SVC	100	100	100	100	100	100	100
	KNN	100	100	100	100	100	100	100
	RF	100	100	100	100	100	100	100
	GNB	100	100	100	100	100	100	100
	GB	96.00	92.00	100	92.59	100	92.29	96.15
	DT	98.00	100	96.00	100	96.15	96.07	97.95
	MLP	100	100	100	100	100	100	100
Case IV	SVC	100	100	100	100	100	100	100
	KNN	100	100	100	100	100	100	100
	RF	100	100	100	100	100	100	100
	GNB	100	100	100	100	100	100	100
	GB	100	100	100	100	100	100	100
	DT	99.33	100	98.00	100	99.00	98.50	98.98
	MLP	100	100	100	100	100	100	100
Case V	SVC	100	100	100	100	100	100	100
	KNN	99.33	99.66	99.33	99.33	99.66	98.99	99.33
	RF	98.66	99.33	98.66	98.66	99.33	97.99	98.66
	GNB	99.33	99.66	99.33	99.33	99.66	98.99	99.33
	GB	97.33	98.66	97.33	97.33	98.66	95.99	97.33
	DT	96.66	98.33	96.66	96.66	98.33	94.99	96.66
	MLP	99.33	99.66	99.33	99.33	99.66	98.99	99.33

achieved 94.75% accuracy by using a recurrent neural network (RNN) embedded an LSTM network. Jiang et al. (2020), Supriya et al. (2021), and Prabhakar and Lee (2022) also studied on this classification issue and reported 99.33, 100, and 97.57% accuracies, respectively. Our method, on the other hand, achieves 100% accuracy.

Regarding Case 3 (AB-E), EEG clips are divided into two types: non-ictal and ictal. It was reported that Swami et al. (2016) had an accuracy rate of 99.2%. Sharma et al. (2017) used analytic time-frequency flexible wavelet transform (ATFFWT) and fractal dimensions to export features and put them into a least squares support vector machine (LS-SVM). Afterwards, a 100% accuracy rate was reported as the best. Jiang et al. (2020) and Prabhakar and Lee (2022) also studied on this classification issue and reported 100 and 97.84% accuracies, respectively. Our method, on the other hand, achieves 100% accuracy.

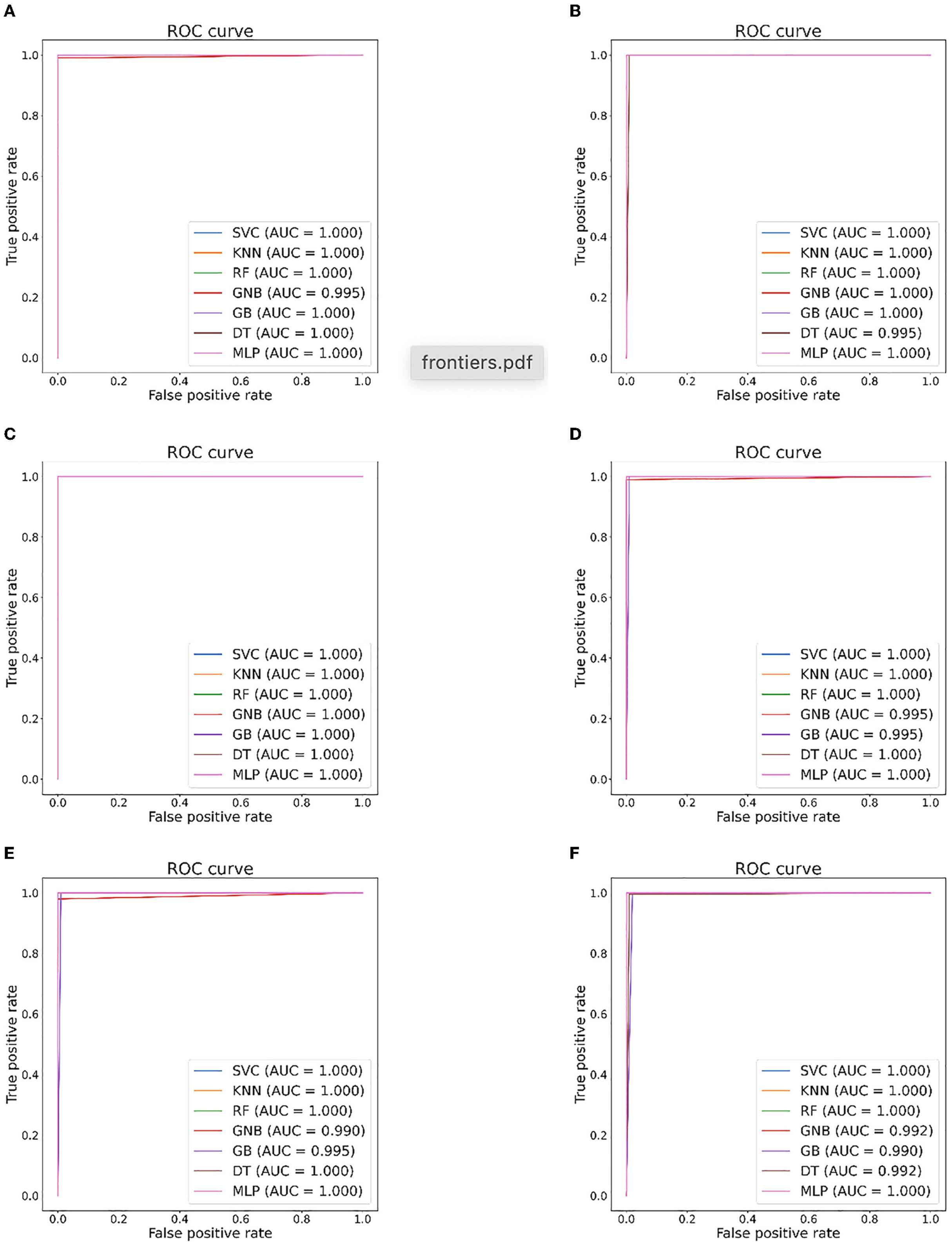


FIGURE 6 (Continued)

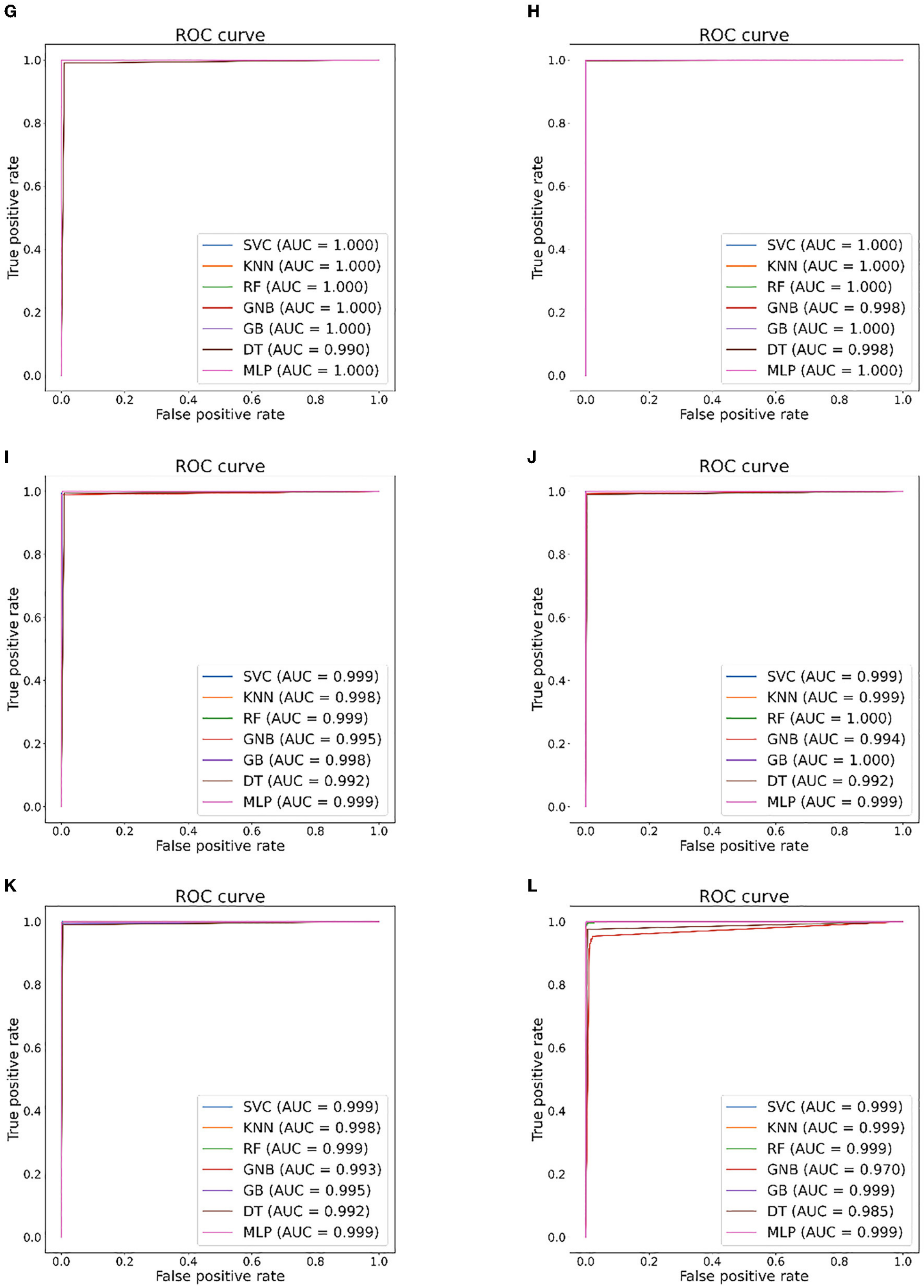


FIGURE 6 (Continued)

Seizure detection ROC curves and AUC from Bonn dataset. **(A)** Case 1: A-E. **(B)** Case 2: B-E. **(C)** Case 3: AB-E. **(D)** Case 4: C-E. **(E)** Case 5: D-E. **(F)** Case 6: CD-E. **(G)** Case 7: A-D. **(H)** Case 8: ABCD-E. **(I)** Case 9: AB-CDE. **(J)** Case 10: A-C-E. **(K)** Case 11: AB-CD-E. **(L)** Case 12: A-B-C-D-E.

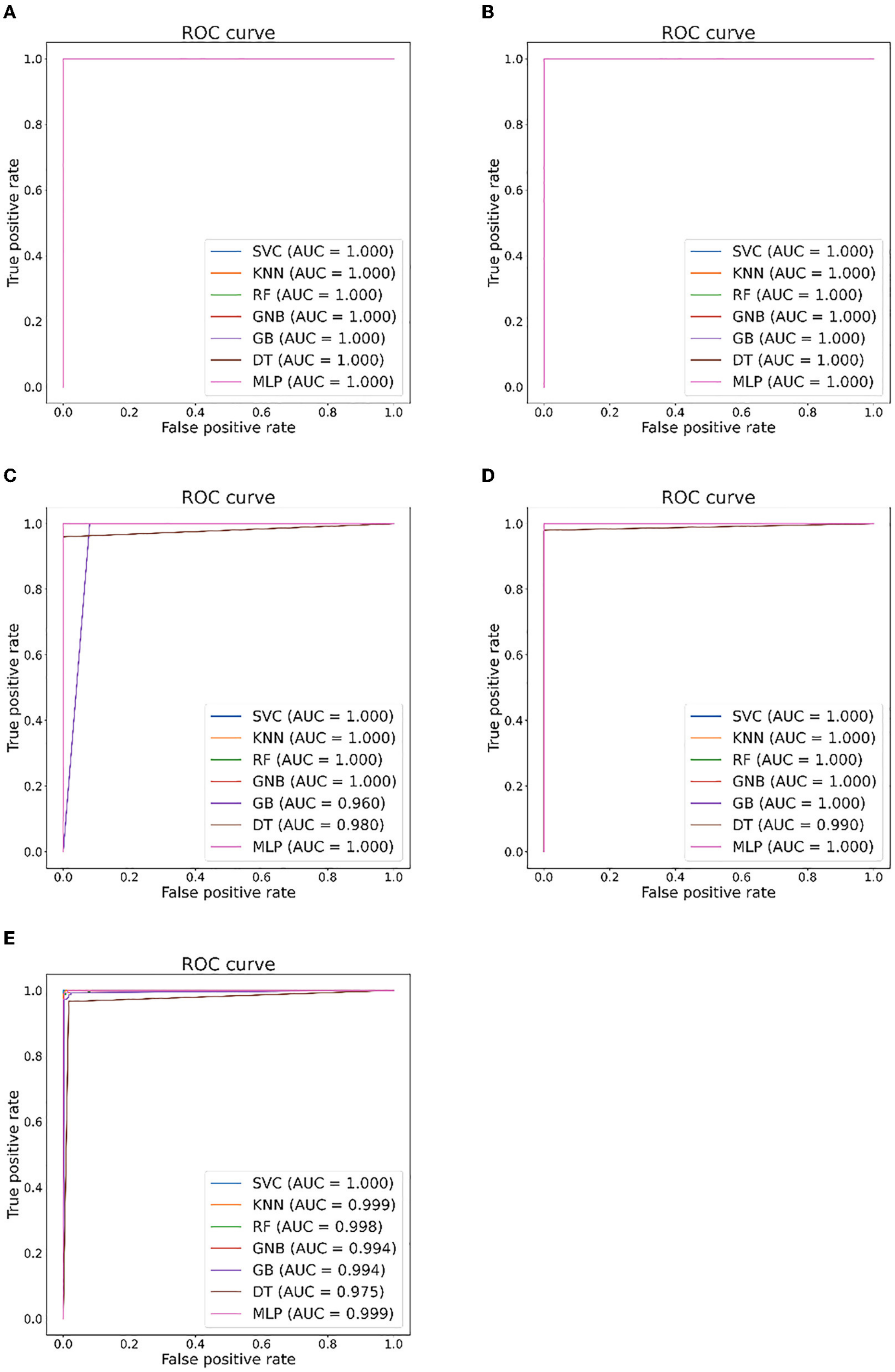


FIGURE 7
Seizure detection ROC curves and AUC from EEG Epilepsy dataset. **(A)** Case I: F-G. **(B)** Case II: F-H. **(C)** Case III: G-H. **(D)** Case IV: F-G. **(E)** Case V: F-G-H.

TABLE 7 Summary of literature comparison results (10-fold cross-validation style).

References	Methodology	Cases	ACC (%)	Our ACC (%)
Cetin et al. (2015)	Autoregressive coefficients with BP + Elman neural networks	A-E	98.3	100
Jiang et al. (2020)	Symplectic geometry eigenvalues + SVM	A-E	100	100
Subasi et al. (2019)	GA, PSO and SVM	A-E	99.38	100
Supriya et al. (2021)	MG, EWF and AWD + SVM	A-E	100	100
Prabhakar and Lee (2022)	K-SVD, SOM + ELM, deep learning, transfer learning	A-E	98.35	100
Swami et al. (2016)	DT-CWT + GRNN	B-E	98.9	100
Ahmedt-Aristizabal et al. (2018)	End-to-end data and RNNs + LSTM	B-E	94.75	100
Jiang et al. (2020)	Symplectic geometry eigenvalues + SVM	B-E	99.33	100
Supriya et al. (2021)	MG, EWF and AWD + SVM	B-E	100	100
Prabhakar and Lee (2022)	K-SVD, SOM + ELM, deep learning, transfer learning	B-E	97.57	100
Swami et al. (2016)	DT-CWT + GRNN	AB-E	99.2	100
Sharma et al. (2017)	ATFFWT + fractal dimension + LS-SVM	AB-E	100	100
Jiang et al. (2020)	Symplectic geometry eigenvalues + SVM	AB-E	100	100
Prabhakar and Lee (2022)	K-SVD, SOM + ELM, deep learning, transfer learning	AB-E	97.84	100
Swami et al. (2016)	DT-CWT + GRNN	C-E	98.7	100
Sharma et al. (2017)	ATFFWT + fractal dimension + LS-SVM	C-E	99	100
Raghu et al. (2019)	Matrix determinant feature + MLP classifier	C-E	97.60	100
Jiang et al. (2020)	Symplectic geometry eigenvalues + SVM	C-E	99.33	100
Supriya et al. (2021)	MG, EWF and AWD + SVM	C-E	100	100
Swami et al. (2016)	DT-CWT + GRNN	D-E	93.3	100
Sharma et al. (2017)	ATFFWT + fractal dimension + LS-SVM	D-E	98.5	100
Raghu et al. (2019)	Matrix determinant feature + MLP classifier	D-E	97.60	100
Jiang et al. (2020)	Symplectic geometry eigenvalues + SVM	D-E	100	100
Supriya et al. (2021)	MG, EWF and AWD + SVM	D-E	100	100
Swami et al. (2016)	DT-CWT + GRNN	CD-E	95.2	100
Sharma et al. (2017)	ATFFWT + fractal dimension + LS-SVM	CD-E	98.67	100
Raghu et al. (2019)	Matrix determinant feature + MLP classifier	CD-E	96.85	100
Jiang et al. (2020)	Symplectic geometry eigenvalues + SVM	CD-E	99.28	100
Gupta et al. (2018)	DCT, Hurst exponent and ARMA + SVM	A-D	98.4	100
Tuncer et al. (2019)	Local senary pattern + SVM	A-D	99.5	100
Hassan et al. (2020)	CEEMDAN + Adaptive Boosting	ABCD-E	99.2	100
Mursalin et al. (2017)	ICFS + RF classifier	ABCD-E	97.4	100
Jiang et al. (2020)	Symplectic geometry eigenvalues + SVM	ABCD-E	99.97	100
Peng et al. (2021)	Stein kernel-based SR	AB-CDE	98.2	99.80
Acharya et al. (2018)	13-layer CNN without performing feature extraction and selection	AB-CDE	88.7	99.80
Jiang et al. (2020)	Symplectic geometry eigenvalues + SVM	AB-CDE	99.17	99.80
Jaiswal and Banka (2017)	LNDP and 1D-LGP + ANN	A-C-E	98.22	100
Gupta and Banka (2019)	WMRPE, rhythms of FBE + LS-SVM	A-C-E	97.3	100
Jiang et al. (2020)	Symplectic geometry eigenvalues + SVM	A-C-E	99.22	100
Zhang et al. (2021)	FSWT-based subbands and CSoS, FuzzyEn, HFD, t-SNE + KNN	A-C-E	99.69	100

(Continued)

TABLE 7 (Continued)

References	Methodology	Cases	ACC (%)	Our ACC (%)
Bhardwaj et al. (2016)	EMD + Constructive Genetic Programming	AB-CD-E	98.33	99.80
Peker et al. (2015)	DT-CWT + CVANN	AB-CD-E	97.79	99.80
Raghu et al. (2019)	Matrix determinant feature + MLP classifier	AB-CD-E	96.5	99.80
Jiang et al. (2020)	Symplectic geometry eigenvalues + SVM	AB-CD-E	99.80	99.80
Zarei and Asl (2021)	DWT and OMP + SVM	AB-CD-E	99.33	99.80
Sharma et al. (2020)	ToC + deep neural network	A-B-C-D-E	97.2	99.80
Zahra et al. (2017)	MVEMD + ANN	A-B-C-D-E	87.2	99.80
Zhang et al. (2021)	FSWT-based subbands and CSoS, FuzzyEn, HFD, t-SNE + KNN	A-B-C-D-E	93.62	99.80
Zhou et al. (2020)	SSA + SVM, ELM and ANN	F-G	94	100
Peng et al. (2021)	Stein kernel-based SR	F-G	98.00	100
Wang et al. (2021)	TVAR-MWBF-UROFR + SVM	F-G	98.18	100
Sukriti et al. (2021)	EMD-MSPCA, RCMSE, RCMFE, RCMPE + SVM	F-G	96.38	100
Tajmirriahi and Amini (2021)	SDE + SVM	F-G	99.1	100
Zhou et al. (2020)	SSA + SVM, ELM and ANN	F-H	95	100
Peng et al. (2021)	Stein kernel-based SR	F-H	99	100
Wang et al. (2021)	TVAR-MWBF-UROFR + SVM	F-H	100	100
Sukriti et al. (2021)	EMD-MSPCA, RCMSE, RCMFE, RCMPE + SVM	F-H	100	100
Tajmirriahi and Amini (2021)	SDE + SVM	F-H	96.8	100
Zhou et al. (2020)	SSA + SVM, ELM and ANN	G-H	93	100
Wang et al. (2021)	TVAR-MWBF-UROFR + SVM	G-H	88.95	100
Sukriti et al. (2021)	EMD-MSPCA, RCMSE, RCMFE, RCMPE + SVM	G-H	97.15	100
Tajmirriahi and Amini (2021)	SDE + SVM	G-H	91.5	100
Zhou et al. (2020)	SSA + SVM, ELM and ANN	F-GH	91	100
Peng et al. (2021)	Stein kernel-based SR	F-GH	97.5	100
Wang et al. (2021)	TVAR-MWBF-UROFR + SVM	F-GH	98.08	100
Peng et al. (2021)	Stein kernel-based SR	F-G-H	97.21	100
Sukriti et al. (2021)	EMD-MSPCA, RCMSE, RCMFE, RCMPE + SVM	F-G-H	93.49	100

Regarding Cases 4 through 6, EEG signals are divided into interictal and ictal types (C-E, D-E, and CD-E). It was reported that Swami et al. (2016) had 98.7, 93.3, and 95.2% accuracies. Sharma et al. (2017) indicated 99, 98.5, and 98.67% accuracy rates. In Raghu et al. (2019), descriptive and bivariate histogram analysis, and polar histogram were used to provide matrix determinant features. The effectiveness was verified on three cases, using the MLP classifier to achieve accuracies of 97.60, 97.60, and 96.85%, respectively. Jiang et al. (2020) also studied on these issues and reported accuracies of 99.33, 100, and 99.28%, respectively. In contrast, our proposed method achieves 100, 100, and 100% accuracies, respectively.

Case 7 (A-D) addresses the classification of normal vs. interictal. Gupta et al. (2018) utilized discrete cosine transform (DCT) to build a multirate filterbank structure, which decomposed EEG signals into their respective brain rhythms. Then, the Hurst exponent together with the autoregressive moving average (ARMA) parameters were derived from the statistical results of the brain rhythms as features. The SVM classifier reported an accuracy of 98.4%. Using Local Military Patterns (LSPs), Tuncer et al. (2019) extracted binary features through EEG signals. A standard deviation based strategy was used to deal with threshold value problems of ternary functions. Then, extracted features were put into SVM for classification with an accuracy rate of 99.5%. Unlike previous methods, ours is 100% accurate.

In Case 8, the EEG is classified as seizure or non-seizure (ABCD-E). An objective method of identifying intrinsic modes was proposed in Hassan et al. (2020) by using complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN). Modeling these mode functions with normal inverse Gaussian (NIG) parameters follows. They employed Adaptive Boosting to classify EEG signals and reported 99.2% accuracy. For feature derivation, Mursalin et al. (2017) examined an improved correlation-based feature selection method (ICFS). A 97.4% accuracy rate was reported for an RF classifier. Jiang et al.

(2020) also focused on this issue and a 99.97% accuracy rate was reported as the best. Our method, on the other hand, achieves 100% accuracy.

Regarding Case 9 (AB-CDE), EEG signals are divided into normal and epileptic types. In Peng et al. (2021), EEG signals were classified in symmetric positive definite (SPD) matrix spaces by using Stein kernel-based sparse representations (SR). They reported accuracy rate of 98.20%. Acharya et al. (2018) developed a multiple-layer CNN algorithm to avoid feature extraction and selection. They reported 88.7% accuracy. Jiang et al. (2020) studied on this classification issue with an accuracy of 99.17%. Unlike previous methods, ours is 99.8% accurate.

In Cases 10 and 11, ternary classification is addressed by both A-C-E and AB-CD-E. We report 100% and 99.80% classification accuracies, respectively. To deal with Case 10, in Jaiswal and Banka (2017), the Local Neighborhood Description Pattern (LNDP) together with the 1D Local Gradient Pattern (1D-LGP) was utilized to export features. An adaptive neural network (ANN) was designed for classification, reporting 98.22% accuracy. Gupta and Banka (2019) achieved feature extraction of rhythms based on a combination of Weighted Multiscale Renyi Permutation Entropy (WMRPE) and Fourier-Bessel Series Expansion (FBSE). To classify these features, LS-SVM was used, and the best accuracy rate was 97.3%. Zhang et al. (2021) proposed a fusion method for feature extraction based on Frequency Sliced Wavelet Transform (FSWT). Then, these feature were fed into a KNN classifier with a classification accuracy of 99.69%. Regarding Case 11, according to Bhardwaj et al. (2016), EEG recordings were split into multiple IMFs, each with a set of bandwidth parameters extracted. They constructed genetic programming for classification and a 98.33% accuracy rate was reported as the best. Peker et al. (2015) used DT-CWT to extract features from EEG signals. EEG data was classified using a complex-valued adaptive neural network (CVANN) and a 97.79% accuracy rate was reported. In Raghu et al. (2019), a 96.5% accuracy rate was reported as the best. Jiang et al. (2020) studied on these classification issues with reported accuracies of 99.22 and 99.80%, respectively. Zarei and Asl (2021) exported different coefficients from EEG signals using DWT and Orthogonal Matching Pursuit (OMP) techniques. Then, some non-linear features and several statistical features were computed using DWT and OMP coefficients. They were put into an SVM classifier, which reported 99.33% accuracy.

In Case 12, the EEG is separated into five categories (A-B-C-D-E). Sharma et al. (2020) used third-order cumulants (ToC) to export features from EEG recording and put them into deep neural networks for classification, reporting 97.2% accuracy. In Zahra et al. (2017), using the MVEMD algorithm, the EEG recordings were decomposed into multiple intrinsic scales. An ANN model was created to classify valid IMFs with a reported accuracy of 87.2%. Zhang et al. (2021) reported 93.62% accuracy. In contrast, our proposed method achieves 99.80% accuracy.

For EEG Epilepsy dataset, Cases I to IV deal with binary classification. Zhou et al. (2020) decomposed the EEG recordings into singular values using singular spectrum analysis (SSA). Then, the log-normalized function values are calculated, forming the eigenvector. They were fed into shallow classifiers, including SVM, ELM, and ANN, to perform with the highest accuracy of 94, 95, 93, and 91% in the four cases. Wang et al. (2021) proposed an autoregressive (AR) model based time-varying (TV) modeling framework to describe EEG recordings. The multiwavelet basis function expansion (MWBF) method was used to approximate the TV parameters of the AR model (TVAR). Afterwards, the resulting extended model was reduced and refined using the Ultra-regularized Orthogonal Regression (UROFR) algorithm. The SVM achieved the highest accuracies of 98.18, 100, 88.95, and 98.08% for the four cases, respectively. Peng et al. (2021) also dealt with Cases I, II and IV and reported accuracies of 98.00, 99, and 97.5%, respectively. The EMD-MSPCA method, developed by Sukriti et al. (2021), combined empirical mode decomposition with multiscale PCA, to denoise EEG recordings. Following that, three complexity measures were used as features. DT, LDA, SVM, and KNN shallow classifiers were used for classification of Cases I, II, and III. The documented accuracy for each is 96.38, 100, and 97.15%. Due to its inherent self-similarity, Tajmirriahi and Amini (2021) used stochastic differential equations (SDEs) to model EEG signals with self-similar fractional Levy stabilization processes. They Fit the probability distribution to the derived EEG signal histogram, and extracted the parameters of the fitted histogram. A SVM classifier was used to classify them, with 99.1, 96.8, and 91.5% accuracies for cases I, II, and III, respectively. In contrast, our approach reports 100, 100, 100, and 100% accuracies for the four cases, respectively.

Case V address ternary classification. Peng et al. (2021) and Sukriti et al. (2021) reported accuracies of 97.21 and 93.49%, respectively. We report the accuracy of 100%, which also outperforms other approaches.

Unlike the aforementioned algorithms, this study designs an DNN model to automatically extract deep features from layer outputs during raining. Afterwards, extracted features are filtered by PCA for dimensionality reduction and directly put into seven shallow classifiers to classify EEG signals. The process is simple, high efficient along with high accuracy. Table 7 illustrates the comparison results on the classification performance between our approaches and other approaches recently proposed. Our method illustrates superior performance and has potential for serving as an adjunct to fMRI in epilepsy diagnosis.

Our experimental results have indicated that the proposed method is highly accurate in detecting epilepsy for binary, three-class, and five-class classification problems, illustrating the suitability of our scheme for solving problems involving multiple classes. The clinical potential of automated analysis of epileptic seizure activity is significant. Additionally, once high-performance computers are utilized, its computational simplicity is enhanced, allowing it to be deployed in clinical applications. As a result, this new approach is better equipped to satisfy clinical demands in terms of efficiency, functionality, universality, and simplicity, while providing satisfactory accuracy. These traits make it an appealing alternative option for clinical diagnosis. Real-time seizure detection for smart healthcare and Internet of Medical Things (IoMT) applications is a potential use case for the proposed method.

5. Conclusion

This study uses different kinds of machine learning classifiers to detect seizure with features derived from the max pooling layers of a DNN model. The suggested algorithm separates EEG recordings into two, three and five classes. The results show that performance of the advised classifier is promising for seizure detection. This model may provide neurologists with additional assistance when diagnosing epilepsy. The work in the future will incorporate a number of handcrafted features (such as intrinsic fuzzy entropy, Lyapunov exponent, and Lempel-Ziv complexity) as well as deep features to design deep learning models and compare them with current model performance. In conclusion, the proposed protocol will speed up epilepsy diagnosis, assist clinicians to implement clinical epilepsy monitoring devices with less burden.

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

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

Author contributions

WZ and SD contributed to the study concept and design. WZ, LS, and BS performed the experiments and data analysis and prepared the draft manuscript. All authors participated manuscript organization and approved the submitted version.

Funding

This work was supported by the Natural Science Foundation of Fujian Province (Grant No. 2022J011146).

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

Abbasi, B., and Goldenholz, D. M. (2019). Machine learning applications in epilepsy. *Epilepsia* 60, 2037–2047. doi: 10.1111/epi.16333

Acharya, U. R., Oh, S. L., Hagiwara, Y., Tan, J. H., and Adeli, H. (2018). Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals. *Comput. Biol. Med.* 100, 270–278. doi: 10.1016/j.compbiomed.2017.09.017

Ahmedt-Aristizabal, D., Fookes, C., Nguyen, K., and Sridharan, S. (2018). "Deep classification of epileptic signals," in *40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (Honolulu, HI), 332–335. doi: 10.1109/EMBC.2018.8512249

Al Ghayab, H. R., Li, Y., Siuly, S., and Abdulla, S. (2018). Epileptic EEG signal classification using optimum allocation based power spectral density estimation. *IET Signal Process.* 12, 738–747. doi: 10.1049/iet-spr.2017.0140

Amin, H. U., Yusoff, M. Z., and Ahmad, R. F. (2020). A novel approach based on wavelet analysis and arithmetic coding for automated detection and diagnosis of epileptic seizure in EEG signals using machine learning techniques. *Biomed. Signal Process. Control* 56, 101707. doi: 10.1016/j.bspc.2019.101707

Andrzejak, R. G., Lehnertz, K., Mormann, F., Rieke, C., David, P., and Elger, C. E. (2001). Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: dependence on recording region and brain state. *Phys. Rev. E* 64, 061907. doi: 10.1103/PhysRevE.64.061907

Anter, A. M., Abd Elaziz, M., and Zhang, Z. (2022). Real-time epileptic seizure recognition using Bayesian genetic whale optimizer and adaptive machine learning. *Future Gen. Comput. Syst.* 127, 426–434. doi: 10.1016/j.future.2021.09.032

Anuragi, A., Sisodia, D. S., and Pachori, R. B. (2022). Epileptic-seizure classification using phase-space representation of FBSE-EWT based EEG sub-band signals and ensemble learners. *Biomed. Signal Process. Control* 71, 103138. doi: 10.1016/j.bspc.2021.103138

Arab, M. R., Suratgar, A. A., and Ashtiani, A. R. (2010). Electroencephalogram signals processing for topographic brain mapping and epilepsies classification. *Comput. Biol. Med.* 40, 733–739. doi: 10.1016/j.compbiomed.2010.06.001

Beniczky, S., Karoly, P., Nurse, E., Ryvlin, P., and Cook, M. (2021). Machine learning and wearable devices of the future. *Epilepsia* 62, S116–S124. doi: 10.1111/epi.16555

Bhardwaj, A., Tiwari, A., Krishna, R., and Varma, V. (2016). A novel genetic programming approach for epileptic seizure detection. *Comput. Methods Prog. Biomed.* 124, 2–18. doi: 10.1016/j.cmpb.2015.10.001

Biau, G., and Scornet, E. (2016). A random forest guided tour. *Test* 25, 197–227. doi: 10.1007/s11749-016-0481-7

Cetin, G. D., Cetin, O., and Bozkurt, M. R. (2015). The detection of normal and epileptic EEG signals using ANN methods with matlab-based GUI. *Int. J. Comput. Appl.* 114, 45–50. doi: 10.5120/20034-2145

Cura, O. K., and Akan, A. (2021). Analysis of epileptic EEG signals by using dynamic mode decomposition and spectrum. *Biocybern. Biomed. Eng.* 41, 28–44. doi: 10.1016/j.bbe.2020.11.002

Griffis, J. C., Allendorfer, J. B., and Szaflarski, J. P. (2016). Voxel-based Gaussian naive Bayes classification of ischemic stroke lesions in individual T1-weighted MRI scans. *J. Neurosci. Methods* 257, 97–108. doi: 10.1016/j.jneumeth.2015.09.019

Gupta, A., Singh, P., and Karlekar, M. (2018). A novel signal modeling approach for classification of seizure and seizure-free EEG signals. *IEEE Trans. Neural Syst. Rehabil. Eng.* 26, 925–935. doi: 10.1109/TNSRE.2018.2818123

Gupta, V., and Pachori, R. B. (2019). Epileptic seizure identification using entropy of FBSE based EEG rhythms. *Biomed. Signal Process. Control* 53, 101569. doi: 10.1016/j.bspc.2019.101569

Hassan, A. R., Subasi, A., and Zhang, Y. (2020). Epilepsy seizure detection using complete ensemble empirical mode decomposition with adaptive noise. *Knowledge Based Syst.* 191, 105333. doi: 10.1016/j.knosys.2019.105333

Jaiswal, A. K., and Banka, H. (2017). Local pattern transformation based feature extraction techniques for classification of epileptic EEG signals. *Biomed. Signal Process. Control* 34, 81–92. doi: 10.1016/j.bspc.2017.01.005

Jia, G., Lam, H. K., and Althoefer, K. (2022). Variable weight algorithm for convolutional neural networks and its applications to classification of seizure phases and types. *Pattern Recogn.* 121, 108226. doi: 10.1016/j.patcog.2021.108226

Jiang, Y., Chen, W., and Li, M. (2020). Symplectic geometry decomposition-based features for automatic epileptic seizure detection. *Comput. Biol. Med.* 116, 103549. doi: 10.1016/j.compbiomed.2019.103549

Jolliffe, I. T., and Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* 374, 20150202. doi: 10.1098/rsta.2015.0202

Kramer, O. (2013). "K-nearest neighbors," in *Dimensionality Reduction With Unsupervised Nearest Neighbors*, eds J. Kacprzyk, and L. C. Jain (Berlin; Heidelberg: Springer), 13–23. doi: 10.1007/978-3-642-38652-7_2

Lau, K. W., and Wu, Q. H. (2003). Online training of support vector classifier. *Pattern Recogn.* 36, 1913–1920. doi: 10.1016/S0031-3203(03)00038-4

Li, C., Zhou, W., Liu, G., Zhang, Y., Geng, M., Liu, Z., et al. (2021). Seizure onset detection using empirical mode decomposition and common spatial pattern. *IEEE Trans. Neural Syst. Rehabil. Eng.* 29, 458–467. doi: 10.1109/TNSRE.2021.3055276

Li, M., and Chen, W. (2021). FFT-based deep feature learning method for EEG classification. *Biomed. Signal Process. Control* 66, 102492. doi: 10.1016/j.bspc.2021.102492

Li, Y., Cui, W. G., Huang, H., Guo, Y. Z., Li, K., and Tan, T. (2019). Epileptic seizure detection in EEG signals using sparse multiscale radial basis function networks and the Fisher vector approach. *Knowledge Based Syst.* 164, 96–106. doi: 10.1016/j.knosys.2018.10.029

Martis, R. J., Tan, J. H., Chua, C. K., Loon, T. C., Yeo, S. W. J., and Tong, L. (2015). Epileptic EEG classification using nonlinear parameters on different frequency bands. *J. Mech. Med. Biol.* 15, 1550040. doi: 10.1142/S0219519415500402

Murat, F., Yildirim, O., Talo, M., Demir, Y., Tan, R. S., Ciaccio, E. J., et al. (2021). Exploring deep features and ECG attributes to detect cardiac rhythm classes. *Knowledge Based Syst.* 232, 107473. doi: 10.1016/j.knosys.2021.107473

Mursalin, M., Zhang, Y., Chen, Y., and Chawla, N. V. (2017). Automated epileptic seizure detection using improved correlation-based feature selection with random forest classifier. *Neurocomputing* 241, 204–214. doi: 10.1016/j.neucom.2017.02.053

Murtagh, F. (1991). Multilayer perceptrons for classification and regression. *Neurocomputing* 2, 183–197. doi: 10.1016/0925-2312(91)90023-5

Natekin, A., and Knoll, A. (2013). Gradient boosting machines, a tutorial. *Front. Neurorobot.* 7, 21. doi: 10.3389/fnbot.2013.00021

Peker, M., Sen, B., and Delen, D. (2015). A novel method for automated diagnosis of epilepsy using complex-valued classifiers. *IEEE J. Biomed. Health Inform.* 20, 108–118. doi: 10.1109/JBHI.2014.2387795

Peng, H., Lei, C., Zheng, S., Zhao, C., Wu, C., Sun, J., et al. (2021). Automatic epileptic seizure detection via Stein kernel-based sparse representation. *Comput. Biol. Med.* 132, 104338. doi: 10.1016/j.compbiomed.2021.104338

Prabhakar, S. K., and Lee, S. W. (2022). ENIC: ensemble and nature inclined classification with sparse depiction based deep and transfer learning for biosignal classification. *Appl. Soft Comput.* 117, 108416. doi: 10.1016/j.asoc.2022.108416

Raghu, S., Sriraam, N., Hegde, A. S., and Kubben, P. L. (2019). A novel approach for classification of epileptic seizures using matrix determinant. *Expert Syst. Appl.* 127, 323–341. doi: 10.1016/j.eswa.2019.03.021

Rout, S. K., Sahani, M., Dash, P. K., and Biswal, P. K. (2021). Multifuse multilayer multikernel RVFLN+ of process modes decomposition and approximate entropy data from iEEG/sEEG signals for epileptic seizure recognition. *Comput. Biol. Med.* 132, 104299. doi: 10.1016/j.compbiomed.2021.104299

Safavian, S. R., and Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE Trans. Syst. Man Cybernet.* 21, 660–674. doi: 10.1109/21.97458

Sahani, M., Rout, S. K., and Dash, P. K. (2021). FPGA implementation of epileptic seizure detection using semisupervised reduced deep convolutional neural network. *Appl. Soft Comput.* 110, 107639. doi: 10.1016/j.asoc.2021.107639

Shankar, A., Khaing, H. K., Dandapat, S., and Barma, S. (2021). Analysis of epileptic seizures based on EEG using recurrence plot images and deep learning. *Biomed. Signal Process. Control* 69, 102854. doi: 10.1016/j.bspc.2021.102854

Sharma, M., Pachori, R. B., and Acharya, U. R. (2017). A new approach to characterize epileptic seizures using analytic time-frequency flexible wavelet transform and fractal dimension. *Pattern Recogn. Lett.* 94, 172–179. doi: 10.1016/j.patrec.2017.03.023

Sharma, R., Pachori, R. B., and Sircar, P. (2020). Seizures classification based on higher order statistics and deep neural network. *Biomed. Signal Process. Control* 59, 101921. doi: 10.1016/j.bspc.2020.101921

Sharmila, A., and Geethanjali, P. (2019). A review on the pattern detection methods for epilepsy seizure detection from EEG signals. *Biomed. Eng.* 64, 507–517. doi: 10.1515/bmt-2017-0233

Shoeibi, A., Ghassemi, N., Alizadehsani, R., Rouhani, M., Hosseini-Nejad, H., Khosravi, A., et al. (2021). A comprehensive comparison of handcrafted features and convolutional autoencoders for epileptic seizures detection in EEG signals. *Expert Syst. Appl.* 163, 113788. doi: 10.1016/j.eswa.2020.113788

Shoeibi, A., Ghassemi, N., Khodatars, M., Moridian, P., Alizadehsani, R., Zare, A., et al. (2022). Detection of epileptic seizures on EEG signals using ANFIS classifier, autoencoders and fuzzy entropies. *Biomed. Signal Process. Control* 73, 103417. doi: 10.1016/j.bspc.2021.103417

Solaija, M. S. J., Saleem, S., Khurshid, K., Hassan, S. A., and Kamboh, A. M. (2018). Dynamic mode decomposition based epileptic seizure detection from scalp EEG. *IEEE Access* 6, 38683–38692. doi: 10.1109/ACCESS.2018.2853125

Subasi, A., Kevric, J., and Canbaz, M. A. (2019). Epileptic seizure detection using hybrid machine learning methods. *Neural Comput. Appl.* 31, 317–325. doi: 10.1007/s00521-017-3003-y

Sukriti, Chakraborty M., and Mitra, D. (2021). A novel automated seizure detection system from EMD-MSPCA denoised EEG: refined composite multiscale sample, fuzzy and permutation entropies based scheme. *Biomed. Signal Process. Control* 67, 102514. doi: 10.1016/j.bspc.2021.102514

Supriya, S., Siuly, S., Wang, H., and Zhang, Y. (2021). New feature extraction for automated detection of epileptic seizure using complex network framework. *Appl. Acoust.* 180, 108098. doi: 10.1016/j.apacoust.2021.108098

Swami, P., Gandhi, T. K., Panigrahi, B. K., Tripathi, M., and Anand, S. (2016). A novel robust diagnostic model to detect seizures in electroencephalography. *Expert Syst. Appl.* 56, 116–130. doi: 10.1016/j.eswa.2016.02.040

Tajmirriahi, M., and Amini, Z. (2021). Modeling of seizure and seizure-free EEG signals based on stochastic differential equations. *Chaos Solitons Fractals* 150, 111104. doi: 10.1016/j.chaos.2021.111104

Tuncer, T., Dogan, S., and Acharya, U. R. (2021a). Automated EEG signal classification using chaotic local binary pattern. *Expert Syst. Appl.* 182, 115175. doi: 10.1016/j.eswa.2021.115175

Tuncer, T., Dogan, S., and Akbal, E. (2019). A novel local senary pattern based epilepsy diagnosis system using EEG signals. *Austral. Phys. Eng. Sci. Med.* 42, 939–948. doi: 10.1007/s13246-019-00794-x

Tuncer, T., Dogan, S., and Naik, G. R., and Plawiak P. (2021b). Epilepsy attacks recognition based on 1D octal pattern, wavelet transform and EEG signals. *Multimedia Tools Appl.* 80, 25197–25218. doi: 10.1007/s11042-021-10882-4

Wang, Q., Wei, H. L., Wang, L., and Xu, S. (2021). A novel time-varying modeling and signal processing approach for epileptic seizure detection and classification. *Neural Comput. Appl.* 33, 5525–5541. doi: 10.1007/s00521-020-05330-7

Wang, Z., Wu, D., Dong, F., Cao, J., Jiang, T., and Liu, J. (2020). A novel spike detection algorithm based on multi-channel of BECT EEG signals. *IEEE Trans. Circ. Syst. II Exp. Briefs* 67, 3592–3596. doi: 10.1109/TCSII.2020.2992285

Wu, M., Wan, T., Wan, X., Fang, Z., and Du, Y. (2021). A new localization method for epileptic seizure onset zones based on time-frequency and clustering analysis. *Pattern Recogn.* 111, 107687. doi: 10.1016/j.patcog.2020.107687

Ye, J., Janardan, R., and Li, Q. (2004). "Two-dimensional linear discriminant analysis," in *Advances in Neural Information Processing Systems 17*, eds L. Saul, Y. Weiss, and L. Bottou (British Columbia: MIT Press), 1569–1576.

Zahra, A., Kanwal, N., ur Rehman, N., Ehsan, S., and McDonald-Maier, K. D. (2017). Seizure detection from EEG signals using multivariate empirical mode decomposition. *Comput. Biol. Med.* 88, 132–141. doi: 10.1016/j.compbiomed.2017.07.010

Zarei, A., and Asl, B. M. (2021). Automatic seizure detection using orthogonal matching pursuit, discrete wavelet transform, and entropy based features of EEG signals. *Comput. Biol. Med.* 131, 104250. doi: 10.1016/j.compbiomed.2021.104250

Zhang, T., Han, Z., Chen, X., and Chen, W. (2021). Subbands and cumulative sum of subbands based nonlinear features enhance the performance of epileptic seizure detection. *Biomed. Signal Process. Control* 69, 102827. doi: 10.1016/j.bspc.2021.102827

Zhou, X., Ling, B. W. K., Li, C., and Zhao, K. (2020). Epileptic seizure detection via logarithmic normalized functional values of singular values. *Biomed. Signal Process. Control* 62, 102086. doi: 10.1016/j.bspc.2020.102086

TYPE Original Research
PUBLISHED 15 June 2023
DOI 10.3389/fnins.2023.1212049

Check for updates

OPEN ACCESS

EDITED BY
Zhengwang Wu,
University of North Carolina at Chapel Hill,
United States

REVIEWED BY
Ling Ma,
Nankai University, China
Hailong Huang,
Hong Kong Polytechnic University,
Hong Kong SAR, China

*CORRESPONDENCE
Shaoyi Du
✉ dushaoyi@gmail.com
Meifeng Xu
✉ xumf96@163.com

†These authors have contributed equally to this work

RECEIVED 25 April 2023
ACCEPTED 31 May 2023
PUBLISHED 15 June 2023

CITATION
Wang X, Han H, Xu M, Li S, Zhang D, Du S and Xu M (2023) STNet: shape and texture joint learning through two-stream network for knowledge-guided image recognition. *Front. Neurosci.* 17:1212049. doi: 10.3389/fnins.2023.1212049

STNet: shape and texture joint learning through two-stream network for knowledge-guided image recognition

Xijing Wang[1†], Hongcheng Han[1†], Mengrui Xu[1,2], Shengpeng Li[1], Dong Zhang[1,3], Shaoyi Du[1]* and Meifeng Xu[4]*

[1]National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, National Engineering Research Center for Visual Information and Applications, Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, China, [2]The School of Software Engineering, Xi'an Jiaotong University, Xi'an, China, [3]The School of Automation Science and Engineering, Xi'an Jiaotong University, Xi'an, China, [4]The Second Affiliated Hospital of Xi'an Jiaotong University (Xibei Hospital), Xi'an, China

Introduction: The human brain processes shape and texture information separately through different neurons in the visual system. In intelligent computer-aided imaging diagnosis, pre-trained feature extractors are commonly used in various medical image recognition methods, common pre-training datasets such as ImageNet tend to improve the texture representation of the model but make it ignore many shape features. Weak shape feature representation is disadvantageous for some tasks that focus on shape features in medical image analysis.

Methods: Inspired by the function of neurons in the human brain, in this paper, we proposed a shape-and-texture-biased two-stream network to enhance the shape feature representation in knowledge-guided medical image analysis. First, the two-stream network shape-biased stream and a texture-biased stream are constructed through classification and segmentation multi-task joint learning. Second, we propose pyramid-grouped convolution to enhance the texture feature representation and introduce deformable convolution to enhance the shape feature extraction. Third, we used a channel-attention-based feature selection module in shape and texture feature fusion to focus on the key features and eliminate information redundancy caused by feature fusion. Finally, aiming at the problem of model optimization difficulty caused by the imbalance in the number of benign and malignant samples in medical images, an asymmetric loss function was introduced to improve the robustness of the model.

Results and conclusion: We applied our method to the melanoma recognition task on ISIC-2019 and XJTU-MM datasets, which focus on both the texture and shape of the lesions. The experimental results on dermoscopic image recognition and pathological image recognition datasets show the proposed method outperforms the compared algorithms and prove the effectiveness of our method.

KEYWORDS

computer-aided diagnosis, image recognition, feature fusion, joint learning, two-stream network, brain-like information processing

1. Introduction

Computer-aided diagnosis (CAD) has been a research hotspot for the past few decades. CAD automatically analyzes the patient data through machine learning methods to make an assessment of the patient's condition (Yanase and Triantaphyllou, 2019; Chan et al., 2020). Medical image analysis is one of the most important fields in CAD technologies, it helps read imaging data to improve the diagnosis efficiency. An intelligent medical image analysis model can share the workload of radiologists and pathologists, and enables areas with underdeveloped medical resources to achieve high-level imaging analysis at low cost (Shen et al., 2017; Kurc et al., 2020).

In the past decade, medical image analysis methods have grown by leaps and bounds due to the development of deep learning and computer vision algorithms. Powerful feature representation ability enables deep neural networks to learn complex hidden features from a large amount of training data, which overcomes the difficulty of manual feature design in traditional medical image analysis methods. However, there are still challenges to be addressed in current deep learning-based algorithms for medical image analysis, with weak shape representation being one of the most critical issues. On the one hand, in the commonly used convolutional neural network (CNN), the limited receptive field of kernels tends to fit local features during kernel parameter learning. Although the range of the receptive field of deep convolutional kernels on original images gradually increases as layers deepen, deeper layers weaken their connection with original images, which limits networks in modeling shape features at larger scales (Luo et al., 2016; Araujo et al., 2019). On the other hand, pre-trained parameters are frequently employed in medical image recognition techniques to expedite convergence during training and potentially enhance model performance. Given the paucity of annotated data in medical images, large-scale natural image datasets such as ImageNet (Deng et al., 2009; Russakovsky et al., 2015) are commonly utilized as pre-training datasets. However, the research of Geirhos et al. (2018) indicates that the deep neural network pre-trained on ImageNet is biased to focus on the texture features and has relatively weak shape feature representation ability.

The weak representation of shapes, caused by the limitations of the model and pre-training datasets, significantly impacts the performance of the model on certain shape-dependent medical image tasks. As, Figure 1 shows, cascade segmentation and classification model (Chang, 2017) can solve the problem in some scenarios, it uses a segmentation network to obtain the mask of a lesion, and then use the segmented lesion image as the input of the classification network, providing shape information for classification, eliminating the background noise. However, the lack of sufficient training data is a prevalent issue in various medical image analysis tasks, resulting in inadequate precision of the trained segmentation task. Inaccurate segmentation can provide erroneous shape information for classification. In addition, the cascade segmentation and classification model contains two encoders and one decoder, and they are cascaded, the research of He et al. (2017) indicates that repetitive encoding and decoding operations yield minimal improvements to the quality of extracted features.

In order to solve the above problems, we proposed a shape-and-texture-biased two-stream network to enhance the shape feature representation in knowledge-guided medical image analysis. The human brain processes shape and texture information separately through different neurons in the visual system, inspired by that, first, the two-stream network shape-biased stream and a texture-biased stream are constructed through classification and segmentation multi-task joint learning. Second, we propose pyramid-grouped convolution (PGC) to enhance the texture feature representation, and introduce deformable convolution (DC) to enhance the shape feature extraction. Third, we used a channel-attention-based feature selection module in shape and texture feature fusion to focus on the key features and eliminate information redundancy caused by feature fusion. Finally, aiming at the problem of model optimization difficulty caused by the imbalance in the number of benign and malignant samples in medical images, an asymmetric loss function was introduced to improve the robustness of the model. We applied our method to the melanoma recognition task on ISIC-2019 (Rotemberg et al., 2021) and XJTU-MM datasets, which focuses on both the texture and shape of the lesions. The experimental results on dermatoscopic image recognition and pathological image recognition show that the proposed method outperforms the compared algorithms and prove the effectiveness of our method.

The main contributions of this work are enumerated as follows:

- We propose the shape and texture joint learning two-stream network for knowledge-guided medical image recognition, taking into account the learning of shape features and texture features by the network, addressing the weak shape representation problem of existed methods.
- We propose pyramid-grouped convolution to enhance the texture feature representation, and introduce deformable convolution to address the limitation of fixed respective fields, enhancing the shape feature extraction.
- We construct the shape and texture fusion module based on channel attention mechanism to focus on the essential features and eliminate the noise, reducing the information redundancy caused by feature fusion.
- We introduce the asymmetric loss function for optimization, reducing the impact of commonly existed sample imbalance problem in medical image datasets.

2. Related work

2.1. Knowledge-guided medical image analysis

Most of the key technologies in medical image analysis come from general computer vision algorithms, however, the image characteristics and the data distribution are different between natural images and medical images. Constructing appropriate deep neural network model with the guidance of the prior knowledge from pathology and radiology is important for improving model performance in specific medical analysis tasks.

Fan et al. (2017) proposed a novel automatic segmentation algorithm using saliency combined with Otsu threshold for dermoscopy images, which extracted prior information on healthy skin to construct the color saliency map and brightness saliency map respectively. Ahn et al. (2017) proposed a saliency-based lesion segmentation method in dermoscopic images, using the reconstruction errors derived from a sparse representation model coupled with a novel background detection. Yang et al. (2023) proposed a Multi-scale Fully-shared Fusion Network (MFF-Net) that gathers features of dermoscopic images and clinical images for skin lesion classification. Zhang et al. (2018a) used deep learning algorithms to help diagnose four common cutaneous diseases based on dermoscopic images and summarized classification/diagnosis scenarios based on domain expert knowledge and semantically represented them in a hierarchical structure to improve the accuracy of the algorithm. Clinical prior knowledge is also widely applied to the analysis of ultrasound images and other medical images. Liu et al. (2019b) proposed a novel deep-learning-based CAD system, guided by task-specific prior knowledge, for automated nodule detection and classification in ultrasound images. Chen et al. (2021) proposed a knowledge-guided data augmentation framework for breast lesion classification, which consists of a modal translater and a semantic inverter, achieving cross-modal and semantic data augmentation simultaneously. Shi et al. (2020) proposed a knowledge-guided synthetic medical image adversarial augmentation method for ultrasonography thyroid nodule classification, extracting domain knowledge from standardized terminology to improve the classification performance. Yang et al. (2021) proposed a multi-task cascade deep learning model (MCDLM), which integrates radiologists' various domain knowledge (DK) and used multimodal ultrasound images for automatic diagnosis of thyroid nodules. Han et al. (2020) proposed an ensemble learning method for panoramic radiographs recognition based on the characteristics of each stage of tooth growth. Ni et al. (2013) proposed a novel learning-based automatic method to detect the fetal head for the measurement of head circumference from ultrasound images and used prior knowledge and online imaging parameters to guide the sliding window-based head detection. Pan et al. (2022) proposed a two-stage network with prior knowledge guidance for medullary thyroid carcinoma recognition in ultrasound images. Meanwhile, extracting and fusing semantic features of solid tissues and calcification for better recognizing the segmented nodules. Zhou et al. (2022) proposed a rheumatoid arthritis knowledge-guided (RATING) system for scoring rheumatoid arthritis activity from multimodal ultrasound images, leveraging diagnostic paradigm and experience to enhance the robustness. Lu et al. (2023) proposed a Prior Knowledge-based Relation Transformer Network (PKRT-Net), which employed the clinical prior knowledge to assist OC segmentation. Gao et al. (2021) proposed a medical-knowledge-guided one-class classification approach that leverages domain-specific knowledge of classification tasks to boost the model's performance and showed superior model performance on three different clinical image classification tasks. Zhang et al. (2023) proposed coarse-to-fine method for melanoma and nevi recognition according to distribution of inter-class and intra-class differences as summarized by dermatologists.

Prior knowledge provides inspiration for medical image analysis design, in this paper, we innovate a novel method for shape-relied medical image recognition.

2.2. Shape and texture feature fusion

Aiming at the problem of weak shape representation of existing CNN-based medical image recognition models, we investigate

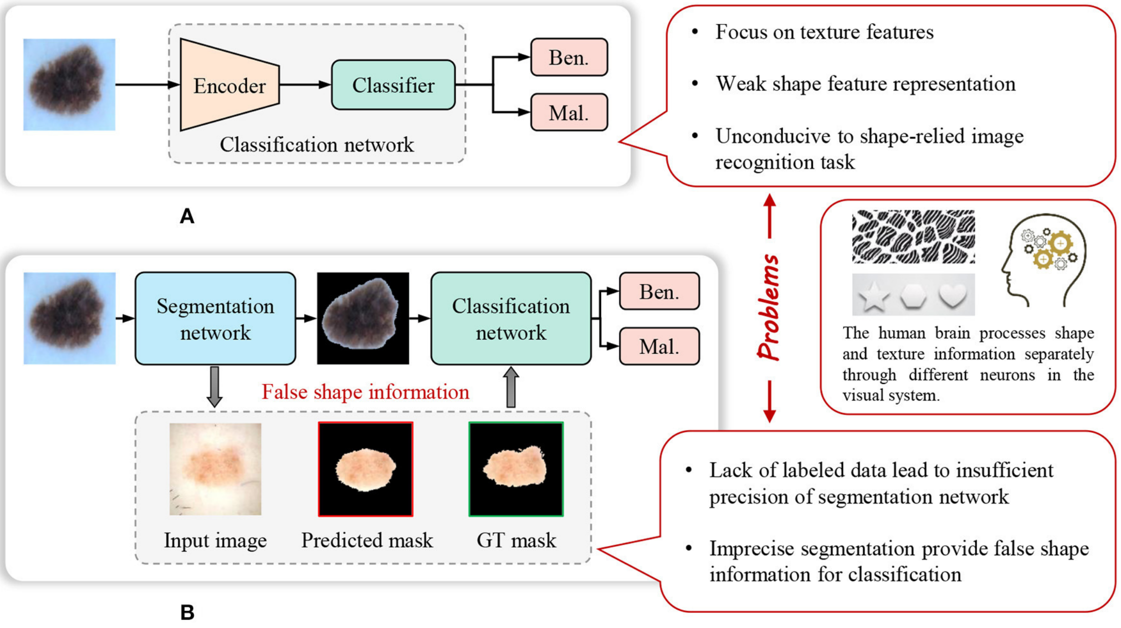

Encoder
Classifier
Classification network
Ben.
Mal.
A
Focus on texture features
Weak shape feature representation
Unconducive to shape-relied image recognition task
Problems
The human brain processes shape and texture information separately through different neurons in the visual system.
Segmentation network
Classification network
Ben.
Mal.
False shape information
Input image
Predicted mask
GT mask
B
Lack of labeled data lead to insufficient precision of segmentation network
Imprecise segmentation provide false shape information for classification

FIGURE 1
Weak feature representation problem of many existing methods for image recognition in computer-aided diagnosis. **(A)** Common image recognition model. **(B)** Cascade segmentation and classification model.

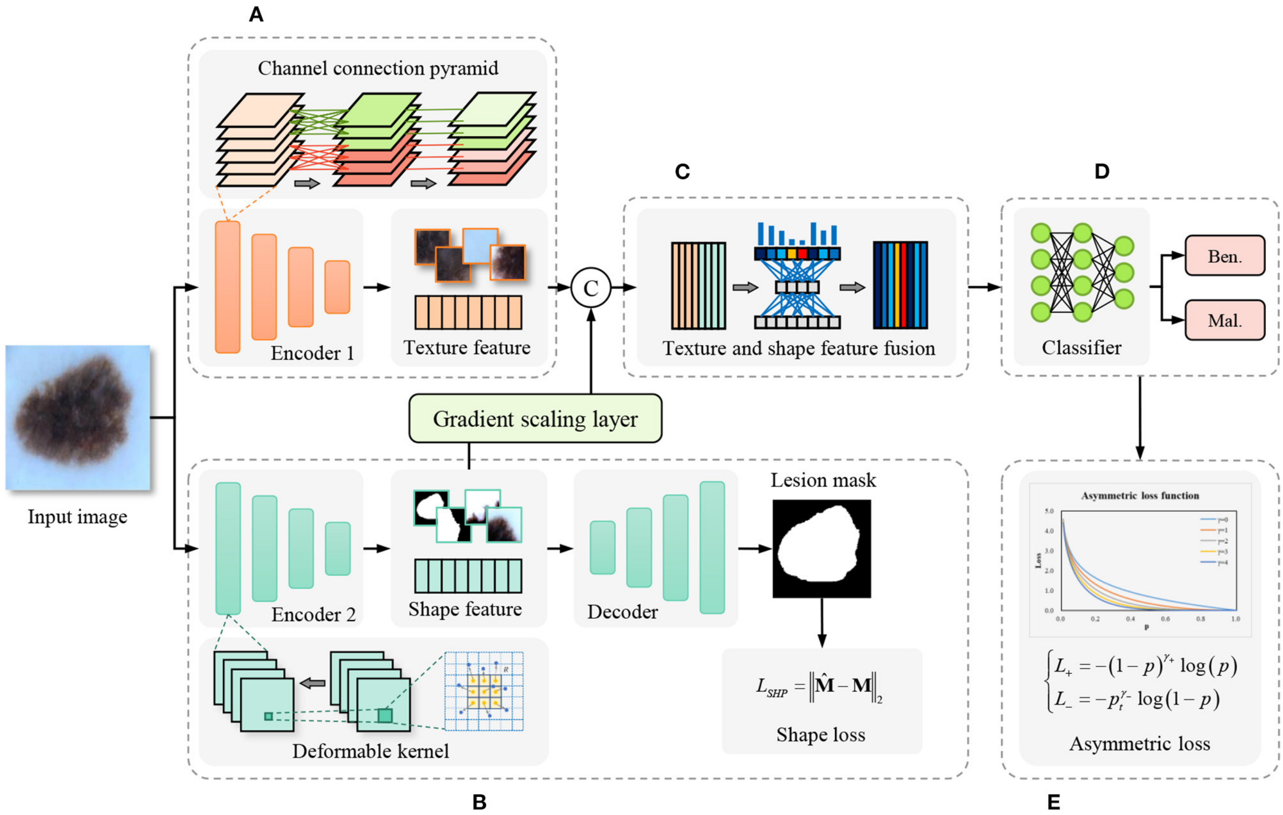


FIGURE 2
Framework of the proposed shape and texture joint learning two-stream network. **(A)** Texture-biased stream. **(B)** Shape-biased stream. **(C)** Feature fusion module. **(D)** Classifier. **(E)** Asymmetric loss.

the texture and shape feature fusion algorithms designed for various tasks.

Al-Osaimi et al. (2011) proposed spatially optimized data/pixel-level fusion of 3-D shape and texture for face recognition. Lu et al. (2017) proposed a face image retrieval method based on shape and texture feature fusion, which used accurate facial landmark locations as shape features and utilized shape priors to provide discriminative texture features. Kotsia et al. (2008) proposed a novel method based on the fusion of texture and shape information for facial expression and Facial Action Unit (FAU) recognition from video sequences and used various approaches to perform texture and shape feature fusion, among which were SVMs and Median Radial Basis Functions (MRBFs). Anantharatnasamy et al. (2013) proposed a content-based image retrieval system based on three major types of visual information including color, texture, shape, and their distances to the origin in a three dimensional space for the retrieval. Sumathi and Kumar (2012) extracted edge and texture features using Gabor filter and fused them for plant leaf classification. Xiong et al. (2007) proposed a Statistical Shape and Radio texture fusion model for facial expression sequence synthesis, processing facial shape and texture separately and fusing them together to synthesize the final result. Jo et al. (2014) proposed a new method for eye state classification to detect diver drowsiness, which extracted and fused features from both eyes. Zhang et al. (2020) proposed two-stream networks to enhance the extraction of shape and texture respectively for clothing classification and attribute recognition.

These researches use various of methods to enhance the texture and shape feature learning on specific data. For shape-relied medical image recognition tasks, we design the model to realize that with the guidance of the prior knowledge, such as visual characteristics and category distribution.

3. Methodology

3.1. Framework

In contrast to the cascade segmentation and classification model, our proposed model employs a two-stream network for joint learning of shape and texture, mitigating the impact of imprecise segmentation on shape information in the former. The overall framework of the proposed method is shown as Figure 2, the input image is fed into the parallel texture-biased stream and shape-biased stream. First, the texture-biased stream consists of a feature encoder, which is pre-trained on texture-biased large-scale dataset, such as ImageNet. To further enhance the texture feature representation ability of the texture feature encoder, we reconstruct the convolutional block using the proposed channel connection pyramid mechanism. Second, the shape-biased stream

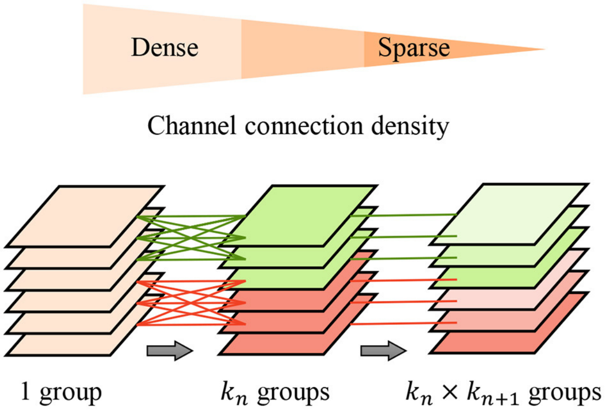


FIGURE 3
Pyramid-grouped convolution. In each pyramid, the density of channel connection changes layer by layer, and from dense to sparse.

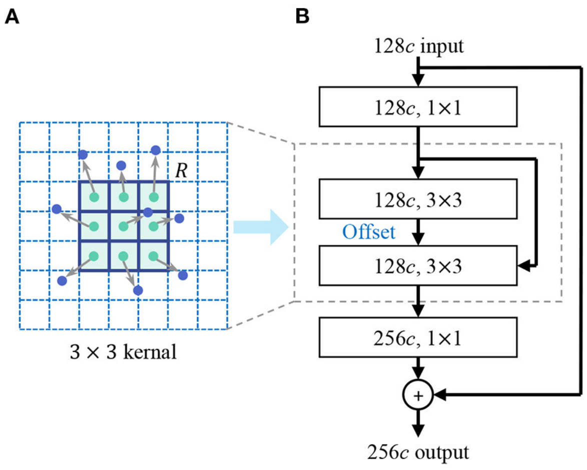


FIGURE 4
Deformable convolution. **(A)** Deformable kernel. **(B)** Deformable convolutional layer. An offset layer is inserted to learn the offset to transform the rectangular kernel to a kernel with an irregular shape that better match the extracted features. The feature map in the deformable receptive field is resampled through bilinear interpolation according to the parameters of the learned offset.

contains an encoder-decoder based network, the encoder extracts the shape features and the decoder generates the lesion mask, the quality of the extracted shape features is supervised by *L2* loss function between the predicted mask and the ground truth mask. Third, the texture feature and the shape feature are concatenated and input to the feature fusion module, to address the information redundancy problem in feature fusion, we construct the feature fusion module based on channel attention mechanism to focus on the essential features and eliminate the effects of noise. In addition, to balance the texture-biased learning and shape-biased learning, the gradient scaling layer is added between the shape feature map and the concatenation operation to weight the gradient in the back propagation. Then, the fully connected layer classifier is used to output the classification results. Finally, to overcome the optimization difficulty caused by the problem of imbalanced samples in medical image datasets, we introduce the asymmetric loss to enhance the attention of the model to the categories with smaller numbers of samples.

3.2. Texture-biased stream

The texture-biased stream is constructed by the texture feature encoder pre-trained on texture-biased dataset ImageNet. To enhance the texture feature representation, we improve the channel connections in convolutional blocks. In the standard convolution operation, each kernel is connected to every channel of the input feature map. However, while the large number of learnable parameters provides a powerful fitting ability for the network, overly dense connections can lead to significant information redundancy and unnecessary computational burden (Huang et al., 2017; Ma et al., 2018; Zhang et al., 2018b). Grouped convolution mechanism (Xie et al., 2017; Zhang H. et al., 2022) provides an efficient way to solve the problem, it divides the input feature map into several groups in the channel dimension, each kernel has connections to the specific group only rather than all channels of the input feature map. With the same number of output feature map channels, channel-wise connections become sparser, thereby enhancing diagonal correlations between channels. Depth-wise convolution (Chollet, 2017) even makes the connections more sparse, which regards each channel of the input feature map as one group to perform grouped convolution. With fewer learnable kernel parameters, depth-wise convolution even shows stronger low-level texture feature representation ability (Guo et al., 2019; Tan and Le, 2019). However, grouped convolution and depth-wise convolution still have problems in balancing the learning of low-level and high-level texture features.

To further improve the feature extraction quality and efficiency, we propose the pyramid-grouped convolution(PGC) mechanism to enhance the feature representation of the texture-biased stream. As Figure 3 shows, In each pyramid-convolutional block, the density of channel connections varies layer by layer, transitioning from dense to sparse. This results in a transition of the channel-wise receptive field of each kernel from large to small, leading to sparser feature encoding compared to conventional grouped convolution and more appropriate channel-wise receptive fields than depth-wise convolution. The PGC blocks are embedded in the backbone network to construct feature encoder of texture-biased stream, enhancing the texture feature representation.

3.3. Shape-biased stream

Pixel-wise semantic segmentation model is a learning paradigm conducive to modeling shape features (Long et al., 2015; Guo et al., 2018). In the proposed method, the shape-biased stream is constructed using an encoder-decoder based segmentation network, the decoder generates the lesion mask based on the features extracted from the input image. With the supervision of the *L2* loss between the predicted mask and the ground truth mask, the encoder is encouraged to learn the shape-biased features. Many encoder-decoder based semantic segmentation models add shortcut connections between encoders and decoders to enhance the contributions of low-level features extracted by shallow layers in encoders to mask generation, which are usually called U-shape networks (Ronneberger et al., 2015; Oktay et al., 2018; Zhou et al., 2018; Zhang et al., 2021). But in the shape-biased stream of our method, all we need is to improve the shape feature representation of the feature map extracted by feature encoder, all the information flow is expected to pass through the deepest feature map, so we did not add any shortcut connection between the encoder and the decoder.

In the design of the shape encoder network, we introduce the deformable kernel to address the limitation of the rectangular receptive field of the convolution kernel. Irregular-shaped visual features are common in lesion images, for example, the irregular-shape boundary of the lesion in dermoscopic images (Celebi et al., 2019), the irregular-shaped cells in pathological images (Zhang D. et al., 2022). Rectangular convolutional kernels have limitation in extracting these features, especially in extracting low-level shape features. As Figure 4 shows, the discrete feature map is regarded as a continuous two-dimensional distribution, we insert an offset layer to learn a offset to transform the rectangular kernel to an kernel with irregular shape that better match the extracted features. The feature map in the deformable receptive field is resampled through bilinear interpolation according to the parameters of the learned offset. deformable convolution is calculated by

$$y(p) = \sum_{p_k \in \boldsymbol{R}} \boldsymbol{w}(p_k) \cdot \boldsymbol{x}(p + p_k + \Delta p_k) \ , \tag{1}$$

where $\boldsymbol{y}(p)$ indicates the feature obtained by the convolution on one sampling point p of the feature map. $\boldsymbol{R}$ is the receptive field size of the regular kernel. p_k donates the difference between the sampling points and $\boldsymbol{y}(p)$, $k = 1, 2, 3...N, N = |\boldsymbol{R}|$, Δp_k is the learned offset, and $\boldsymbol{w}$ is the kernel parameter. We reconstruct the backbone network of feature encoder using deformable convolution layers, enhancing the representation of irregular-shaped features.

3.4. Channel-attention-based texture and shape feature fusion

The feature maps extracted from the texture-biased and shape-biased streams are concatenated to fuse texture and shape features, which expands the scope of the extracted features. However, this also results in a certain degree of information redundancy. Some irrelevant features not only fail to contribute to improving model performance but also increase the risk of overfitting and negatively impact model robustness. To select essential features for lesion recognition and eliminate irrelevant features and noise, we design the texture and shape feature fusion module based on channel attention mechanism.

Each kernel represents a specific hidden feature, having a specific correlation with lesion recognition, feature selection is equivalent to kernel selection, which can also be regarded as the selection of channels of feature map. We introduce the channel attention mechanism to highlight the essential channels and suppress noise through learning the channel weights based on the global representation of each channel. As Figure 5 shows, for the $w \times h \times c$ input feature map $\mathbf{Z}$, it is first transformed into a $1 \times 1 \times c$ feature vector $\boldsymbol{g}$ through global pooling, which combines average pooling and max pooling to balance average and peak characterization, calculating by

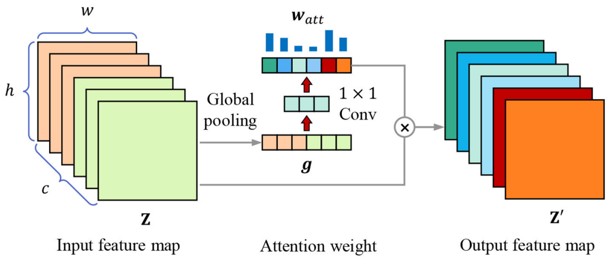


FIGURE 5
Channel attention mechanism. The attention weight vector w_{att} is calculated through global pooling and 1 × 1 convolutional layers, then the input feature map Z is weighted to obtain the output feature map Z′.

$$g_k = \frac{1}{2}\left(\frac{1}{wh}\sum_{i=1}^{h}\sum_{j=1}^{w} z_{i,j,k} + \max_{i,j}(z_{i,j,k})\right), \tag{2}$$

where g_k is the element in feature vector $\boldsymbol{g}$, $z_{i,j,k}$ is the element in k-th channel of feature map $\mathbf{Z}$. Then we use two 1×1 convolutional layers to obtain the attention weight of each channel, calculating through

$$\boldsymbol{w}_{att} = \delta\left(\boldsymbol{w}_{Conv2}^{T} \cdot \delta\left(\boldsymbol{w}_{Conv1}^{T} \cdot \boldsymbol{g}\right)\right), \tag{3}$$

where $\boldsymbol{w}_{Conv1}$ and $\boldsymbol{w}_{Conv2}$ are the weight parameters of two 1×1 convolutional layers, $\delta(\cdot)$ is the sigmoid activation function. Finally, the original input feature map is weighted by the weight vector,

$$\mathbf{Z}' = \boldsymbol{w}_{att} \otimes \mathbf{Z}, \tag{4}$$

where $\otimes$ means to multiply $\boldsymbol{w}_{att}$ and $\mathbf{Z}$ channel by channel.

In optimization, the channels that are highly relevant to lesion recognition are highlighted, which eliminates the information redundancy caused by the feature fusion of texture-biased stream and shape-biased stream, and selects the features conductive to lesion recognition, improving the robustness of the model.

3.5. Joint learning loss function and optimization

Due to the characteristics of the disease, training data often contains more benign lesions than malignant ones, resulting in insufficient attention given to malignant samples during network training and negatively impacting model optimization (Liu et al., 2019a) and (Xu et al., 2020). If the number of benign samples is forcibly reduced to balance the number of benign and malignant samples, it will lead to insufficient training data.

To address the problem of sample imbalance, we design the asymmetric loss function for medical image recognition with a large amount of negative samples and few positive samples. Different from the commonly used cross-entropy loss shown in Equation (5),

$$\mathcal{L}_{CE} = -y\log(p) - (1-y)\log(1-p), \tag{5}$$

where $y_{\in}\{0,1\}$ means the ground truth label of the sample, $p \in (0,1)$ is the predicted score, when $p > 0.5$, the sample is predicted as the positive category, the asymmetric loss decouples the loss of positive and negative categories, reducing the impact of sample imbalance through asymmetric focusing and asymmetric probability transfer, for each sample, the new loss function for classification $\mathcal{L}_{CLS}$ is calculated through

$$\mathcal{L}_{CLS} = -y(1-p)^{\gamma_+}\log(p) - (1-y)p^{\gamma_-}\log(1-p), \tag{6}$$

where γ_+ and γ_- are the exponential decay factors, the larger the value of the decay factor, the greater the attenuation effect. The adaptive weight factors $(1-p)^{\gamma_+}$ and p^{γ_-} are added to original cross-entropy loss function to asymmetrically scale the loss of positive samples and negative samples, which is better for the optimization in the case of unbalanced samples. We set $\gamma_+ < \gamma_-$ to reduce the gradient of the negative samples, strengthening the attention of the model optimization to the positive samples.

In addition, with typical characteristics, some negative samples are easy to identify, to constrain the model to focus on hard samples, we add the probability transfer to the loss function, directly discarding samples which have a low predicted p value. The weight factor of $\mathcal{L}_-$ is reconstructed with the transfer probability p_t, which is calculated by

$$p_t = \max(p - \varphi, 0), \tag{7}$$

where φ is the probability cutoff threshold, when the predicted p is lower than μ, p_t is set to 0. The final asymmetric classification loss function is

$$\mathcal{L}_{CLS} = -y(1-p)^{\gamma_+}\log(p) - (1-y)p_t^{\gamma_-}\log(1-p), \tag{8}$$

which enables the model to overcome the imbalance of samples in training, and focus on the difficult samples near the discrimination interface, enhancing the robustness of the trained model.

In the optimization of the shape-biased stream, we use $L2$ loss, which is the pixel-wise mean square error between the predicted mask $\hat{\mathbf{M}}$ and the ground truth mask $\mathbf{M}$, the shape loss $\mathcal{L}_{SHP}$ is

$$\mathcal{L}_{SHP} = \|\hat{\mathbf{M}} - \mathbf{M}\|_2, \tag{9}$$

In joint learning, texture feature encoder parameter $\boldsymbol{\theta}^*_{TE}$ is supervised by L_{CLS}, shape feature decoder parameter $\boldsymbol{\theta}^*_{SD}$ is supervised by L_{SHP}, shape feature encoder parameter $\boldsymbol{\theta}^*_{SE}$ is supervised by L_{CLS} and L_{SHP} to encourage learning shape features that are conductive to lesion classification. In summary, they are optimized by

$$\boldsymbol{\theta}^*_{TE} = \underset{\boldsymbol{\theta}_{TE}}{\arg\min}\,\mathcal{L}_{CLS} \tag{10}$$

$$\boldsymbol{\theta}^*_{SE} = \underset{\boldsymbol{\theta}_{SE}}{\arg\min}\,(\alpha\mathcal{L}_{CLS} + \beta\mathcal{L}_{SHP}) \tag{11}$$

$$\boldsymbol{\theta}^*_{SD} = \underset{\boldsymbol{\theta}_{SD}}{\arg\min}\,\mathcal{L}_{SHP} \tag{12}$$

where α and β is the scaling coefficient to balance $\mathcal{L}_{CLS}$ and $\mathcal{L}_{SHP}$, which is realized through the gradient scaling layer. Through the cooperative optimization of each module, the proposed method realizes texture and shape joint learning, improving the performance on shape-relied medical image recognition tasks.

TABLE 1 Number of samples in each dataset.

Dataset	Malignant	Benign	Total	Mask label*
ISIC-2019	4,522	12,875	17,397	2,671
XJTU-MM	2,170	6,928	9,098	726

*Due to not all samples having corresponding mask label, the shape-biased learning is only optimized when the input images have corresponding mask labels.

4. Experiments

4.1. Experimental setup

4.1.1. Data preparation

We use two medical image datasets to verify the effectiveness of the proposed method.

- **ISIC-2019:** A public and commonly used dermoscopic image dataset for dermatological diagnose. According to the advice from dermatologists, the malignant melanoma is one of the most dangerous skin cancer, and the melanoma lesions have similar visual characteristics to nevus. Therefore, we focus on the melanoma and nevi recognition task on this dataset. We use 12,875 nevi images and 4,522 malignant melanoma images, of which 2,671 images have corresponding lesion mask labels.
- **XJTU-MM:** A skin pathological image dataset collected from the Second Affiliated Hospital of Xi'an Jiaotong University(Xibei Hospital). It contains 9,098 images of RoI regions cropped from the whole slide histopathological images by pathologists, of which 2,170 images are malignant melanoma lesions and 6,928 images are benign nevus. And 726 of them have cell-wise masks labeled by pathologists.

The sample number of three datasets are shown in Table 1. Each dataset is divided into training set, validation set, and test set according to the ratio of 6:2:2, the images of malignant lesions are positive samples and the images of benign lesions are negative samples. Due to not all samples having the corresponding mask label, the shape-biased learning is only optimized when the input images have the corresponding mask labels.

4.1.2. Evaluation metrics

To quantitatively evaluate the performance of the model, we use accuracy(*Acc.*), precision(*Pre.*), recall(*Rec.*), and F1 score(*F1*) as evaluation metrics. They are calculated by

$$\begin{aligned} Acc. &= \frac{TP+TN}{TP+FP+TN+FN}, \\ Pre. &= \frac{TP}{TP+FP}, \\ Rec. &= \frac{TP}{TP+FN}, \\ F1 &= \frac{2\times Pre.\times Rec.}{Pre.+Rec.}, \end{aligned} \tag{13}$$

where *TP* (true positive) means the number of samples categorized to positive correctly, *TN* (true negative) means the number of samples categorized to negative correctly, *FP* (false positive) means the number of samples misclassified to malignant, *FN* (false negative) means the number of samples misclassified to negative. Higher accuracy reflects better overall performance of the model on all samples, higher precision means fewer malignant lesions are miss detected, and higher recall means higher sensitivity of the model to malignant lesions, F1 score is the combination of precision and recall. The four metrics provide a comprehensive evaluation of the medical image recognition models.

4.1.3. Implementation

In the proposed STNet-50, ResNet-50 is used as the baseline backbone of texture encoder and shape encoder, the shape feature decoder in the shape-biased stream is constructed using deconvolution operations and referring to the structure of ResNet-18. The texture encoder is pre-trained on ImageNet-1K. We implement the network using pytorch, opencv, scikit-learn and the libraries they depend on based on Python, and train the model on 2 RTX3090-24GB GPUs. All images are resized to 224 × 224, random rotation and random cropping are used for data augmentation. Batch size is set to 64, initial learning rate is set to $5e-4$, weight decay is set to $1e-5$, RMSprop (Hinton et al., 2012) is used as the optimization algorithm and the momentum is set to 0.9. The exponential decay factors in asymmetric loss is set to $\lambda_{+} = 1$, $\lambda_{-} = 3$.

4.2. Comparison results

We compared the proposed method with some popular general vision models, including the ResNeSt (Zhang H. et al., 2022), which is the latest iteration of ResNet, and ConvNeXt (Liu et al., 2022), which is regarded as CNN for 2020s. We also added some models designed for specific medical image recognition tasks to the comparative experiment, including DeMAL-CNN (He et al., 2022) for skin lesion classification in dermoscopy images, and MPMR (Zhang D. et al., 2022), which is a multi-scale-feature-based melanoma recognition method in pathological images.

The results are shown in Table 2, which indicate that the proposed STNet outperforms compared algorithms on two datasets and on all evaluation metrics. ConvNeXt series models show generally better performance than ResNeSt-50 on two datasets, which confirms the progress from split-attention block to ConvNet block. DeMAL-CNN shows a similar ability to ConvNeXt on ISIC-2019 dataset, considering that it uses standard ResNet as the backbone, the framework design of DeMAL-CNN has considerable contributions to enhance the dermoscopic image feature representation. MPMR shows better performance than ConvNeXt, which indicates that enhancing multi-scale features is effective in skin pathology image recognition. In addition, in each series of models, the increase in network layers does not bring about significant performance improvements, it is difficult to significantly improve the recognition accuracy of the model simply by increasing the number of layers. Furthermore, in four evaluation metrics, precision and recall are obviously lower than accuracy, which is caused by the sample imbalance of malignant and benign samples. In this case, accuracy cannot comprehensively reflect the performance of the model, it is necessary to add other three metrics.

Some difficult samples in the test set of XJTU-MM dataset are visualized and shown in Figure 6, where difficult samples mean the samples near the discriminant hyperplane. According to the results, The proposed STNet-50 correctly recognizes all of these samples. ResNeSt-50, ConvNeXt-S, and MPMR-50 all fail to recognition the first sample and the second sample, which contains rich irregular-shaped features. The fourth sample and the sixth sample have relatively distinct texture features distinct from melanoma, which is relatively easy to identify. The texture and feature joint learning enhances the shape feature representation, and the proposed asymmetric loss guides model to focus on difficult samples, so STNet has advantages on recognizing these difficult samples.

In summary, the results of comparative experiments on ISIC-2019 and XJTU-MM datasets proves the effectiveness of our method.

4.3. Ablation analysis

To further study the contribution of each module in our method, we design ablation experiments to analyze the effect of pyramid-grouped convolution(PGC), deformable convolution(DC) and channel-attention-based feature fusion(CAFF) on model performance. we remove all of these modules from the proposed STNet-50 and use it as the baseline model (first row in Table 3). And then PGC, DC and CAFF are rejoined to baseline model one by one (row 2–4 in Table 3). According to the results shown in Table 3, all the three modules bring performance improvement to model, especially in the increase of precision and recall. It indicates that PGC in the texture-biased stream and DC in shape-biased stream can both enhance the feature representation, and CAFF can select features that are more conducive to lesion identification. Additionally, these three modules are portable and can be plugged to other methods.

To further study the feature selection effect of CAFF in texture and shape feature fusion, we construct STNet-50 with CAFF and without CAFF respectively, and feed 500 malignant samples and 500 benign sample to them, for each sample, the feature vector in front of the classifier is input to t-SNE (Van der Maaten and Hinton, 2008) manifold learning model to study the separability of the extracted features. Through t-SNE, the input feature vectors are transformed into two dimensions and visualized in Figure 7. The comparison of Figures 7A, B show that the feature vector of the model with CAFF is more separable, which is conductive to

TABLE 2 Quantitative results of the proposed method and the comparison method on ISIC-2019 and XJTU-MM datasets.

Dataset	Model	*Acc.* ↑	*Pre.* ↑	*Rec.* ↑	*F1* ↑
ISIC-2019	ResNeSt-50	0.925	0.813	0.923	0.865
	ResNeSt-101	0.927	0.816	0.929	0.869
	ConvNeXt-S	0.949	0.858	0.964	0.908
	ConvNeXt-B	0.957	0.881	0.965	0.921
	DeMAL-50	0.952	0.864	0.967	0.913
	DeMAL-101	0.954	0.878	0.955	0.915
	STNet-50 (ours)	0.967	0.904	0.977	0.939
	STNet-101 (ours)	0.971	0.916	0.978	0.946
XJTU-MM	ResNeSt-50	0.929	0.828	0.885	0.855
	ResNeSt-101	0.933	0.846	0.880	0.863
	ConvNeXt-S	0.945	0.868	0.908	0.887
	ConvNeXt-B	0.946	0.875	0.901	0.888
	MPMR-50	0.958	0.894	0.935	0.914
	MPMR-101	0.961	0.910	0.929	0.919
	STNet-50 (ours)	0.979	0.954	0.959	0.956
	STNet-101 (ours)	0.985	0.963	0.972	0.968

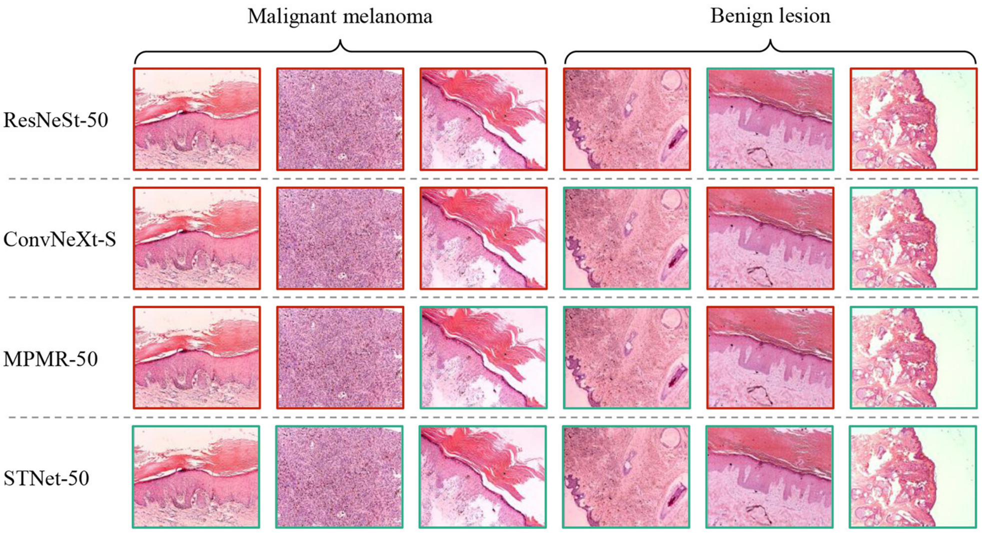


FIGURE 6
Visualized results of comparative experiment on XJTU-MM dataset. The green boxes mean correctly classified samples, the red boxes mean misclassified samples.

TABLE 3 Results of ablation analysis of pyramid-grouped convolution(PGC), deformable convolution(DC) and channel-attention-based feature fusion(CAFF) on ISIC-2019 dataset.

Module			*Acc.* ↑	*Pre.* ↑	*Rec.* ↑	*F1* ↑
PGC	DC	CAFF				
-	-	-	0.944	0.874	0.915	0.894
✓	-	-	0.951	0.884	0.933	0.908
✓	✓	-	0.959	0.895	0.955	0.924
✓	✓	✓	0.967	0.904	0.977	0.939

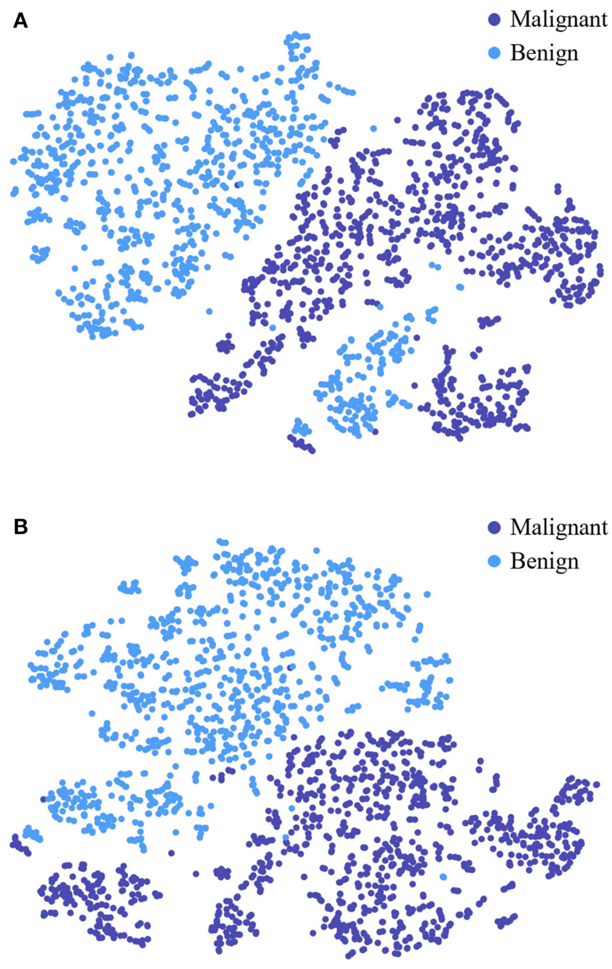


FIGURE 7
Visualized feature separability analysis through t-SNE. **(A)** Visualized result of STNet-50 without CAFF. **(B)** Visualized result of STNet-50 with CAFF. The feature vectors of STNet-50 with CAFF and STNet-50 without CAFF are transformed to two dimensions, respectively.

classification. The results indicate that the introduction of CAFF module is effective to select features relevant to lesion recognition.

Due to the available data is limited, to verify performance of the proposed model more rigorously, we conducted five-fold cross-validation on both ISIC-2019 and XJTU-MM datasets. Each dataset was divided into five mutually exclusive parts, with four used for training the STNet-50 model and one remaining part used for testing. Because of the sample imbalance problem, we use $F1$ score as the evaluation metric. The cross-validation results are shown in Table 4, STNet-50 shows consistent performance in each fold of the cross-validation, which proves the stability and reliability of the results.

4.4. Discussion on shape and texture joint learning framework

We propose the two-stream network for texture and shape joint learning, compared to single-stream network, an extra shape feature encoder is introduced. To analyze the contributions to performance improvements are provided by texture and shape joint learning or just the extra feature encoder, three control group models are designed for the comparative experiment. The first model uses the texture encoder only for feature extraction. The second model cascades the segmentation network and the classification network in the proposed method, the segmented lesion is used as the input of the classification network. The third model is constructed by removing the feature decoder of the shape-biased stream in our method, which is a two-stream network but without shape and texture joint learning. ISIC-2019 dataset is used for this experiment, the results are shown in Table 5, compared to the single-stream model, the cascade classification and segmentation model does not show obvious performance improvement and even have a performance drop on recall. It means that when the lesion mask labels are not sufficient, cascading the segmentation network and the classification network has limitation in solving weak shape representation problems. Two-stream network with joint learning shows better performance than that without joint learning, it indicates that the performance improvement of the proposed method is not simply brought by the extra shape feature encoder but by shape and texture joint learning, which proves the effectiveness of our method.

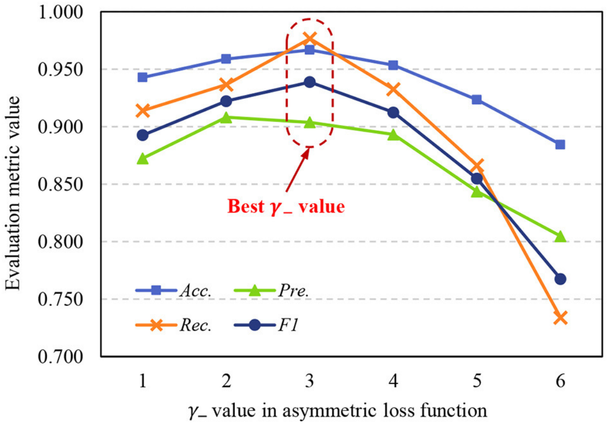


FIGURE 8
Variation of evaluation metrics with γ_- when $\gamma_+ = 1$. *Acc.*, accuracy; *Pre.*, precision; *Rec.*, recall; *F*1, F1 score.

4.5. Discussion on parameters of asymmetric loss

The asymmetric loss function in the proposed method is designed to address the sample imbalance problem, we use exponential decay factors γ_+ and γ_- to adjust the attention of the model to positive and negative classes. Due to in medical image datasets, malignant samples are usually much fewer than benign samples, γ_- should achieve a stronger decay effect, so $\gamma_+ < \gamma-$. To further study the effects of γ_+ and γ_- to model performance, we set $\gamma_+ = 1$, and use different γ_- to train the STNet-50 on ISIC-2019 dataset, the test results are shown

TABLE 4 Five-fold cross-validation results of the proposed STNet-50 model on ISIC-2019 and XJTU-MM datasets.

Datasets	*F*1 score ↑				
	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
ISIC-2019	0.939	0.930	0.932	0.939	0.935
XJTU-MM	0.956	0.953	0.955	0.953	0.952

TABLE 5 Experiments of discussion on shape and texture joint learning.

Backbone layers	Structure	*Acc.* ↑	*Pre.* ↑	*Rec.* ↑	*F*1 ↑
50	Single-stream[a]	0.950	0.872	0.945	0.907
	Cascade Cls. and Seg.[b]	0.950	0.886	0.928	0.907
	Two-stream without joint learning[c]	0.960	0.909	0.939	0.924
	Two-stream with joint learning[d]	0.967	0.904	0.977	0.939
101	Single-stream[a]	0.952	0.877	0.950	0.912
	Cascade Cls. and Seg.[b]	0.955	0.888	0.945	0.916
	Two-stream without joint learning[c]	0.961	0.911	0.944	0.927
	Two-stream with joint learning[d]	0.971	0.916	0.978	0.946

[a] Single-stream: only use the texture encoder in the proposed method for feature extraction.
[b] Cascade Cls. and Seg.: cascading segmentation network in front of classification network.
[c] Two-stream without joint learning: removing the feature decoder in the shape-biased stream of our method.
[d] Two-stream with joint learning: the proposed framework.

in Figure 8. Despite the model achieving the highest *Pre.* value When $\gamma_{-} = 2$, taking into account the four metrics, the model has the best performance when $\gamma_{-} = 3$. When γ_{-} is too small, exponential decay is not enough to eliminate the impacts of sample imbalance. When γ_{-} is too large, the effect of exponential decay is so strong that the model tends to ignore negative samples, and the performance of the model drops significantly. According to the results in Figure 8, choosing an appropriate value of the exponential decay factor is important to train a good-performance model.

5. Conclusion

In this paper, we propose the two-stream shape and texture joint learning network to address the weak shape feature representation problem of existing medical image recognition methods. According to the experiments on ISIC-2019 and XJTU-MM datasets, the proposed two-stream network is an effective method to combine texture and shape features. In addition, the proposed pyramid-grouped convolution enhances the texture feature representation, and deformable convolution enhances the shape feature representation. Furthermore, the channel-attention-based feature fusion module effectively eliminates redundant information and selects essential features. The asymmetric loss function addresses the problem of sample imbalance. The proposed method improves the model performance on shape-relied medical image recognition tasks, and provides support for computer-aided imaging diagnosis. Additionally, in our method, to enhance shape feature representation, an extra feature encoder is introduced, which increase the computation requirements, although the computation. Although inference speed is not the most critical concern in medical image analysis, we aim to enhance shape and texture feature representation by avoiding the use of additional encoders in future work, enhancing shape feature representation and texture feature representation within a single encoder.

Data availability statement

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

Author contributions

XW provided some ideas for this work. HH designed and implemented the models, ran the experiments, and wrote the manuscript. MenX analyzed the experimental data and visualized the results. SL helped write a part of the manuscript. DZ helped analyze the data and checked the manuscript writing. SD was in charge of project management. MeiX helps manage the project and provided advice for data analysis. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the National Key Research and Development Program of China under Grant No.

2017YFA0700800, the Natural Science Basic Research Plan in Shaanxi Province of China under Grant No. 2022JM-324, the Social Science Foundation of Shaanxi Province of China under Grant No. 2021K014, and the Key Project of Shaanxi Province under Grant No. 2018ZDCXLGY-06-07.

Acknowledgments

Thanks to Longfei Zhu, an attending doctor from the Department of Dermatology of the Second Affiliated Hospital of Xi'an Jiaotong University (Xibei Hospital), for his valuable advice on analysis of dermoscopic image data and pathology image data.

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

Ahn, E., Kim, J., Bi, L., Kumar, A., Li, C., Fulham, M., et al. (2017). Saliency-based lesion segmentation via background detection in dermoscopic images. *IEEE J. Biomed. Health Inform.* 21, 1685–1693. doi: 10.1109/JBHI.2017.2653179

Al-Osaimi, F. R., Bennamoun, M., and Mian, A. (2011). Spatially optimized data-level fusion of texture and shape for face recognition. *IEEE Trans. Image Process.* 21, 859–872. doi: 10.1109/TIP.2011.2165218

Anantharatnasamy, P., Sriskandaraja, K., Nandakumar, V., and Deegalla, S. (2013). "Fusion of colour, shape and texture features for content based image retrieval," in *2013 8th International Conference on Computer Science & Education* (Colombo: IEEE), 422–427.

Araujo, A., Norris, W., and Sim, J. (2019). Computing receptive fields of convolutional neural networks. *Distill* 4, e21. doi: 10.23915/distill.00021

Celebi, M. E., Codella, N., and Halpern, A. (2019). Dermoscopy image analysis: overview and future directions. *IEEE J. Biomed. Health Inform.* 23, 474–478. doi: 10.1109/JBHI.2019.2895803

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

Chang, H. (2017). Skin cancer reorganization and classification with deep neural network. *arXiv preprint arXiv:1703.00534.* doi: 10.48550/arXiv.1703.00534

Chen, K., Guo, Y., Yang, C., Xu, Y., Zhang, R., Li, C., et al. (2021). "Enhanced breast lesion classification via knowledge guided cross-modal and semantic data augmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference* (Strasbourg: Springer), 53–63.

Chollet, F. (2017). "Xception: deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI), 1251–1258.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). "ImageNet: a large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (Miami, FL: IEEE), 248–255.

Fan, H., Xie, F., Li, Y., Jiang, Z., and Liu, J. (2017). Automatic segmentation of dermoscopy images using saliency combined with otsu threshold. *Comput. Biol. Med.* 85, 75–85. doi: 10.1016/j.compbiomed.2017.03.025

Gao, L., Liu, C., Arefan, D., Panigrahy, A., Zuley, M. L., and Wu, S. (2021). Medical knowledge-guided deep learning for imbalanced medical image classification. *arXiv preprint arXiv:2111.10620.* doi: 10.48550/arXiv.2111.10620

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. (2018). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231.* doi: 10.48550/arXiv.1811.12231

Guo, Y., Li, Y., Wang, L., and Rosing, T. (2019). "Depthwise convolution is all you need for learning multiple visual domains," in *Proceedings of the AAAI Conference on Artificial Intelligence* (Honolulu, HI), 8368–8375.

Guo, Y., Liu, Y., Georgiou, T., and Lew, M. S. (2018). A review of semantic segmentation using deep neural networks. *Int. J. Multimedia Inform. Retrieval* 7, 87–93. doi: 10.1007/s13735-017-0141-z

Han, H., Du, S., Zhang, D., Long, H., and Guo, Y. (2020). "Precise dental staging method through panoramic radiographs based on deep learning," in *2020 Chinese Automation Congress (CAC)* (Shanghai), 7406–7411. doi: 10.1109/CAC51589.2020.9327719

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). "Mask r-CNN," in *Proceedings of the IEEE International Conference on Computer Vision* (Venice), 2961–2969.

He, X., Wang, Y., Zhao, S., and Yao, C. (2022). Deep metric attention learning for skin lesion classification in dermoscopy images. *Complex Intell. Syst.* 8, 1487–1504. doi: 10.1007/s40747-021-00587-4

Hinton, G., Srivastava, N., and Swersky, K. (2012). *Neural Networks for Machine Learning Lecture 6a Overview of Mini-Batch Gradient Descent*. Department of Computer Science, Toronto University, Toronto, ON, Canada. Available online at: https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Hawaii, HI), 4700–4708.

Jo, J., Lee, S. J., Park, K. R., Kim, I.-J., and Kim, J. (2014). Detecting driver drowsiness using feature-level fusion and user-specific classification. *Expert Syst. Appl.* 41, 1139–1152. doi: 10.1016/j.eswa.2013.07.108

Kotsia, I., Zafeiriou, S., and Pitas, I. (2008). Texture and shape information fusion for facial expression and facial action unit recognition. *Pattern Recogn.* 41, 833–851. doi: 10.1016/j.patcog.2007.06.026

Kurc, T., Bakas, S., Ren, X., Bagari, A., Momeni, A., Huang, Y., et al. (2020). Segmentation and classification in digital pathology for glioma research: challenges and deep learning approaches. *Front. Neurosci.* 14, 27. doi: 10.3389/fnins.2020.00027

Liu, T., Fan, W., and Wu, C. (2019a). A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset. *Artif. Intell. Med.* 101, 101723. doi: 10.1016/j.artmed.2019.101723

Liu, T., Guo, Q., Lian, C., Ren, X., Liang, S., Yu, J., et al. (2019b). Automated detection and classification of thyroid nodules in ultrasound images using clinical-knowledge-guided convolutional neural networks. *Med. Image Anal.* 58, 101555. doi: 10.1016/j.media.2019.101555

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. (2022). "A ConvNet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (New Orleans, LA), 11976–11986.

Long, J., Shelhamer, E., and Darrell, T. (2015). "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Boston, MA), 3431–3440.

Lu, S., Zhao, H., Liu, H., Li, H., and Wang, N. (2023). PKRT-Net: prior knowledge-based relation transformer network for optic cup and disc segmentation. *Neurocomputing* 538, 126183. doi: 10.1016/j.neucom.2023.03.044

Lu, Z., Yang, J., and Liu, Q. (2017). Face image retrieval based on shape and texture feature fusion. *Comput. Visual Media* 3, 359–368. doi: 10.1007/s41095-017-0091-7

Luo, W., Li, Y., Urtasun, R., and Zemel, R. (2016). "Understanding the effective receptive field in deep convolutional neural networks," in *30th Conference on Neural Information Processing Systems (NIPS 2016)*, eds D. Lee, M. Sugiyama, U. Luxburg, I. Guyon and R. Garnett [Barcelona: Neural Information Processing Systems Foundation, Inc. (NeurIPS)], 4898–4906.

Ma, N., Zhang, X., Zheng, H.-T., and Sun, J. (2018). "ShuffleNet V2: practical guidelines for efficient CNN architecture design," in *Proceedings of the European Conference on Computer Vision (ECCV)* (Munich), 116–131.

Ni, D., Yang, Y., Li, S., Qin, J., Ouyang, S., Wang, T., et al. (2013). "Learning based automatic head detection and measurement from fetal ultrasound images via prior knowledge and imaging parameters," in *2013 IEEE 10th International Symposium on Biomedical Imaging* (San Francisco, CA: IEEE), 772–775.

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., et al. (2018). Attention U-Net: learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*. doi: 10.48550/arXiv.1804.03999

Pan, L., Cai, Y., Lin, N., Yang, L., Zheng, S., and Huang, L. (2022). A two-stage network with prior knowledge guidance for medullary thyroid carcinoma recognition in ultrasound images. *Med. Phys.* 49, 2413–2426. doi: 10.1002/mp.15492

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-Net: convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference* (Munich: Springer), 234–241.

Rotemberg, V., Kurtansky, N., Betz-Stablein, B., Caffery, L., Chousakos, E., Codella, N., et al. (2021). A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Sci. Data* 8, 34. doi: 10.1038/s41597-021-00815-z

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

Shen, D., Wu, G., and Suk, H.-I. (2017). Deep learning in medical image analysis. *Annu. Rev. Biomed. Eng.* 19, 221–248. doi: 10.1146/annurev-bioeng-071516-044442

Shi, G., Wang, J., Qiang, Y., Yang, X., Zhao, J., Hao, R., et al. (2020). Knowledge-guided synthetic medical image adversarial augmentation for ultrasonography thyroid nodule classification. *Comput. Methods Prog. Biomed.* 196, 105611. doi: 10.1016/j.cmpb.2020.105611

Sumathi, C., and Kumar, A. S. (2012). Edge and texture fusion for plant leaf classification. *Int. J. Comput. Sci. Telecommun.* 3, 6–9.

Tan, M., and Le, Q. V. (2019). Mixconv: Mixed depthwise convolutional kernels. *arXiv preprint arXiv:1907.09595*. doi: 10.48550/arXiv.1907.09595

Van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605.

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. (2017). "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Hawaii, HI), 1492–1500.

Xiong, L., Zheng, N., You, Q., and Liu, J. (2007). "Facial expression sequence synthesis based on shape and texture fusion model," in *2007 IEEE International Conference on Image Processing* (San Antonio, TX: IEEE), 4–473.

Xu, Z., Shen, D., Nie, T., and Kou, Y. (2020). A hybrid sampling algorithm combining m-smote and ENN based on random forest for medical imbalanced data. *J. Biomed. Inform.* 107, 103465. doi: 10.1016/j.jbi.2020.103465

Yanase, J., and Triantaphyllou, E. (2019). A systematic survey of computer-aided diagnosis in medicine: past and present developments. *Expert Syst. Appl.* 138, 112821. doi: 10.1016/j.eswa.2019.112821

Yang, W., Dong, Y., Du, Q., Qiang, Y., Wu, K., Zhao, J., et al. (2021). Integrate domain knowledge in training multi-task cascade deep learning model for benign–malignant thyroid nodule classification on ultrasound images. *Eng. Appl. Artif. Intell.* 98, 104064. doi: 10.1016/j.engappai.2020.104064

Yang, Y., Xie, F., Zhang, H., Wang, J., Liu, J., Zhang, Y., et al. (2023). Skin lesion classification based on two-modal images using a multi-scale fully-shared fusion network. *Comput. Methods Prog. Biomed.* 229, 107315. doi: 10.1016/j.cmpb.2022.107315

Zhang, D., Han, H., Du, S., Zhu, L., Yang, J., Wang, X., et al. (2022). MPMR: multi-scale feature and probability map for melanoma recognition. *Front. Med.* 8, 775587. doi: 10.3389/fmed.2021.775587

Zhang, D., Yang, J., Du, S., Han, H., Ge, Y., Zhu, L., et al. (2023). Coarse-to-fine feature representation based on deformable partition attention for melanoma identification. *Pattern Recogn.* 136, 109247. doi: 10.1016/j.patcog.2022.109247

Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Lin, H., Zhang, Z., et al. (2022). "Resnest: split-attention networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2736–2746.

Zhang, J., Li, C., Kosov, S., Grzegorzek, M., Shirahama, K., Jiang, T., et al. (2021). LCU-Net: a novel low-cost u-net for environmental microorganism image segmentation. *Pattern Recogn.* 115, 107885. doi: 10.1016/j.patcog.2021.107885

Zhang, X., Wang, S., Liu, J., and Tao, C. (2018a). Towards improving diagnosis of skin diseases by combining deep neural network and human knowledge. *BMC Med. Inform. Decis. Mak.* 18, 59. doi: 10.1186/s12911-018-0631-9

Zhang, X., Zhou, X., Lin, M., and Sun, J. (2018b). "ShuffleNet: an extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT), 6848–6856.

Zhang, Y., Zhang, P., Yuan, C., and Wang, Z. (2020). "Texture and shape biased two-stream networks for clothing classification and attribute recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Seattle, WA), 13538–13547.

Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., and Liang, J. (2018). "UNet++: a nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018* (Granada: Springer), 3–11.

Zhou, Z., Zhao, C., Qiao, H., Wang, M., Guo, Y., Wang, Q., et al. (2022). Rating: medical knowledge-guided rheumatoid arthritis assessment from multimodal ultrasound images via deep learning. *Patterns* 3, 100592. doi: 10.1016/j.patter.2022.100592

TYPE Methods
PUBLISHED 19 June 2023
DOI 10.3389/fnins.2023.1201370

A semi-independent policies training method with shared representation for heterogeneous multi-agents reinforcement learning

Biao Zhao [1†], Weiqiang Jin [1†], Zhang Chen[2] and Yucheng Guo[3,4*]

[1]School of Information and Communications Engineering, Xi'an Jiaotong University, Xi'an, China, [2]School of Software Engineering, Xi'an Jiaotong University, Xi'an, China, [3]Key Laboratory of Shaanxi Province for Craniofacial Precision Medicine Research, College of Stomatology, Xi'an Jiaotong University, Xi'an, China, [4]Department of Orthodontics, Stomatological Hospital of Xi'an Jiaotong University, Xi'an, China

OPEN ACCESS

EDITED BY
Shihui Ying,
Shanghai University, China

REVIEWED BY
Rui Qin,
Manchester Metropolitan University, United Kingdom
Babak Rahi,
University of Nottingham, United Kingdom
Yanhua Liu,
Shandong University, China

*CORRESPONDENCE
Yucheng Guo
✉ xjtu-guoyucheng@163.com

[†]These authors have contributed equally to this work and share first authorship

RECEIVED 06 April 2023
ACCEPTED 29 May 2023
PUBLISHED 19 June 2023

CITATION
Zhao B, Jin W, Chen Z and Guo Y (2023) A semi-independent policies training method with shared representation for heterogeneous multi-agents reinforcement learning. *Front. Neurosci.* 17:1201370. doi: 10.3389/fnins.2023.1201370

Humans do not learn everything from the scratch but can connect and associate the upcoming information with the exchanged experience and known knowledge. Such an idea can be extended to cooperated multi-reinforcement learning and has achieved its success on homogeneous agents by means of parameter sharing. However, it is difficult to straightforwardly apply parameter sharing when dealing with heterogeneous agents thanks to their individual forms of input/output and their diverse functions and targets. Neuroscience has provided evidence that our brain creates several levels of experience and knowledge-sharing mechanisms that not only exchange similar experiences but also allow for sharing of abstract concepts to handle unfamiliar situations that others have already encountered. Inspired by such a brain's functions, we propose a semi-independent training policy method that can well tackle the conflict between parameter sharing and specialized training for heterogeneous agents. It employs a shared common representation for both observation and action, enabling the integration of various input and output sources. Additionally, a shared latent space is utilized to maintain a balanced relationship between the upstream policy and downstream functions, benefiting each individual agent's target. From the experiments, it can approve that our proposed method outperforms the current mainstream algorithms, especially when handling heterogeneous agents. Empirically, our proposed method can also be improved as a more general and fundamental heterogeneous agents' reinforcement learning structure for curriculum learning and representation transfer. All our code is open and released on https://gitlab.com/reinforcement/ntype.

KEYWORDS

brain function, knowledge-sharing institution, multi-agent reinforcement learning, parameters sharing, representation transferability

1. Introduction

The attention on Multi-agent reinforcement learning (MARL) is booming largely since a lot of real-world cooperatives challenges can be properly solved. The scenarios such as distributed network routers, sensor networks (Zhang and Lesser, 2011), traffic management (Singh et al., 2020), and coordination of robot swarms (Hüttenrauch et al., 2017), etc. can be better modeled as MARL where the decision on controlling and management are distributed made. However, the introduction of multi-agent reinforcement learning has also brought in 2 challenges: increased computation requirements due to the larger observation and action spaces, and difficulty in convergence during training due to the presence of other agents.

Multi-Agent Reinforcement Learning (MARL) methods can be classified into two categories based on the level of centralization in decision-making and learning (centralized or decentralized). In decentralized systems, each agent makes decisions and learns on its own, without accessing the observations, actions, or policies of other agents. However, decentralized learning lacks the guarantee of convergence due to the non-stationary caused by other agents. Therefore, most modern MARL research follows the paradigm of Centralized Training and Decentralized Execution (CTDE), where agents have access to other agents' observations during training but execute their own policies separately. Examples of CTDE include MADDPG (Lowe et al., 2017), COMA (Foerster et al., 2018), and QMIX (Rashid et al., 2020).

Based on such a paradigm, the idea of parameter sharing is naturally born following the merging of multi-reinforcement learning. It coheres to the human intuition that knowledge sharing can make better learning and judgment. Humans do not learn everything from scratch but exchange knowledge when learning from experience. This idea was first introduced for classical RL (Tan, 1993) and later extended to cooperative multi-agent reinforcement learning (Chu and Ye, 2017; Gupta et al., 2017). Homogeneous multi-reinforcement learning has achieved success when utilizing parameter sharing. They leverage an identical policy trained with all the trajectories. This method is more efficient compared to training multiple independent policies, as only one policy is employed for both learning and training, reducing the high computational demands, and difficulties in achieving convergence.

The application of parameter sharing to heterogeneous agents is limited in its effectiveness due to the homogenizing effect it has on agents' behavior, particularly at the early stages. Additionally, the shared policy results in a fixed observation and action space size. To address this, some algorithms utilize zero-padding to standardize inputs and outputs, and allow a single policy to serve multiple agents (Gupta et al., 2017; Foerster et al., 2018). These strategies have helped to reduce the obstacles to further extension to hetero agents. It works well for agents with fewer functional and targeting variations or for environments easy to normalize the input and output but not for an abundance diversity of agents. However, this adaptation may not be suitable for all situations, such as when there are different dimensions of inputs and outputs that are not easy to be unified through extra padding of inputs or outputs. The policy for diverse agents also results in slow convergence. Therefore, a more flexible parameter-sharing and policy training strategy is desirable for the real-world application.

Neuroscience has provided evidence that our brain establishes various levels of experience and knowledge-sharing institutions that not only exchange similar experiences but also allow for the exchange of abstract concepts to tackle novel situations that others have already encountered. Inspired by this, we propose a semi-independent training policy method that applies identical policies among the same type of agents and semi-independent parameter-sharing schemes between different types for tackling the conflict between parameter sharing and specialized training for heterogeneous agents. This method also utilizes a common shared representation, generated by supervised learning, to formalize the observations and actions of the agents, allowing it to handle all types of inputs and outputs. An intrinsic reward is also introduced to speed up the environmental exploration. Experimental results demonstrate that our proposed method outperforms the current mainstream algorithms, particularly when dealing with heterogeneous agents. In advance, our proposed method can be considered as a more general and fundamental structure for heterogeneous agent reinforcement learning, incorporating curriculum learning and representation transferring.

This paper is organized as follows. In Section 2, we provides some background on Multi-agent Reinforcement Learning (MARL) and recent advances in Deep Reinforcement Learning (DRL) relevant to the proposal. Section 3 presents the proposal in detail. In Section 4, we will detail the experiments performed and their results. In Section 5, we will review the related work concerning our proposed MARL, including curriculum learning and representation transferring. Lastly, in Section 6, we will summarize the conclusions and suggests future research directions.

1.1. Main contribution

This paper presents three main contributions we have made to our proposal.

First, we introduce and adopt a hard-parameter-sharing scheme to MARL in order to balance the conflicting requirements of agents' specialization and network fast convergence. This scheme was originally proposed for multi-task networks, which take a parameter-shared base to process the input and multiple-task terminals to handle different tasks. This structure accounts for specialization among heterogeneous agents while still attempting for the maximum level of experience sharing. Based on our knowledge, there is no other literature currently existing for this approach, and our work is the first made such attempt to introduce the multi-task network parameter-sharing scheme to multi-agent reinforcement learning.

Second, we invent a supervised learning method to generate a general input and output representation shared with all agents. The shared common representation facilitates the formalization of input and output, thus resolving the diversity of heterogeneous agents' input and output issues, and making it easier to incorporate the hard-parameter-sharing scheme. Thanks to this common shared input/output representation, all the agents will be equally treated after the input/output processing regardless of the types of agents. Empirically, We carried out such an approach by simultaneously training with reinforcement learning to ensure that the representation is both accurate and precise. It can be approved such a training schedule can fast generate the representation to facilitate policy training.

Third, we introduce an extra intrinsic reward to encourage more exploration of the environment initially. Unlike traditional intrinsic rewards which are based on a comparison of trajectories, our proposed intrinsic reward is based on the prediction of supervised learning and its input/output representation. Such a tactic can help to stimulate more exploration right away without requiring extra effort and well incorporate the representation generation process.

2. Background

2.1. Reinforcement learning

Reinforcement Learning (RL) methods attempt to identify an optimal policy (a function that takes an observation and returns an action) that maximizes the expected total reward from an environment. Commonly, such environments are modeled as a Markov Decision Process (MDP) or Partially-Observable Markov Decision Process (POMDP) (Boutilier, 1996). MDPs characterize decision-making as a repetitive process whereby an agent takes an action, receives a reward, and transitions to a new state (with perfect knowledge of the state). POMDP extends this to include environments in which the agent may not be able to observe the full state information.

In Deep Reinforcement Learning (DRL), a neural network is used to represent the policy. These methods are typically divided into two categories: Q-learning methods and policy gradient (PG) methods. The first deep Q learning method was Deep Q Network (DQN) (Mnih et al., 2013), and the first widely-used PG method was Deep Deterministic Policy Gradient (DDPG) (Lillicrap et al., 2015). Subsequently, various newer, more powerful methods were developed, including Soft Actor-Critic (SAC) (Haarnoja et al., 2018), TD3 (Fujimoto et al., 2018), Proximal Policy Optimization (PPO) (Chu and Ye, 2017), (the synchronous version of Asynchronous Advantage Actor-Critic (A3C) (Mnih et al., 2016), Rainbow DQN (Hessel et al., 2018) etc., and more advanced deep reinforcement learning methods is on the way of development for the real-world applications.

Multi-agent reinforcement learning (MARL) can be deemed as an extension of RL that considers the interactions between multiple agents in a changing environment. The agents must learn to adjust their actions based on changes not only in the environment but also in the behavior of other agents. MARL can lead to distributed intelligent decision-making and has applications in game theory and robotics. Our proposed method focuses on developing a fast and accurate MARL algorithm for practical use.

2.2. Brain's transfer learning on the new tasks

Learning is not a process that begins from scratch, as people can connect and relate new information to their existing experiences and knowledge. Recent neuroscience research has shown that the brain has the capacity to transfer knowledge from one task to another, even if the tasks appear dissimilar. The brain's ability to extract and store abstract representations of information is the reason behind this transfer. When confronted with a new task, the brain first looks for similarities with past experiences, allowing individuals to learn how to handle the new task quickly. These abstract experiences can also be shared and learned by others, highlighting the importance of utilizing past experiences and knowledge to facilitate learning.

2.3. Dec-POMDP

Decentralized Partially Observable Markov Decision Processes (Dec-POMDPs) are a probabilistic framework for enabling distributed decision-making among multiple agents. It has been commonly utilized for decision-making in cooperated large-scale multi-agent settings, originally proposed in the literature on autonomous multi-agent systems (Lillicrap et al., 2015). In this framework, each agent has a set of actions and observations defined in mathematics that it can take in order to achieve a goal. The environment is represented as a stochastic process that is partially observable to the agents.

A Dec-POMDP on MARL can be formally defined as a tuple $(N, \mathrm{S}, \mathbf{A}, \mathbf{P}, \mathbf{R}, \Omega, \mathrm{O}, \mathrm{n}, \gamma)$, where:

- N is a finite state of n agents where $i \in N \equiv \{1, \ldots, n\}$;
- S is the global state of the environment where $s \in S$;
- $\mathbf{A}$ is a set of joint actions, $\mathbf{A} = \mathcal{A}_{\mathbf{1}} \times \cdots \times \mathcal{A}_{\mathbf{N}}$ where $\mathcal{A}_i$ is the set of actions that the i-th agent can choose from;
- P is a state transition probability function where $P(s^{'}|s, \mathbf{A}) : \mathrm{S} \times \prod_{\mathbf{i} \in \mathbf{N}} \mathcal{A}_{\mathbf{i}} \times \mathrm{S} \to [\mathbf{0}, \mathbf{1}]$;
- $\mathbf{R}$ is a reward function, often can be modeled as $\mathbf{R} = \mathrm{R}(\mathrm{S}, \mathbf{A})$, where $R_i \in \mathbf{R} : \mathrm{S} \times \prod_{\mathbf{i} \in \mathbf{N}} \mathcal{A}_{\mathbf{i}} \times \mathrm{S} \to \mathbb{R}$ is the reward function for agent i;
- Ω is the set of observations, where $\Omega_i \in \Omega$ is the possible observation for agent i;
- O is the observation function, normally modeled as $O(S, i)$. According to the settings of partial observation, the agent cannot access the global state but samples local observations according to the observation function where $S \times \mathcal{A}_i \equiv \Omega_i$, which can generate the set of observation that ith agent can receive;
- γ is the discount factor, where $\gamma \in [0, 1)$. The utilization of the discount factor is to compromise for the reward one can receive a few steps later than immediately.

The set of agents A comprises the agents that are involved in the decision-making problem, each of which has its own set of decisions and observations. The set of observed states S represents the states of the environment, which are partially observed by the agents. Finally, the set of joint actions A contains the joint actions taken by all, which are finally to determine the probability of transitioning to different states. The Dec-POMDP framework allows agents to make optimal decisions in a partially observable environment by combining their observations and taking into account their own rewards and the rewards of their peers (Oliehoek, 2012).

Such a framework can be perfectly utilized to describe the decision-making in cooperated large-scale multi-agent settings, thus we will also apply the above-mentioned mathematics definitions in this paper to describe our proposal.

2.4. Parameters sharing

The concept of parameter sharing is a widespread practice in the field of deep learning. It refers to an approach where a single set of parameters is shared among multiple components of a neural network, such as layers or sub-networks. In the context of multi-agent reinforcement learning, parameter sharing involves an algorithm that learns from the experiences of all agents and updates a collectively shared policy. Parameter sharing, which involves representing all policies with a single neural network that shares the same set of parameters, was first introduced by Tan (1993) for classical reinforcement learning. Later, it was concurrently introduced to cooperative multi-agent deep reinforcement learning by Chu and Ye (2017) and Gupta et al. (2017). This straightforward approach has proven to be highly effective in various applications, including those presented in Zheng et al. (2018), Chen et al. (2021), and Yu et al. (2022). This paper will discuss parameter sharing in detail and make a proposal based on that with a more general framework and structure from the common representation and semi-independent training and will further analyze the effectiveness and utilization of representation transferring and curriculum learning.

2.5. Coping with heterogeneity

Heterogeneity in agents is a common challenge in multi-agent systems, which can arise due to various reasons, such as differences in the physical capabilities or perceptual abilities of the agents. Addressing this issue is crucial to ensure that the agents can effectively cooperate and achieve their goals. To address such a challenge, two methods have been proposed. The first method is to add an indication of observations to enable a single policy to serve multiple agents, accommodating different action and observation spaces. However, since there is only one neural network, the observation spaces of all agents must be the same size especially when the observation spaces of agents are vastly different, as the neural network may struggle to learn from a disparate input. The second method proposes "padding" observations and action spaces to a uniform size, which allows agents to ignore any actions outside their "true" action space. By standardizing the observation and action spaces, the agents can effectively communicate with each other, and the neural network can learn from these inputs more efficiently. However, this approach may introduce redundant or irrelevant information, leading to additional computational overhead. Also the initial policies it generated with the unified neutral network will be also less efficient and mislead to sub-optimal when the network cannot well recognize the correct information and "padding".

3. Preliminary

3.1. Representation learning

Reinforcement learning (RL) involves training an agent through interactions with an environment. This formalism is powerful in its generality, but poses an open-ended problem: how can we design agents that learn efficiently and generalize well, given only sensory information and a scalar reward signal? One solution that is becoming increasingly popular is introducing self-supervised learning. Applying self-supervised learning in RL can help solve problems with high-dimensional state-action spaces and improve sample efficiency by incorporating inductive biases, such as structural information about tasks anden vs, into the representations for better performance.

The UNREAL agent (Jaderberg et al., 2016) introduced unsupervised auxiliary tasks to deep RL, including the Pixel Control task, a Q-learning method that requires predictions of screen changes in discrete control environments, which has become a standard in DMLab (Hessel et al., 2019). CPC (Oord et al., 2018) applied contrastive losses over multiple time steps as an auxiliary task for the convolutional and recurrent layers of RL agents, and it has been extended with future action-conditioning (Guo et al., 2018). Recently, PBL (Guo et al., 2020) surpassed these methods with an auxiliary loss of forward and backward predictions in the recurrent latent space using partial agent histories. A small number of model-free methods have attempted to decouple encoder training from the RL loss as ablations, but have met reduced performance relative to end-to-end RL (Laskin et al., 2020). Examples of works that pre-train encoder features in advance using image reconstruction losses, such as the VAE (Kingma and Welling, 2013), PR2 (Finn et al., 2016), and World models (Ha and Schmidhuber, 2018). Other works (Devin et al., 2018; Kipf et al., 2019), apply pre-trained object-centric representations that learn a forward model through contrasting losses. CFM (Yan et al., 2021) introduced a similar technique to learn encoders that support the manipulation of deformable objects through traditional control methods. In this paper, we will leverage an encoder-decoder framework to formalize the various inputs and output for heterogeneous agents.

3.2. Hard/soft parameter sharing

Hard parameter sharing is a fundamental scheme that enables domains to share some of their model parameters to reduce storage costs and improve prediction accuracy. This approach originated from multi-task learning (MTL), which aims to support multiple downstream tasks on devices. While recent advancements in model compression have made deploying a single model easier, supporting multiple models on devices remains challenging due to increased bandwidth, energy, and storage costs. To address this challenge, the hard/soft parameter-sharing approach has been employed. Unlike soft parameter sharing, where each task keeps its own model and parameters, hard parameter sharing allows multiple tasks to share some of the model parameters. As depicted in Figure 1, this sharing is commonly applied by sharing the bottom layers among all tasks while keeping several top layers and an output layer task-specific (Ruder, 2017). Hard parameter sharing is often used in designing multi-task deep neural network models (Long et al., 2017; Ruder et al., 2019).

Given its effectiveness in MTL, we believe that utilizing hard parameters can also be a viable solution for sharing policies among different types of agents to share the basis while maintaining dependence.

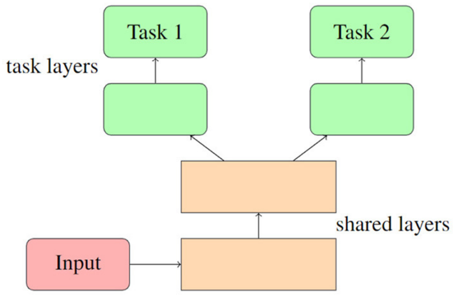


FIGURE 1
A typical structure of hard parameter sharing.

3.3. Role-based learning method

Roles are a fundamental aspect of natural systems, such as ants, bees, and humans, where they are closely related to the division of labor and crucial for labor efficiency. This concept has inspired multi-agent system designers to reduce design complexity by assigning agents with the same roles to specific sub-tasks. However, in such systems, roles and their associated responsibilities are typically predefined using prior knowledge, limiting their generalizability and requiring prior knowledge that may not always be available. To overcome this challenge, Wilson et al. (2010) utilized Bayesian inference to learn a set of roles, while ROMA (Wang et al., 2019) developed a specialization objective to encourage the emergence of roles, method RODE (Wang et al., 2020b) proposes a scalable role-based multi-agent learning method that effectively discovers roles by decomposing the joint action space according to action effects, thereby access to the producing of role selectors and learning of role policies in the reduced spaces. These methods suffer from a limitation in searching for the optimal task decomposition in the full state-action space, resulting in inefficient learning in hard-exploration tasks. Our work is inspired by the concept of role-based policy training, and we propose a method that groups agents by their unit types. Within each group, we implement a full parameter-sharing scheme, while across different groups, we use a semi-sharing parameter scheme. This approach can facilitate faster convergence for agents with similar roles or types while allowing for greater flexibility in learning different strategies or behaviors for agents with different roles or types.

4. Proposal

Based on the preliminary research mentioned above, we propose our semi-independent training policy method with shared representation (STSR) for reinforcement learning. This method comprises three main components: a common inputs/outputs representation derived from supervised learning, a semi-independent policy training scheme that applies full shared parameters among agents of the same type/role and hard sharing among different types, and an intrinsic/diversity-driven extra reward to encourage environment exploration and enhance the representation that can more clearly distinguish the inputs and outputs from different types/roles. Before we delve into each component, Figure 2 depicts the graph illustrating the entire process.

Our idea is to use supervised learning to build a prediction model, which enables us to establish an observation-action embedding to formalize the agent's input and output, regardless of their invariant observation and actions. Based on that, we can extend a hard parameter-sharing scheme to multiple heterogeneous agents, which fully shares the parameters among the same types of agents and employs hard sharing between different type groups. From the learned representation, we will generate an extra intrinsic reward to encourage environment exploration and an identifying reward to enhance the representation difference between agent types. We provide a clear definition of the agent types and representations below.

Definition 1 *Given a cooperative multi-agent task* G=(N,S, $\mathbf{A}, \mathbf{P}, \mathbf{R}, \Omega, \mathrm{O}, \mathrm{n}, \gamma$), *let* K_j *be a set of agent type with the total type accounts for* j, *where each agent* $i \in \mathrm{K}_j$. *Each type with the same policy forms as the tuple* (g_j, π_{K_j}), *where* $g_j = (N_j, \mathrm{S}, \mathbf{A_j}, \mathrm{P_j}, \mathbf{R}, \Omega_\mathbf{j}, \mathrm{O}, \mathrm{n_j}, \gamma, \mathrm{Z}_\mathrm{j}^\mathrm{o}, \mathrm{Z}_\mathrm{j}^\mathrm{a})$ *can be defined as a sub-space for each type,* $\pi_{K_j} : T \times A_j \rightarrow [0, 1]$ *is a full parameter shared type policy, associated with each type.* $\mathrm{Z}_\mathrm{j}^\mathrm{o} = \mathrm{Z}^0(o_i, K_j), \mathrm{Z}_\mathrm{j}^\mathrm{a} = \mathrm{Z}^a(a_i, K_j)$ *are the observation representation function and action representation function, respectively, shared for each type.*

Our aim is to seek a set of hard parameters shared policies $\pi_K j$ that can maximize the expected global return $Q(s_t, \mathbf{a}_t) = \mathbb{E}_{s_{t+1:\infty}, \mathbf{a}_{t+1:\infty}}[\sum_{i=0}^{\infty} \gamma^i r_{t+1} | s_t, \mathbf{a}_t, K(Z^o, Z^a)]$. The policies $\pi_K j$ are also related to each other in terms of basic representation $\mathrm{Z}^a, \mathrm{Z}^o$, and low-level layers. We will now introduce the comprised each component in detail, which is illustrated in Figure 2.

4.1. Common observation and action representation

To well handle the heterogeneous agents and to improve the effectiveness of parameter sharing, we attempt to cluster the agents according to their types and then exert full parameter sharing among unit type. Even though some role-based MARL (Wang et al., 2020b; Christianos et al., 2021) do the partition of the agents according to their representation latent space, we group our agents based on the agent's unit type.

To formalize the input and outputs from different types of agents and to better architecture the hard parameter sharing schemes, we propose a recurrent neural network (RNN) based prediction model for learning the observation and action latent representation that incentivizes including enough information such that the next observations and rewards can be predicted when given the actions and current observations.

As it is depicted in Figure 3, a collection of functions $\mathrm{Z}^0(o_i, K_j, t)$ and $\mathrm{Z}^a(a_i, K_j, t)$ are employed to estimate o_{t+1}^i and r_{t+1}^i, respectively, from the agents' limited view of the world. Due to the fact that an agent cannot perceive the state or actions of another agent, we define $\hat{O}^i : O^i \times A^i \rightarrow \Delta(O^i)$ and $\hat{R}^i : O^i \times A^i \rightarrow \mathbb{R}$ to model the next observation and reward, respectively, based

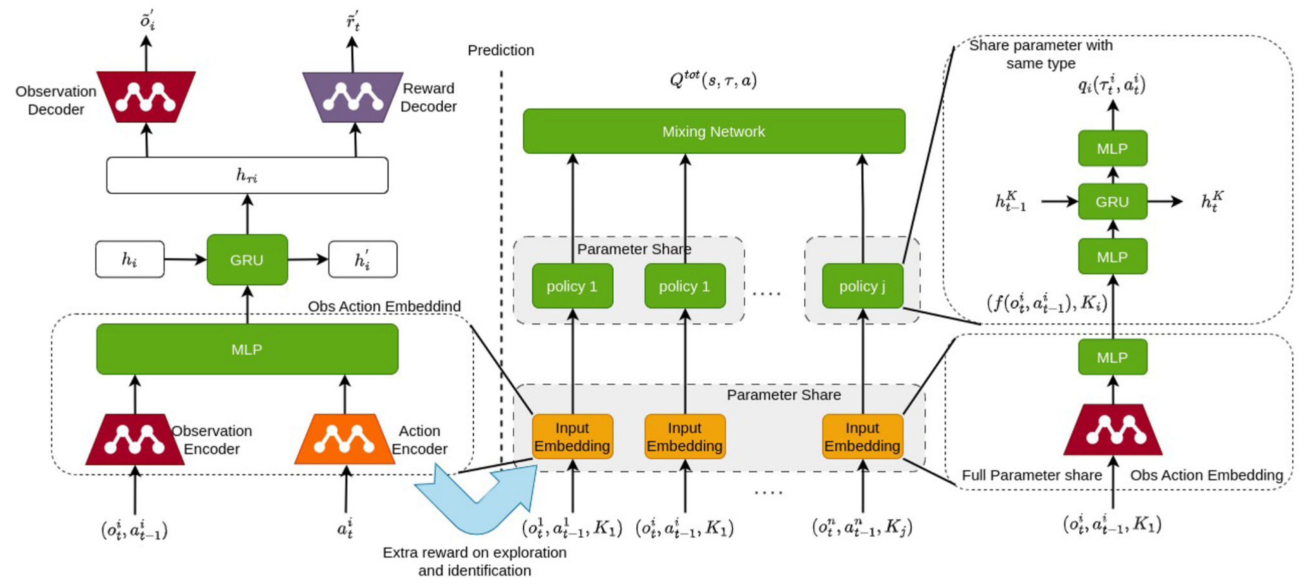


FIGURE 2
The framework of STSR includes a common representation derived from supervised learning, a semi-independent policy training scheme that applies full shared parameters among agents of the same type and hard share among different types, and an intrinsic/diversity-driven extra reward to encourage environment exploration and enhance the representation that can more clearly distinguish the inputs and outputs of different types.

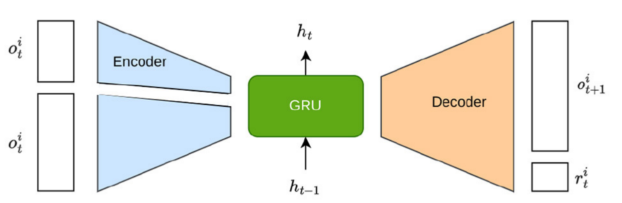


FIGURE 3
RNN-based prediction model from which to learn the observation and action representation embedding.

solely on the action and observation of an agent i. Our purpose in learning these functions is to ensure a wide acceptable input/output approximation and to establish an initial full share basis for a hard parameter sharing scheme for all the agents regardless of their types. Such prediction model training is due to be processed before the reinforcement learning while the full-parameter shared basis will be kept updated throughout the whole training process of reinforcement learning.

In our proposal, we introduce an encoder f_e and a decoder f_p, both parameterized by θ and depicted in Figure 3. The encoder is solely conditioned on the agent's identity. On the other hand, the decoder is split into an observation decoder, $f_{k_j}^o$, and a reward decoder, $f_{k_j}^r$, which receives the observation, action, and sampled encoding z of agent i and try to predict the next observation and reward. Unlike conventional autoencoders, o_t^i and a_t^i bypass the encoder and are only received by the decoder. As a result, due to the bottleneck, z can encode information only about the agent, such as its reward function $\hat{R}^i$ or observation transition model $\hat{\mathcal{O}}^i$.

To formalize the process, we assume that each agent's type denoted as k_j, represents its observation transition distribution and reward function. We also assume that both the agent's identity and its observation transition distribution can be projected in a latent space, z, through the posteriors $q(z|k_j)$ and $p(z|\operatorname{tr} = (o_{t+1}, o_t, r_t, a_t))$. The objective is to find the posterior $q(z|k_j)$.

The encoder-decoder model is trained with samples from all agents to learn from the experience of all agents, and it will represent the collection of the agent-centered transition and reward functions $\hat{\mathcal{P}}^i$ and $\hat{R}^i$ for all $i \in \mathcal{N}$. Given the inputs of the decoder, the information of the agent type can only pass through the sample z.

This model can be interpreted as a forward model, which is trained by minimizing the following loss function:

$$\mathcal{L}_e(\theta_e) = \mathbb{E}_{(\boldsymbol{o},\boldsymbol{a},r,\boldsymbol{o}')\sim\mathcal{D}}\Bigg[\sum_i \left\|f_o\left(\boldsymbol{z}_{a_i}, o_i, \boldsymbol{a}_{-i}\right) - o_i'\right\|_2^2 + \lambda_e \sum_i \left(f_r\left(\boldsymbol{z}_{a_i}, o_i, \boldsymbol{a}_{-i}\right) - r\right)^2\Bigg] \qquad (1)$$

where f_o and f_r are predictors for observations and rewards, respectively, and parameterized by θ_e. λ_e is a scaling factor, $\mathcal{D}$ is a replay buffer, and the sum is carried out over all agents.

Minimizing the model loss can be done prior to reinforcement learning. We sample actions $a^i \sim A^i$ and store the observed trajectories in a shared experience replay with all agents. We have empirically observed that the data required for this procedure is orders of magnitude less than what is usually required for reinforcement learning, and it can even be reused for training the policies, thus not adding to the sample complexity.

4.2. Intrinsic rewards for environment exploration and unit type identification

Multi-agent Reinforcement Learning (MARL) is an effective method for addressing complex decision-making challenges involving multiple agents, where external rewards are present. This approach enables agents to interact with the environment to make optimal decisions, motivated by rewards. A significant challenge for those designing agents is defining a suitable reward function for sequential decision-making tasks in Reinforcement Learning (RL). Additional potential-based rewards, besides extrinsic rewards, do not alter the order of agent behaviors. However, the choice of potential-based or policy-based reward function used to transform the original reward function can impact the sample and computational complexity of RL agents learning from experience in their environment. While this does not change the optimal policy, it can influence the learning process for better or worse.

The aforementioned representation can facilitate the designing of intrinsic rewards on 2 aspects: novelty rewards which encourage the agent to take extra effort on efficient environmental exploration and representability for diversity which can help to form representation more widely identify a different kind of agent.

One of the main challenges in RL is the trade-off between exploitation and exploration: agents must exploit the actions that they know lead to high rewards, but they must also explore new actions and states in order to discover new strategies that may lead to even higher rewards. The data distance between the forward prediction model can provide an additional source of motivation for exploration, beyond the extrinsic rewards provided by the environment.

For simplicity, we can define the state of the environment by combining the observation and rewards of all agents, which can be expressed as $s_t = \{(o_t^i, r_t^i), i \in \mathcal{N}\}, t = 0 \ldots \infty$. Let $d(r_1, r_2)$ be a distance metric between two representation vectors $r_1, r_2 \in \mathbb{R}^d$. One common distance metric is the Euclidean distance.

$$
\begin{aligned}
r_t^e = & \sum_i p_i^m \left\| f_o\left(z_{a_i}, o_t^i, \boldsymbol{a}_t^{-i}\right) - o_t^{i\prime} \right\|_2^2 \\
& + \lambda_e \sum_i p_i^m \left(f_r\left(z_{a_i}, o_t^i, \boldsymbol{a}_t^{-i}\right) - r_t^i \right)^2
\end{aligned} \tag{2}
$$

where p_i^m is weight when calculating Q_{tot} that we can obtain from mixer layer. The reward function 2 assigns a positive reward when the current state s_t is situated in a low-density region of the representation space. This low-density region indicates that the state is unique and hasn't been encountered by the agent before. The value of the reward is modified based on the discrepancy between the density estimate determined by the mixer function, denoted as p_m, and the overall density estimate. This normalization procedure guarantees that the reward stays within acceptable limits and does not become unreasonably high. As a result, we can determine the intrinsic reward of promoting environmental exploration. It is worth noting that the emphasis on exploration will decrease once the environment has been thoroughly explored. Therefore, we will introduce a discount factor that will gradually decrease during the training process.

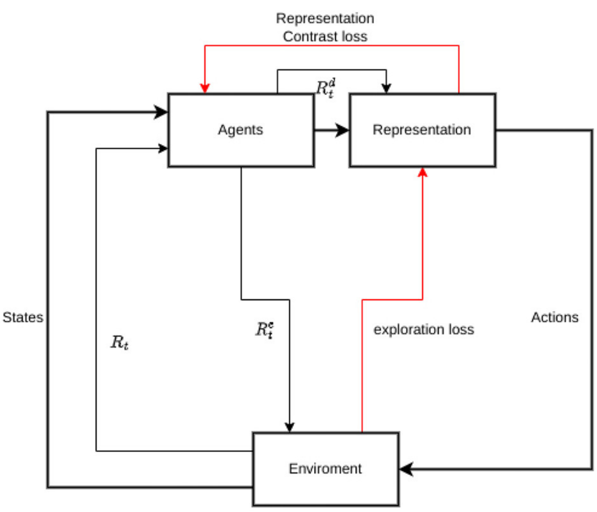


FIGURE 4
Total rewards including both intrinsic rewards to encourage environment exploration and to diverse representation upon agents' type. Thick black lines illustrate data flow, thin black lines illustrate rewards and red lines illustrate the loss to generate the intrinsic rewards.

We have incorporated an additional intrinsic reward to our design which aims to promote diversity in the representation of the agent's type. One of our key concepts is to implement a specialization policy for agents of the same type. To encourage this behavior, we implement an additional intrinsic reward system that incentives the agent to have similar representations for the same type when having the same kind of inputs and different representations for different types. In order to create a reliable representation-intrinsic reward, we utilize a method that involves calculating the average representation of agents that are of the same type when they receive a positive input. Conversely, we calculate the average representation of the different types of agents to serve as the negative input. By subtracting the negative reward from the positive reward, we obtain a final representation reward. This representation reward can be expressed in the following form:

$$
\begin{aligned}
r_t^d = & \sum_i p_i^m \left[\frac{1}{\mathcal{N}_i} \left(\left\| f_o\left(z_{a_i}, o_t^i, \boldsymbol{a}_t^{-i}\right) - o_t^{i\prime} \right\|_2^2 \right.\right. \\
& \left. + \lambda_e \left(f_r\left(z_{a_i}, o_t^i, \boldsymbol{a}_t^{-i}\right) - r_t^i \right)^2 \right) \\
& - \frac{\lambda_h}{\mathcal{N}_j} \left(\left\| f_o\left(z_{a_j}, o_t^i, \boldsymbol{a}_t^{-i}\right) - o_t^{i\prime} \right\|_2^2 \right. \\
& \left.\left. + \lambda_e \left(f_r\left(z_{a_j}, o_t^i, \boldsymbol{a}_t^{-i}\right) - r_t^i \right)^2 \right) \right], t = 0 \ldots \infty
\end{aligned} \tag{3}
$$

Our total reward after accounting for both these 2 intrinsic rewards is:

$$
r_t^{tot} = r_t + \lambda_e r_t^e + \lambda_d r_t^d, t = 0 \ldots \infty \tag{4}
$$

This representation, which is demonstrated in Figure 4, incentives intrinsic reward will be taken throughout the whole process of training accompanying the building up with the policy common representation basis.

4.3. Common representation based semi-independent policy training

The approach of full parameter sharing has shown remarkable achievements among homogeneous agents. Nevertheless, when extending it to a heterogeneous multi-agent environment, challenges arise regarding how to handle different types of agents with the same policy network that shares all parameters. This extension creates a dilemma since sharing parameters among agents with different characteristics can limit their potential and hinder policy optimization. On the other hand, avoiding parameter sharing altogether requires creating a complex decision-making system with multiple policy networks, each with isolated parameters. This alternative approach leads to slow convergence and inefficient use of experience.

In order to effectively address this issue, we propose utilizing full parameter sharing among agents of the same type, while applying semi-parameter sharing to agents of different types. Agents of the same type share inherent similarities, which enables them to be scaled up with a consistent range of decision-making capabilities. The success of parameter sharing among homogeneous agents supports its application among agents of the same type in a heterogeneous agent system, where the group of heterogeneous agents can be viewed as a collection of multiple sub-groups of homogeneous agents with varying types.

To well utilize the similarities between different sub-groups, we propose to apply hard parameter-sharing schemes. Hard parameter sharing is a technique used in multi-task learning, where a single neural network is trained to perform multiple tasks simultaneously by sharing some of its layers among the tasks. This approach can be effective and efficient because it allows the network to learn and generalize across multiple related tasks, while also reducing the total number of parameters needed to train the model.

Mathematically, hard parameter sharing can be represented as follows: Let x be the input to the network, y_1 and y_2 be the outputs of two related tasks, and f be the shared layers of the network. Then, the network can be represented as: $y_1 = g_1(f(x))$ and $y_2 = g_2(f(x))$ where g_1 and g_2 are task-specific output layers. In this way, the shared layers are trained to extract relevant features from the input that are useful for both tasks, while the task-specific output layers are trained to map these features to the desired outputs for each task. By sharing the parameters of the network across tasks, the model can learn to generalize better and improve performance on all tasks.

In the context of multi-agent reinforcement learning, hard parameter sharing can also be useful when different agents share common tasks or goals. For example, in a multi-agent scenario where agents must cooperate to achieve a common objective, such as in a game or robotics application, the agents may share some common knowledge or features that can be learned through a shared network. In our framework, multi-agent reinforcement learning with hard-parameter sharing can be expressed as:

$$\max_{\mathbf{a}_i} Q(s, \mathbf{a}_i) = \max_{\mathbf{a}_i} \sum_{t=0}^{\infty} \sum_{i=1}^{n} \mathbb{E}_{\pi_j} Q_i(f(s_t, \mathbf{a}_i); \theta_j) \tag{5}$$

f denotes the shared layers employed for hard-parameter sharing, while $\pi_{j=1\ldots k}$ represents the policies employed for all agents, where agents of the same type apply the identical policy with full parameter shared.

As illustrated in Figure 5, our proposed shared layer embedding is identical to the common representation latent. The representation latent handles all agent inputs and outputs regardless of unit type, reflecting its parameter-sharing is applicable among all agents. In this case, the representation latent can be selected as the shared layer, initialized with its current parameters. Empirically, this shared layer can be deemed as a separate branch of the common representation, training via unit type based reinforcement learning with the purpose to maximize overall value. In the experiments section, we can prove the shared layer updated with the reinforcement learning outperforms the one updated with the representation latent. Meanwhile, the representation latent is under training with the environment predictor for better representation.

5. Experiment and results

In this section, we thoroughly evaluate our proposed method from various perspectives. Firstly, we provide a comprehensive assessment of its overall performance in different scenarios and compare it with other mainstream algorithms to gauge its effectiveness. Secondly, we conduct experiments with different alternative flows and perform ablation studies to assess the impact of each component. Thirdly, we conduct a detailed analysis of the intermediate results to gain a better understanding of the underlying principles, including the initial representation and its subsequent updates, their distribution and representativeness, shared layers, and the course of its training. Finally, we attempt to validate the framework's generalizability by testing its representation transferability and its curriculum learning capacity.

5.1. Experiment setting

We have chosen the StarCraft II micromanagement (SMAC) benchmark (Samvelyan et al., 2019) as our test-bed due to its rich environments and high complexity of control. The SMAC benchmark presents a series of challenging tasks, as agents must learn policies in a large action space that includes four cardinal directions, stop, take noop, or select an enemy to attack at each time step. If there are n_e enemies in the map, each ally unit's action space contains $n_e + 6$ discrete actions. SMAC environment is rich in all kinds of settings including a lot of homogeneous agents. It is also a widely used setting where multiple agents of distinct types coexist and must learn together, for which our proposed method is mainly focused. The `MMM2` is an example of

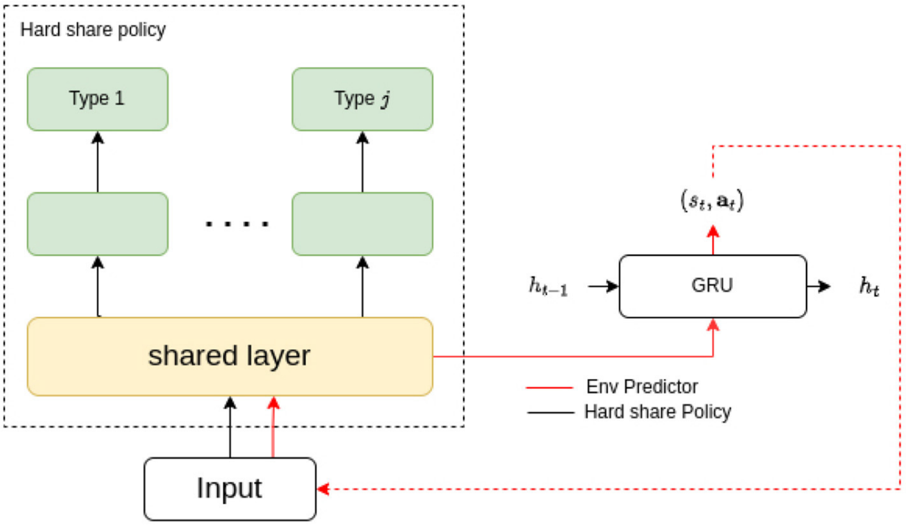


FIGURE 5
The flow and structure for the hard-parameter-sharing based policy generation scheme, where the initial parameter is from the supervised env prediction and will be later updated along with the type/agent layer.

such an environment that contains three types of units (marines, marauders, and medivacs) with distinct attributes. One of the unit types medivacs is particularly different, as it needs to learn how to heal friendly units instead of attacking enemies.

Although our proposal is mainly concerned with heterogeneous agents, it is quite capable to handle all kinds of environments. To conduct a full assessment of our proposal, we carry out tests on all kinds of settings, respectively, regardless of either homogeneous agents or heterogeneous agents settings and we compare the improvements in different settings.

SMAC consists of various maps which have been classified as *easy*, *hard*, and *super hard*. It also contains variate group agents of homogeneous or heterogeneous. Even though our main proposal is aimed at heterogeneous scenarios, the method is also applicable to the homogeneous and can also outperform its original method.

To fully evaluate its overall performance on different scenarios, we have conducted a thorough evaluation of our approach by benchmarking it across all 14 scenarios within the SMAC suite. This allows us to assess its performance across a range of settings. Additionally, we present some of the results obtained from this evaluation. Furthermore, we have compared our proposal with other value-based MARL algorithms that are considered state-of-the-art, including VDN (Sunehag et al., 2018), QMIX (Rashid et al., 2020), QPLEX (Wang et al., 2021), some role-based MARL method including ROMA (Wang et al., 2020a), and RODE (Wang et al., 2020b) and an agent-specific modules based parameter-sharing algorithm CDS (Li et al., 2021).

To better understand the contribution of each component, we conducted an ablation study by comparing the performance with and without various components. This series of tests were assigned different names: *STSR full* denotes the setting where all components were included, *STSR No Representation Learning* excluded the common representation as the basis for hard-parameter sharing, instead using a random basis initially. The *STSR No Representation Later-update* setting did not update or learn the hard-parameter sharing basis but only utilized the initial common representation. Additionally, we examined the settings of *STSR No r_e Reward* and *STSR No r_d Reward*, which, respectively, excluded the exploration reward and representation reward. Finally, the *STSR No Hard-Parameter-Share* setting did not apply the hard parameter sharing scheme and did not share parameters among different types of agents.

In the next section, we will present and discuss the results of these thorough evaluation and ablation tests.

5.2. Results and discussion

5.2.1. Overall performance

To assess the performance of the models or algorithms, the experiments in this section were conducted 4 times using different random seeds. The median performance is reported as performance metrics. These metrics provide a comprehensive understanding of the models or algorithms' performance and account for the variability that can occur due to stochasticity.

We conducted a comprehensive evaluation of our approach by benchmarking it across all 14 scenarios, categorized in Table 1. Due to space limitations, we present examples of one easy map (`3s vs. 5z`) and all the super hard maps in Figure 6. Among the tests presented, our proposed method STSR demonstrated the best performance in scenarios `3s5z vs. 3s6z` and `MMM2`, and ranked second in scenarios `3s5z` and `27m vs. 30m`. These results are not surprising, as our proposal primarily focuses on heterogeneous agents' settings. Compared to role-based methods that cluster agents based on their properties CDS (Li et al., 2021) which seeks to achieve the maximum diversity among

TABLE 1 Categories of the SMAC scenarios and their corresponding difficulties, ally units, and agents type.

Difficulties	Name	Ally units	Agents type
Easy	*2s3z*	2 Stalkers & 3 Zealots	Heterogeneous
	3s5z	3 Stalkers & 5 Zealots	Heterogeneous
	1c3s5z	1 Colossus, 3 Stalkers & 5 Zealots	Heterogeneous
	5 m_vs._6m	5 Marines	Homogeneous
	10 m_vs._11m	10 Marines	Homogeneous
Hard	2s_vs._1sc	2 Stalkers	Homogeneous
	3s_vs._5z	3 Stalkers	Homogeneous
	2c vs. 64zg	2 Colossi	Homogeneous
	Bane vs. bane	20 Zerglings & 4 Banelings	Heterogeneous
Super hard	3s5z_vs_3s6z	3 Stalkers & 5 Zealots	Heterogeneous
	6h^{-}vs. 8*z*	6 Hydralisks	Homogeneous
	27 m_*vs*_30 m	27 Marines	Homogeneous
	Corridor	6 Zealots	Homogeneous
	MMM2	1 Medivac, 2 Marauders & 7 Marines	Heterogeneous

individualized behaviors from the shared network., our proposal outperforms in heterogeneous settings, particularly in the speed of convergence. Clustering agents of the same kind and sharing parameters among them is a natural choice. We believe that our proposed agent clustering method is more stable and consistent, enabling more efficient use of generated experience to train policy networks. In contrast, role-based methods may require more interactions with the environment to better understand the agents' properties and assign roles, which may cause a delay in convergence. The size of the agents in these 2 scenarios may be well-suited for our proposed method. In the map 3s5z vs. 3s6z there are 3 Stalkers and 5 Zealots, while in the map MMM2 there is 1 Medivac, 2 Marauders, and 7 Marines. The size of each agent type is not too large or too small, making it appropriate to share the same type of parameters. In contrast, in the map bane vs. bane there are 20 Zerglings and four Banelings. The size of the Zerglings is too large and may require clustering in advance. One surprising outlier is the *easy* scenario 3s5z for which QPLEX exhibits the best performance, surpassing our proposal and the role-based method by a large margin. We hypothesize that this is because these maps do not require significant exploration or distributed policy training. The limited experience can be better utilized by training on a single, fully-parameter-shared network.

In contrast to achieving the best performance on heterogeneous agent scenarios, our proposed STSR is less efficient in homogeneous agent settings compared to its counterparts from role-based algorithms such as RODE (Wang et al., 2020b) and ROMA (Wang et al., 2020a), and diversity oriented parameter sharing algorithm CDS (Li et al., 2021). Role-based algorithms employ different principles in clustering small groups of agents automatically and then apply role-based policies to improve the overall performance, whileour approach relies purely on the agents' unit types. CDS (Li et al., 2021) leverages information-theoretical regularization to maximize the mutual information between agents' identities and their trajectories with the purpose to promote learning sharing among agents while keeping necessary diversity. Thus for the scenarios with homogeneous agents which cannot be clustered and achieve sufficient diversity from the environmental exploration and agents behavior the performance of our approach is comparatively lower than the aforementioned counterparts. Empirically, we have observed that the performance on scenarios with homogeneous agents can be enhanced by employing random clustering as an initial step. We plan to conduct a detailed investigation of this phenomenon in our forthcoming research on clustering size, the initial settings, etc.

Our method introduces a hierarchical parameter sharing scheme, wherein parameters are fully shared among agents of the same type and partially shared among agents of different types through hard parameter sharing. By sharing parameters, agents can leverage each other's experiences and exploit common patterns in the environment based on their similarities. This approach simplifies training and enables efficient knowledge transfer. In contrast, role-based MARL assigns specific roles or tasks to individual agents, defining their unique responsibilities and objectives. Each agent possesses its own set of parameters optimized for fulfilling its designated role. Roles can be predefined or learned during training. This approach fosters specialization and coordination among agents, as they concentrate on specific tasks or functions. While role-based MARL excels in handling complex scenarios and adapting to diverse environments, it may necessitate more intricate training algorithms and coordination mechanisms. CDS (Li et al., 2021) propose an information-theoretical regularization to maximize the mutual information between agents' identities and their trajectories, which encourages extensive exploration and diverse individualized behaviors. It introduce agent-specific modules in the shared neural network architecture, which are regularized by L1-norm to promote learning sharing among agents while keeping necessary diversity. Compared to our proposed STSR and role-based methods, CDS (Li et al., 2021) allows for more flexibility in fostering agent specialization and achieving diversity in individualized behaviors. However, without clustering-based group tactics, it results in low efficient utilization of experience.

5.2.2. Ablation study

To better understand the contributions of each component, we conducted an ablation study on three scenarios with the best performance: 27m vs. 30m, 3s5z vs. 3s6z, and MMM2. Among these scenarios, 3s5z vs. 3s6z and MMM2 are heterogeneous, while 27m vs. 30m is homogeneous. The performance of the ablation study can be viewed in Figure 7. According to the results we presented, all components make positive contributions to the overall performance. Among all the curves, *STSR No Representation Later-update* had the worst performance, implying that the original representation from

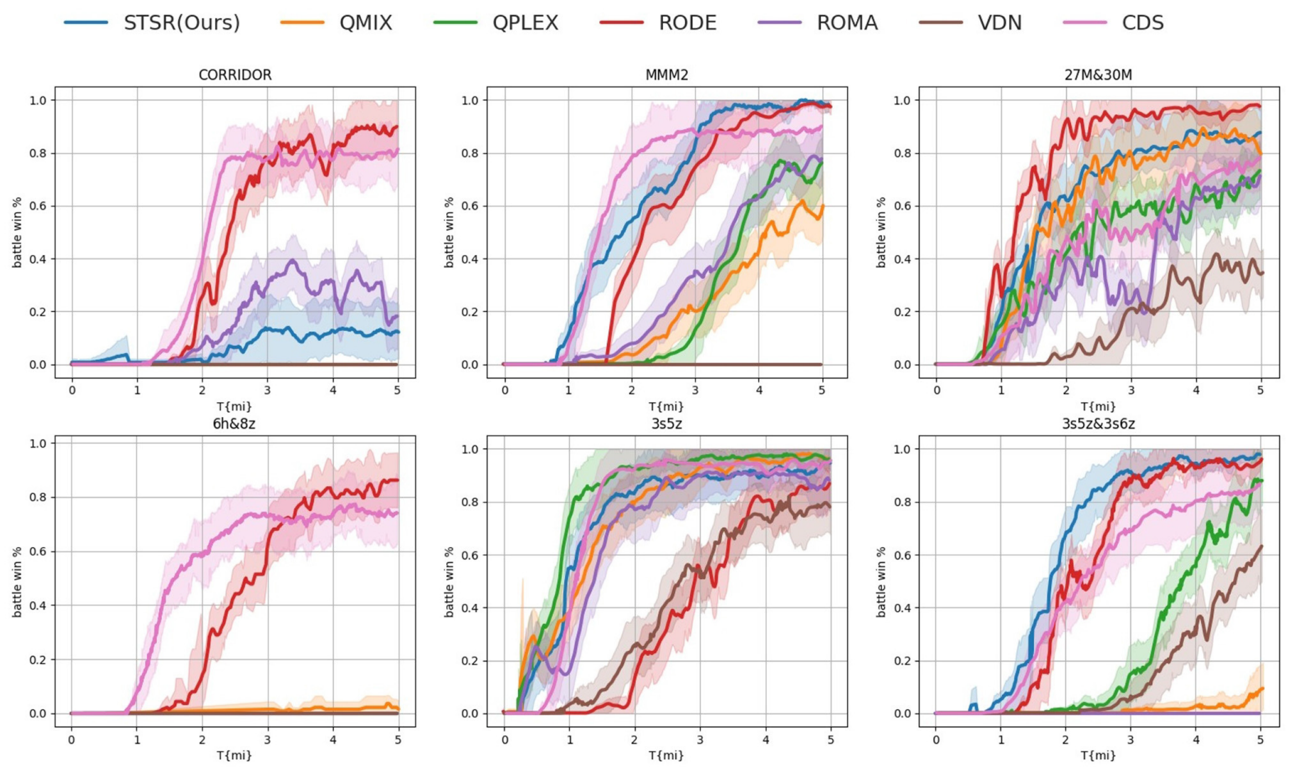


FIGURE 6
Performance comparison with baselines on all super hard maps and one easy map (3s5z). The baselines compromise VDN (Sunehag et al., 2018), QMIX (Rashid et al., 2020), QPLEX (Wang et al., 2021), role based algorithms ROMA (Wang et al., 2020a), RODE (Wang et al., 2020b), and CDS (Li et al., 2021).

supervised learning is not sufficient for a hard-parameter sharing basis, and a later updated data procedure is necessary. Meanwhile, the curves *STSR No Representation Learning* are not as good as *STSR full* on all scenarios, which means that even if an initial value settled on the hard-parameter sharing basis may not be sufficient, it can still help to quickly approach the proper basis. For the scenario `27m vs. 30m`, there is no difference in performance between *STSR full*, *STSR No Hard-Parameter-Share*, and *STSR No r_d Reward*. This result is not surprising since these two components mainly work for heterogeneous agents, and `27m vs. 30m` is a homogeneous scenario. The comparison between *STSR No r_d Reward* and *STSR full* on the other two scenarios shows that the application of r_d can help to more quickly approach the hard-parameter sharing layer, especially at the beginning. Such a contribution is decreased following the later update of the sharing layer. The reward r_e can help achieve better performance on all scenarios regardless of whether they are homogeneous or heterogeneous (presented on *STSR No r_e Reward*), by encouraging environment exploration. The performance of *STSR full* suggests that the utilization of hard-parameter sharing may not approach the capability to largely improve performance, but it does speed up the training process.

In conclusion, the ablation study found that all components make positive contributions to overall performance. The study also showed that even an initial value settled on the hard-parameter sharing basis may not be sufficient, but it can still help to quickly approach the proper basis. The utilization of hard-parameter sharing may not largely improve performance, but it does speed up the training process. The application of r_d can help to more quickly approach the hard-parameter sharing layer, especially at the beginning, and such a contribution decreases following the later update of the sharing layer. The reward r_e can help achieve better performance in all scenarios by encouraging environmental exploration.

5.3. Diverged representation embedding training

In our proposal, we introduce a novel approach for representation embedding. Initially generated through self-supervised learning, the representation embedding is duplicated and diverged into two branches. One branch is updated using reinforcement learning to handle the observation for RL, while the other branch is continuously updated to guide the intrinsic reward. Although these two branches serve different purposes, they function similarly to the representation of the common agent's observation and action. To gain a deeper understanding of the functions and capabilities of these two embedding representations, we conducted an experiment comparing their centralization and clustering properties. To achieve this, we projected the embeddings onto a 2D space, as depicted in Figure 8, using the scenario `MMM2` as an illustrative example.

In the `MMM2` scenario, which comprises a heterogeneous composition of 1 Medivac, 2 Marauders, and 7 Marines facing 1 Medivac, 3 Marauders, and 8 Marines, it is essential to

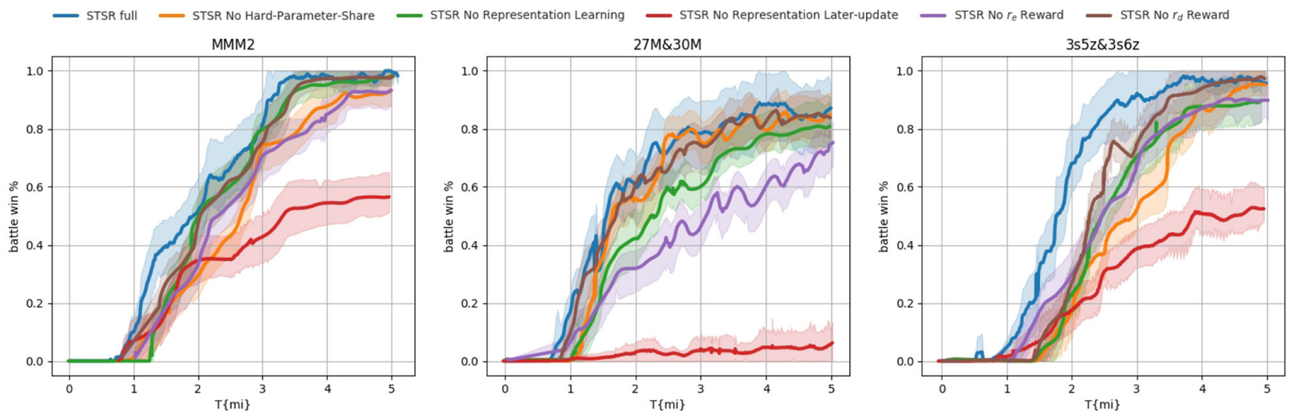


FIGURE 7
Ablation study on 3 best-performing scenarios. All components make positive contributions to the overall performance.

foster effective cooperation among different agent types to fully exploit the advantages of each unit. Our observation revealed that while both forms of embedding clustered within their respective groups, their concentration levels varied. This indicates that both embeddings are capable of effectively distinguishing the observations and actions of different unit types, albeit with varying degrees of concentration, resulting in distinct functional characteristics.

The self-supervised embedding, which is supervised by self-supervised learning, exhibited a higher level of concentration, while the RL-led embedding showed slightly more diversity among individual points. We hypothesize that the self-supervised embedding prioritizes forming distinctive representations for each unit type, reinforced by intrinsic rewards. Hence, the dense concentration in the self-supervised embedding as a result of this objective. On the other hand, the RL embedding focuses on obtaining maximum rewards, the distinctiveness of representation for each individual agent will access a more proper reaction for each agent. Therefore, the RL embedding aims to strike a balance between representing the unit type and the individual agent's characteristics.

5.3.1. Representation transferability and curriculum learning

In this paper, we propose a method that can transfer learned policies to new agents without requiring the entire system to be retrained. This is achieved by duplicating common representations and sharing parameters among agents of the same type. An additional benefit of this approach is that it can be easily applied to tasks involving curriculum learning, where agents of different types are gradually introduced. To accomplish this, we first identify the type of the incoming agents, then average the outputs of agents of the same type from the updated representation. Next, we duplicate the policy parameters of the agent type and apply them to the new agent. By duplicating policies and representations, we ensure that learned policies can be transferred to tasks with varying numbers of agents. This makes our proposed method versatile and applicable to a wide range of tasks without the need for additional training.

We evaluated the transferability of our method on the SMAC benchmark by sorting allies and enemies based on their relative distances to an agent and including information on the nearest ones while keeping the observation length fixed. Figure 9 shows the win rates of the policy learned from the map `3s5z vs. 3s6z` on various maps without further policy training. In the original task, 3 Stalkers and 5 Zealots face 3 Stalkers and 6 Zealots. We designed 2 types of maps which, respectively, increased the number of Stalkers and Zealots for both the number of allies and enemies to test the transferability of different agents' types. We observed that the transferability of STSR was evident from the learned policy and still has a good performance on new maps especially when both sides increase their agents' numbers. Additionally, our proposed method is easy to extend for the transferring to the increased size of agents which may help to provide a promising result in curriculum learning.

6. Conclusion and future work

Overall, this research provides a fresh perspective on addressing the challenges of parameter sharing in multi-agent reinforcement learning, particularly in heterogeneous environments. The proposed approach not only enables agents to learn from each other but also improves the overall performance of the system. These contributions allow for specialization among heterogeneous agents while still promoting experience sharing, and make it easier to incorporate the hard-parameter-sharing scheme. The proposed method outperforms current mainstream algorithms, particularly for heterogeneous agents, and can be considered a more general and fundamental structure for heterogeneous agent reinforcement learning. Our work is the first to introduce a multi-task network parameter-sharing scheme to MARL and to utilize a supervised learning method for generating a shared input/output representation. Additionally, our proposed intrinsic reward is based on the prediction of supervised learning and its input/output representation, which can stimulate more exploration and enhance the representability of this representation without requiring extra effort. Overall, our contributions provide a promising direction for addressing the challenges

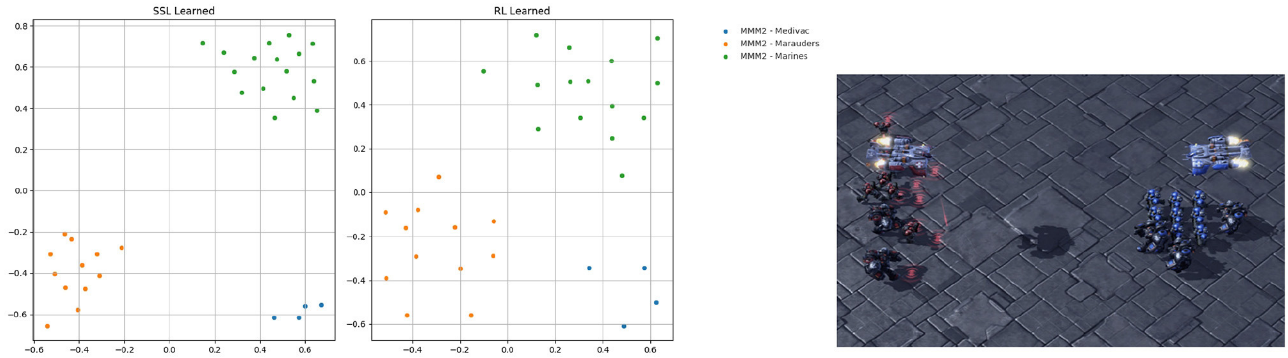
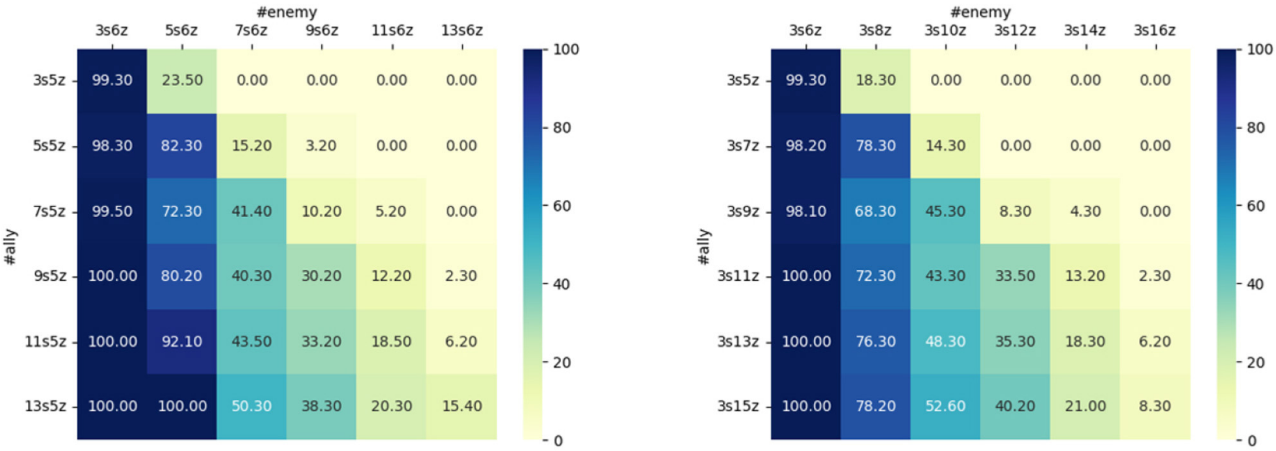


FIGURE 8
Visualization of representation embeddings of scenario MMM2 projected to 2D space on step 2500, updated, respectively, through ever-existing self-supervised learning and RL. There are 3 kinds of points representing 3 kinds of agents type for scenario MMM2, which are marines, marauders, and medivac.



#ally \ #enemy	3s6z	5s6z	7s6z	9s6z	11s6z	13s6z
3s5z	99.30	23.50	0.00	0.00	0.00	0.00
5s5z	98.30	82.30	15.20	3.20	0.00	0.00
7s5z	99.50	72.30	41.40	10.20	5.20	0.00
9s5z	100.00	80.20	40.30	30.20	12.20	2.30
11s5z	100.00	92.10	43.50	33.20	18.50	6.20
13s5z	100.00	100.00	50.30	38.30	20.30	15.40

#ally \ #enemy	3s6z	3s8z	3s10z	3s12z	3s14z	3s16z
3s5z	99.30	18.30	0.00	0.00	0.00	0.00
3s7z	98.20	78.30	14.30	0.00	0.00	0.00
3s9z	98.10	68.30	45.30	8.30	4.30	0.00
3s11z	100.00	72.30	43.30	33.50	13.20	2.30
3s13z	100.00	76.30	48.30	35.30	18.30	6.20
3s15z	100.00	78.20	52.60	40.20	21.00	8.30

FIGURE 9
Transferability on the unseen maps on 3s5z vs. 3s6z without further training on the new maps.

in MARL and improving performance for heterogeneous agents.

Based on our experiments, it was observed that one of the bottlenecks in our work is its focus solely on scenarios with heterogeneous agents. It is not well-suited to scenarios with homogeneous agents, and even for heterogeneous scenarios with a large group of the same kind of agents. In comparison with role-based MARL methods, a smaller clustered parameter-sharing group is required. We have empirically noted that a random clustering of homogeneous agents can outperform the baselines and our proposed work. For our future work, we plan to conduct further research to gain a better understanding of the principles behind this observation and make appropriate improvements to the parameter-sharing groups.

Data availability statement

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

Author contributions

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

Funding

This work was supported in part by the National Key R&D Program of China (Grant No. 2021ZD0112700) and the Natural Science Basic Research Plan in Shaanxi Province of China (Project Codes: 2021JQ-061 and 2021JM-020).

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

Boutilier, C. (1996). "Planning, learning and coordination in multiagent decision processes," in *TARK* (De Zeeuwse Stromen: Citeseer), 195–210.

Chen, D., Li, Z., Wang, Y., Jiang, L., and Wang, Y. (2021). Deep multi-agent reinforcement learning for highway on-ramp merging in mixed traffic. *arXiv preprint arXiv:2105.05701*. doi: 10.48550/arXiv.2105.05701

Christianos, F., Papoudakis, G., Rahman, M. A., and Albrecht, S. V. (2021). "Scaling multi-agent reinforcement learning with selective parameter sharing," in *International Conference on Machine Learning* (PMLR), 1989–1998.

Chu, X., and Ye, H. (2017). Parameter sharing deep deterministic policy gradient for cooperative multi-agent reinforcement learning. *arXiv preprint arXiv:1710.00336*. doi: 10.48550/arXiv.1710.00336

Devin, C., Abbeel, P., Darrell, T., and Levine, S. (2018). "Deep object-centric representations for generalizable robot learning," in *2018 IEEE International Conference on Robotics and Automation (ICRA)* (Brisbane, QLD: IEEE), 7111–7118.

Finn, C., Tan, X. Y., Duan, Y., Darrell, T., Levine, S., and Abbeel, P. (2016). "Deep spatial autoencoders for visuomotor learning," in *2016 IEEE International Conference on Robotics and Automation (ICRA)* (Stockholm: IEEE), 512–519.

Foerster, J., Farquhar, G., Afouras, T., Nardelli, N., and Whiteson, S. (2018). "Counterfactual multi-agent policy gradients," in *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32 (New Orleans, LA: MIT).

Fujimoto, S., Hoof, H., and Meger, D. (2018). "Addressing function approximation error in actor-critic methods," in *International Conference on Machine Learning* (Stockholm: PMLR), 1587–1596.

Guo, Z. D., Azar, M. G., Piot, B., Pires, B. A., and Munos, R. (2018). Neural predictive belief representations. *arXiv preprint arXiv:1811.06407*. doi: 10.48550/arXiv.1811.06407

Guo, Z. D., Pires, B. A., Piot, B., Grill, J.-B., Altché, F., Munos, R., et al. (2020). "Bootstrap latent-predictive representations for multitask reinforcement learning," in *International Conference on Machine Learning* (PMLR), 3875–3886.

Gupta, J. K., Egorov, M., and Kochenderfer, M. (2017). "Cooperative multi-agent control using deep reinforcement learning," in *Autonomous Agents and Multiagent Systems: AAMAS 2017 Workshops, Best Papers* (São Paulo: Springer), 66–83.

Ha, D., and Schmidhuber, J. (2018). World models. *arXiv preprint arXiv:1803.10122*.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. (2018). "Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International Conference on Machine Learning* (Stockholm: PMLR), 1861–1870.

Hessel, M., Modayil, J., Van Hasselt, H., Schaul, T., Ostrovski, G., Dabney, W., et al. (2018). "Rainbow: combining improvements in deep reinforcement learning," in *Proceedings of the AAAI Conference on Artificial Intelligence* (New Orleans, LA: MIT).

Hessel, M., Soyer, H., Espeholt, L., Czarnecki, W., Schmitt, S., and van Hasselt, H. (2019). "Multi-task deep reinforcement learning with popart," in *Proceedings of the AAAI Conference on Artificial Intelligence* (Honolulu, HI: MIT), 3796–3803.

Hüttenrauch, M., Šošić, A., and Neumann, G. (2017). Guided deep reinforcement learning for swarm systems. *arXiv preprint arXiv:1709.06011*. doi: 10.48550/arXiv.1709.06011

Jaderberg, M., Mnih, V., Czarnecki, W. M., Schaul, T., Leibo, J. Z., Silver, D., et al. (2016). Reinforcement learning with unsupervised auxiliary tasks. *arXiv preprint arXiv:1611.05397*. doi: 10.48550/arXiv.1611.05397

Kingma, D. P., and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*. doi: 10.48550/arXiv.1312.6114

Kipf, T., Van der Pol, E., and Welling, M. (2019). Contrastive learning of structured world models. *arXiv preprint arXiv:1911.12247*. doi: 10.48550/arXiv.1911.12247

Laskin, M., Srinivas, A., and Abbeel, P. (2020). "CURL: contrastive unsupervised representations for reinforcement learning," in *International Conference on Machine Learning* (PMLR), 5639–5650.

Li, C., Wang, T., Wu, C., Zhao, Q., Yang, J., and Zhang, C. (2021). "Celebrating diversity in shared multi-agent reinforcement learning," in *Proceedings of the 35th International Conference on Neural Information Processing Systems* (ACM), 3991–4002.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., et al. (2015). Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*. doi: 10.48550/arXiv.1509.02971

Long, M., Cao, Z., Wang, J., and Yu, P. S. (2017). "Learning multiple tasks with multilinear relationship networks," in *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, CA: ACM).

Lowe, R., Wu, Y. I., Tamar, A., Harb, J., Pieter Abbeel, O., and Mordatch, I. (2017). "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, CA: ACM), 6382–6393.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., et al. (2016). "Asynchronous methods for deep reinforcement learning," in *International Conference on Machine Learning* (New York, NY: PMLR), 1928–1937.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., et al. (2013). Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*. doi: 10.48550/arXiv.1312.5602

Oliehoek, F. A. (2012). "Decentralized POMDPs," in *Reinforcement Learning: State-of-the-Art*, eds. M. Wiering and M. van Otterlo (Berlin: Springer), 471–503. doi: 10.1007/978-3-642-27645-3_15

Oord, A. V. D., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*. doi: 10.48550/arXiv.1807.03748

Rashid, T., Samvelyan, M., De Witt, C. S., Farquhar, G., Foerster, J., and Whiteson, S. (2020). Monotonic value function factorisation for deep multi-agent reinforcement learning. *J. Mach. Learn. Res.* 21, 7234–7284.

Ruder, S. (2017). An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*. doi: 10.48550/arXiv.1706.05098

Ruder, S., Bingel, J., Augenstein, I., and Søgaard, A. (2019). "Latent multi-task architecture learning," in *Proceedings of the AAAI Conference on Artificial Intelligence* (Honolulu, HI: MIT), 4822–4829.

Samvelyan, M., Rashid, T., de Witt, C. S., Farquhar, G., Nardelli, N., Rudner, T. G. J., et al. (2019). The StarCraft multi-agent challenge. *arXiv preprint arXiv:1902.04043*. doi: 10.48550/arXiv.1902.04043

Singh, A. J., Kumar, A., and Lau, H. C. (2020). "Hierarchical multiagent reinforcement learning for maritime traffic management," in *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020)* (Auckland).

Sunehag, P., Lever, G., Gruslys, A., Czarnecki, W. M., Zambaldi, V., Jaderberg, M., et al. (2018). "Value-decomposition networks for cooperative multi-agent learning based on team reward," in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems* (Stockholm: ACM), 2085–2087.

Tan, M. (1993). "Multi-agent reinforcement learning: Independent vs. cooperative agents," in *Proceedings of the Tenth International Conference on Machine Learning* (Amherst, MA: PMLR), 330–337.

Wang, J., Ren, Z., Liu, T., Yu, Y., and Zhang, C. (2021). "Qplex: duplex dueling multi-agent q-learning. in *International Conference on Learning Representations* (Vienna).

Wang, T., Dong, H., Lesser, V., and Zhang, C. (2020a). "Roma: multi-agent reinforcement learning with emergent roles," in *Proceedings of the 37th International Conference on Machine Learning* (PMLR), 9876–9886.

Wang, T., Gupta, T., Mahajan, A., Peng, B., Whiteson, S., and Zhang, C. (2020b). Rode: learning roles to decompose multi-agent tasks. *arXiv preprint arXiv:2010.01523*. doi: 10.48550/arXiv.2010.01523

Wang, T., Wang, J., Zheng, C., and Zhang, C. (2019). Learning nearly decomposable value functions via communication minimization. *arXiv preprint arXiv:1910.05366*. doi: 10.48550/arXiv.1910.05366

Wilson, A., Fern, A., and Tadepalli, P. (2010). "Bayesian policy search for multi-agent role discovery," in *Proceedings of the AAAI Conference on Artificial Intelligence* (Atlanta, GA: MIT), 624–629.

Yan, W., Vangipuram, A., Abbeel, P., and Pinto, L. (2021). "Learning predictive representations for deformable objects using contrastive estimation," in *Conference on Robot Learning* (London: PMLR), 564–574.

Yu, J., Vincent, J. A., and Schwager, M. (2022). DiNNO: distributed neural network optimization for multi-robot collaborative learning. *IEEE Robot. Automat. Lett.* 7, 1896–1903. doi: 10.1109/LRA.2022.3142402

Zhang, C., and Lesser, V. (2011). "Coordinated multi-agent reinforcement learning in networked distributed POMDPs," in *Twenty-Fifth AAAI Conference on Artificial Intelligence* (San Francisco, CA: MIT).

Zheng, L., Yang, J., Cai, H., Zhou, M., Zhang, W., Wang, J., et al. (2018). "Magent: a many-agent reinforcement learning platform for artificial collective intelligence," in *Proceedings of the AAAI Conference on Artificial Intelligence* (New Orleans, LA).

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 26 June 2023
DOI 10.3389/fnins.2023.1209132

OPEN ACCESS

EDITED BY
Zhengwang Wu,
University of North Carolina at Chapel Hill,
United States

REVIEWED BY
Zhi Li,
East China Normal University, China
Rui Liu,
Dalian University, China
Hong Zhu,
Jiangsu University, China

*CORRESPONDENCE
Liyan Ma
✉ liyanma@shu.edu.cn

RECEIVED 20 April 2023
ACCEPTED 25 May 2023
PUBLISHED 26 June 2023

CITATION
Cai B, Ma L and Sun Y (2023) Dual consistent pseudo label generation for multi-source domain adaptation without source data for medical image segmentation.
Front. Neurosci. 17:1209132.
doi: 10.3389/fnins.2023.1209132

Dual consistent pseudo label generation for multi-source domain adaptation without source data for medical image segmentation

Binke Cai, Liyan Ma* and Yan Sun

School of Computer Engineering and Science, Shanghai University, Shanghai, China

Introduction: Unsupervised domain adaptation (UDA) aims to adapt a model learned from the source domain to the target domain. Thus, the model can obtain transferable knowledge even in target domain that does not have ground truth in this way. In medical image segmentation scenarios, there exist diverse data distributions caused by intensity in homogeneities and shape variabilities. But multi source data may not be freely accessible, especially medical images with patient identity information.

Methods: To tackle this issue, we propose a new multi-source and source-free (MSSF) application scenario and a novel domain adaptation framework where in the training stage, we only get access to the well-trained source domain segmentation models without source data. First, we propose a new dual consistency constraint which uses domain-intra and domain-inter consistency to filter those predictions agreed by each individual domain expert and all domain experts. It can serve as a high-quality pseudo label generation method and produce correct supervised signals for target domain supervised learning. Next, we design a progressive entropy loss minimization method to minimize the class-inter distance of features, which is beneficial to enhance domain-intra and domain-inter consistency in turn.

Results: Extensive experiments are performed for retinal vessel segmentation under MSSF condition and our approach produces impressive performance. The sensitivity metric of our approach is highest and it surpasses other methods with a large margin.

Discussion: It is the first attempt to conduct researches on the retinal vessel segmentation task under multi-source and source-free scenarios. In medical applications, such adaptation method can avoid the privacy issue. Furthermore, how to balance the high sensitivity and high accuracy need to be further considered.

KEYWORDS

unsupervised domain adaptation, retinal vessel segmentation, semantic segmentation, multi-source, source-free

1. Introduction

Retinal diseases such as glaucoma and diabetic retinopathy often lead to blindness (Wu et al., 2021). It has been estimated that the risk of retinal-related diseases has increased greatly due to increasing pressure, lifestyle changes, and other potential factors. Such a trend pushes more and more researchers dedicated in exploring computer-aided diagnosis (CAD) systems for automatic and accurate diagnosis of retinal pathologies. It is of great significance for CAD systems to segment retinal vessels accurately because the segmentation result can provide the dependable diagnosis basis for examination of retinal diseases. Although there are a great

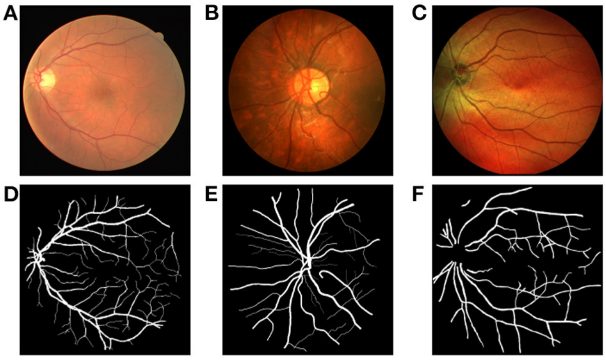


FIGURE 1
Example of fundus images and ground truths from the DRIVE, CHASEDB1, and IOSTAR datasets, respectively. **(A)** Fundus image 1 from DRIVE, **(B)** Fundus image 2 from CHASEDB1, **(C)** Fundus image 3 from IOSTAR, **(D)** ground truth of Fundus image 1, **(E)** ground truth of Fundus image 2, **(F)** ground truth of Fundus image 3.

quantity of classical model based methods about retinal vessel segmentation such as hand-crafted filters and fully connected conditional random fields (CRFs) (Orlando et al., 2016), it still remains challenging due to the large variation of the size of vessels, inhomogeneous lighting conditions, and other interference factors.

Semantic segmentation is one of the hot and widespread concerned topics in computer vision field, which aims to classify each pixel correctly in the whole image. Unprecedented advances in the semantic segmentation technique have been possible owning to the rapid development of convolutional neural networks (CNNs) and the availability of large-scale datasets. CNNs have outstanding ability to provide powerful and meaningful feature representations for medical image segmentation. Guo et al. (2023) proposed a new transformer framework based on CNN with parallax fusion paths for stereo image super-resolution. But there exists an obvious defect in the training process of supervised models that they entail a large training dataset equipped with labor-intensive annotations. The supervised models inevitably face challenges when they deal with new samples that correspond to different distributions with training samples. In medical image segmentation, the differences about the camera type and personal bioinformation lead to a distribution shift which hurts the performance of model in the target domain. Hence, how to transfer the knowledge of source model to the target domain is a significant problem for medical image analysis.

Recently, there has been extensive research about unsupervised domain adaptation (UDA) in the medical image segmentation field. On the one hand, some studies consider making maximum use of multiple source datasets to adapt a model from the source domain to the target domain (Kang et al., 2020; Li et al., 2021). Training with multiple source datasets can ease the condition of scarce expert knowledge ground truth. Furthermore, the adapted model is capable of exploring more essential knowledge with multiple source datasets involved. On the other hand, some studies propose to use the model's knowledge contained in the source model to transfer domain knowledge so as to preserve personal bioinformation in a medical image (Prabhu et al., 2021a; Yang et al., 2022). Medical data often cause problems about privacy as they contain sensitive information. Thus, source-free unsupervised domain adaptation (SFUDA) is a hot pot for medical applications where only the source trained model and target data are available.

Although those existing works have extremely promoted the possibility of real application for medical image segmentation, all of them only focus on one condition either non-source or multiple sources. Ahmed et al. (2021) explored such setting but they were devoted to the classification task. Therefore, we provide a more practical clinical setting where we have access to only the multiple source trained models in the adapting process for a segmentation task. In this multi-source and source-free setting (MSSF), it can not only protect patient's privacy but also make full use of multiple source datasets to learn more effective knowledge to eliminate distribution shift better. We measure the performance of our proposed method under multiple settings on three fundus image datasets. As far as we know, it is the first attempt to conduct researches on the retinal vessel segmentation task under multi-source and source-free scenarios.

2. Related work

Broadly, there are three different categories for unsupervised domain adaptations (UDAs) that include original unsupervised domain adaptation, source-free unsupervised domain adaptation (SFUDA), and multi-source unsupervised domain adaptation (MSUDA). Under the unsupervised domain adaptation scene, the goal of the model is to learn how to obtain more transferable features for the source domain and the target domain. It can be achieved by emphasizing the features of specific channels with less discrepancy between the first-order and second-order statistics of the source domain and target domain (Feng et al., 2021). Prabhu et al. (2021b) evaluated the reliability of a target instance based on its predictive consistency under a committee of random image transformations. Hoyer et al. (2022) proposed masked image consistency (MIC) that forces network to learn to infer the predictions of the masked regions from their context.

Medical data are sensitive, and they contain private bioinformation and identity information. It inevitably leads to privacy concerns during the process of adaptation with source data. Driven by this fact, some pseudo label generation methods use the knowledge of source model to denoise the pseudo label of target samples under source-free conditions (Chen et al., 2021; VS et al., 2022a). Bateson et al. (2022) introduced a label-free entropy loss and a domain-invariant prior that integrated in the form of a Kullback-Leibler divergence in loss function to guide the adaptation process. Yang et al. (2022) designed a Fourier Style Mining generator to inverse source-like images through statistic information. These generated images can simulate source data distribution and benefit the domain alignment. They designed a domain distillation loss to achieve feature-level adaptation and a domain contrastive loss to narrow down the domain shift using a self-supervised mechanism.

As depending on the characters of medical imaging instruments and patient' organs, medical image datasets from different sources follow different distributions. To make full use of the underlying values of multiple source datasets, adversarial learning is introduced to minimize the distribution shift between multiple source domains and target domains (Chen et al., 2021; VS et al., 2022a). He et al. (2021) proposed a simple image translation to align the pixel value distribution to reduce the domain shift. To make full use of unlabeled data, the pseudo labels generated by an ensembled model constrained the outputs of multiple source models. For the classification task, Ahmed et al. (2021) proposed a new domain adaptation strategy that the source models combine with suitable weights to predict a integrated classification result with the best quality than each source model.

3. Methods

3.1. Dataset description

In our experiments, we choose three public fundus image databases for evaluation including the DRIVE, CHASEDB1, and IOSTAR dataset (Figure 1). Each group of experiments chooses two databases as source domain data and the remaining one as target domain data. The DRIVE dataset contains 20 training images and 20 testing images. This dataset provides two labeled ground truths for each image, and we use the first labeled mask for training and testing. The CHASEDB1 (Child Heart and Health Study in England) dataset contains 28 color vascular images with a resolution of 990 × 960. There are two segmentation annotations available, and we adopt the first manual annotation in our study. We follow the setting in Li et al. (2015) and use the first 20 images for training and the remaining eight images for testing. The IOSTAR dataset includes 30 images taken with an EasyScan camera1 based on SLO technology. These high contrast images have a resolution of 1,024 × 1,024 with 45° FOV. The corresponding ground truths of these vessel images are annotated by experts having a good knowledge of retinal image analysis (Abbasi-Sureshjani et al., 2015; Zhang et al., 2016).

3.2. Measurement of performance

The retinal segmentation task is to classify each pixel in the fundus image into vessel pixel or background pixel. Obviously, it is a binary classification task. In order to analyze the performance of our proposed method quantitatively, we use several common metrics, including accuracy (Acc), sensitivity (Sen), specificity (Spe), which are defined as below:

$$Acc = \frac{TP + TN}{TP + FN + TN + FP}, \tag{1}$$

$$Sen = \frac{TP}{TP + FN}, Spe = \frac{TN}{TN + FP}, \tag{2}$$

where TP and FP denote the number of foreground vessel pixels that are correctly segmented and the number of background pixels that are wrongly classified, respectively. TN represents the number of background pixels that are correctly segmented, and FN denotes the number of foreground vessel pixels that are wrongly classified as background class. Moreover, we also calculate the AUC metric (the area under the ROC curve) that is depended on the recall and precision and is more appropriate to measure performance under an unbalanced circumstance.

3.3. Approach

Figure 2 illustrates the whole structure of our multi-source and source-free UDA framework. In this section, we first present the dual consistency mechanism including intra-domain consistency constraint and inter-domain consistency constraint. Next, we propose a progressive entropy loss that can optimize the features in a progressive way. The training procedures are finally presented.

3.3.1. Inter-domain consistency constraint

Because of the presence of distribution shift, the model trained on the source domain tends to be frustrated when facing target sample. In order to deal with such problem, we introduce the inter-domain consistency constraint, which can select those reliable samples to improve the adaptation process.

For each target image, there are two different augmented images as the input of the source model, that is, a weak-augmented image and a strong-augmented image. The weak augmentation operations include intensity normalization, random rotation, and random flip. The strong augmentation operations include random gray scale adjustment and random color jitter besides the operations in weak augmentation. Therefore, we get two different prediction results for each pixel i in the two augmented images,

$$p_i^{\mathrm{ks}} = S_k(x_i^s), \tag{3}$$

$$p_i^{\mathrm{kw}} = S_k(x_i^w), \tag{4}$$

where S_k denotes the source model trained on the kth source domain. The superscript s denotes the strong augmentation and w denotes the weak augmentation. Due to the existence of domain shift between the source domain and the target domain in the early stage of model training, the two prediction results often have certain differences while the differences represent unreliable samples that do harm to the adaptation. Therefore, we introduce the inter-domain consistency constraint to discover credible samples. If the two prediction results at the same pixel position share the same category, this pixel sample is credible, and it can participate in the domain adaptation. On the contrary, this pixel sample is unreliable and should be discarded. Thus, we can obtain a consistency mask indicating the dependability of each pixel,

$$m_i^k = \begin{cases} 1, l_i^{ks} == l_i^{kw}, \\ 0, otherwise, \end{cases} \tag{5}$$

where l_i^{ks} and l_i^{kw} denote the pseudo label of the ith pixel in strong-augmented image and weak-augmented image, and m_i^k indicates whether the ith pixel is selected to adapt the kth source model.

For the approach of pseudo label generation, we proposed a dynamic threshold mechanism with weak-augmented image for each source model given by

$$T_{dyn} = \max_{\gamma\%}(sort(p_i^{kw})). \tag{6}$$

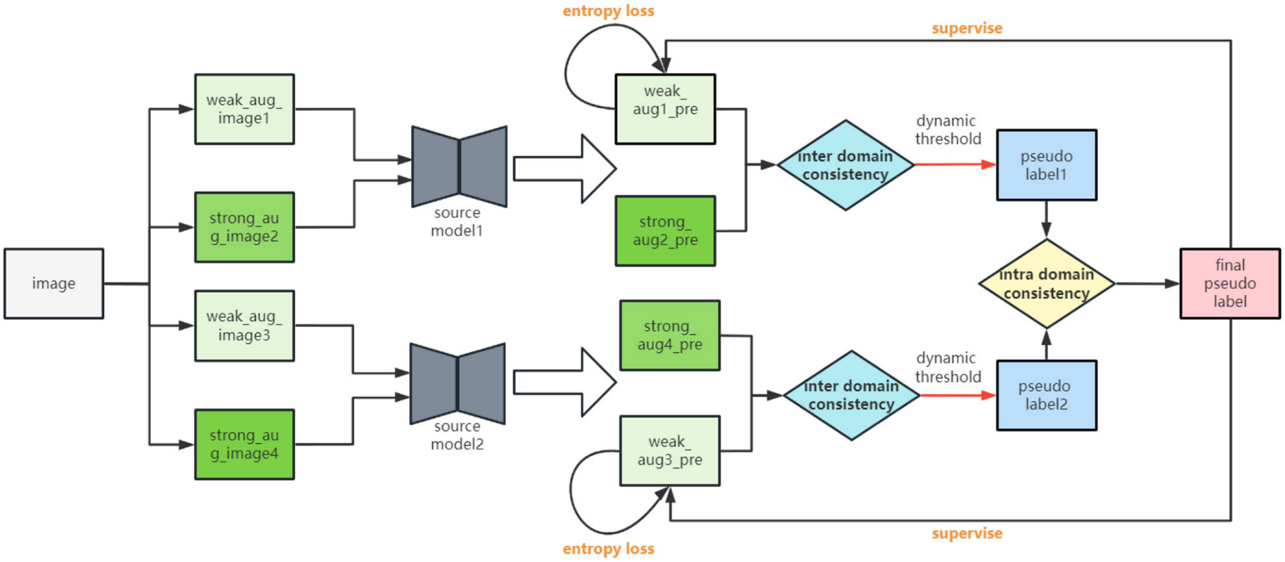


FIGURE 2
Overview of our proposed MSSF framework. Pseudo labels are generated under the guidance of the dual consistency constraint. Inter-domain consistency constraint aims to optimize the intra-class distance of each source model and also ease the distribution shift between the source domain and the target domain with two partially different predictions in a sense. Intra-domain consistency constraint can utilize the knowledge of multiple source models and thus can teach multiple models more essential and transferable knowledge regarding different data domains. Progressive entropy loss is complementary to our proposed dynamic pseudo label generation method, which can optimize the effectiveness of feature step by step.

After ranking the predicted score results in ascending order, the top γ percentage probability value is taken as the dynamic factor. Dynamic factor fits in with the adaptation process that can adjust according to the training epoch,

$$\gamma = \min(a - \frac{b-a}{total\_epoch} * epoch, b), \tag{7}$$

where a and b are the upper and lower bounds of the interval. With the increase of training rounds, the number of credible vessels samples in prediction also increases gradually, and the probability distribution of the prediction results gradually inclines to the high probability area. Thereafter, the dynamic threshold is supposed to be reduced. After the dynamic threshold is obtained, the final pseudo label can be obtained by

$$l_i^{kw} = \begin{cases} 1, p_i^{kw} >= T_{dyn}, \\ 0, otherwise. \end{cases} \tag{8}$$

It should be noted that the generation method of pseudo label can be both applied to the prediction of weak-augmentation image and strong-augmentation image. The pseudo label of strong-augmentation image l_i^{kw} only relates to the consistency mask m_i^k.

3.3.2. Intra-domain consistency constraint

Considering that the inter-domain consistency constraint only focuses on a single source domain, it can only improve the feature compactness of the model in a single source domain. In order to make full use of the information in multiple source domains, we propose that the intra-domain consistency constraint learns the crucial knowledge and gets rid of the domain shift. Given two source domains and corresponding source models S_1 and S_2 as special cases, we can get the pseudo label imposing intra-domain consistency constraint for target sample,

$$t_i = \begin{cases} l_i^{1w}, l_i^{1w} == l_i^{2w}, \\ 2, otherwise. \end{cases} \tag{9}$$

The valid pseudo label t_i will be given to the pixel i only when the pseudo labels of S_1 and S_2 are consistent; otherwise, it will be assigned invalid category 2. The final output pseudo label only depends on the pseudo label of a weak-augmentation image instead of a strong-augmentation image.

Those highly reliable pixels obtained via inter-domain consistency constraint can maintain prediction-invariance on different source domain models for each target domain image. This prediction-invariance character can alleviate the domain shift between multiple source domains and target domains to a certain extent and improve generalization between multiple source domains and target domains. Therefore, for the dual consistency constraints, the intra-domain consistency constraint of a single source domain can reduce the intra-class distance and the feature space, while the intra-domain consistency constraint of multiple source domains can utilize multiple source domain models to ease the domain shift problem.

Thereafter, we introduce a consistency loss to utilize the advantages of both the inter-domain consistency constraint and intra-domain consistency constraint via filtering out samples that do not meet both consistency constraints. For all source domain models, consistency loss $Loss_{con}$ is defined as

$$Loss_{con} = -\frac{1}{KN}\sum_{k=1}^{K}\sum_{i=1}^{N} CE(p_i^{kw}, t_i), \tag{10}$$

where CE denotes cross entropy loss,

$$CE(p_i^{kw}, t_i) = \begin{cases} -[p_i^{kw} log t_i + (1 - p_i^{kw}) log(1 - t_i)], \\ t_i \neq 2 \& m_i^k == 1, \\ 0, otherwise. \end{cases} \quad (11)$$

3.3.3. Progressive entropy loss

Dual consistency constraints can identify valuable pixels for model training from both intra-domain and inter-domain perspectives to deal with the unlabeled data. However, this strategy is not completely satisfied. The pseudo label generated by dynamic threshold mechanism cannot entirely substitute for the real ground truth, which causes the intra-class feature to be not discriminative enough. Therefore, in order to further reduce the distance of the intra-class features, we propose progressive entropy loss.

Entropy minimization is familiar in semi-supervised learning and unsupervised domain adaptation, which essentially supervises model with the help of high probability regions during the training phase. On the other hand, entropy minimization can also be seen as a clustering method to compress the distance within each class, making the features extracted from the model more compact (Chen et al., 2019; Zou et al., 2019). However, there will be some problems occurring when applying this method directly. Due to the lack of ground truths, the model is usually unstable in the early training stage and the prediction is inaccurate. Then, the training model tends to collapse and fall into the local optimal solution. Therefore, we have come up with a progressive entropy loss strategy, which gradually increases the weight of the unsupervised entropy loss during the training phase to avoid the problem of insufficient optimization of the model.

First, the unsupervised entropy loss is calculated based on the prediction results of the weak augmented samples for multi-source models:

$$Loss_{ent} = -\frac{1}{KN} \sum_{k=1}^{K} \sum_{i=1}^{N} p_i^{kw} log(p_i^{kw}), \quad (12)$$

where K denotes the number of source models and N denotes the whole pixel set of target dataset. The dynamic factor β will be adjusted according to the training epoch:

$$\beta = \max(a + \frac{b - a}{total\_epoch} * epoch, b). \quad (13)$$

When the model gradually becomes stable with the increase of training epochs, it gradually strengthens the constraint of entropy minimization in a reasonable manner:

$$Loss_{pro\_ent} = \beta \times [-\frac{1}{KN} \sum_{k=1}^{K} \sum_{i=1}^{N} p_i^{kw} log(p_i^{kw})]. \quad (14)$$

Therefore, the final loss of our proposed method defined as follows:

$$Loss = Loss_{con} + Loss_{pro\_ent}. \quad (15)$$

4. Experiments

4.1. Experiments setting

The implementation of our approach is based on the publicly Pytorch framework. We train our models on a NVIDIA GeForce RTX 3090 graphics card with a memory of 24 GB. We adopt the Adam algorithm as our network optimization method, of which the hyperparameters usually do not need to be adjusted.

Under the multi-source scenario, the number of source domain datasets in our experiments is 2, the batchsize is set to be 2 for both source domains, the training epoch is set to be 10, and the initial learning rate is set to be 0.00002. Because our experiments are conducted on the DRIVE, CHASEDB1, and IOSTAR datasets, in a multi-source scenario, three groups of experiments can be formed: (1) The source domains are the DRIVE and CHASEDB1 datasets, and the target domain is the IOSTAR dataset. (2) The source domains are the DRIVE and IOSTAR datasets, and the target domain is CHASEDB1. (3) The source domains are the CHASEDB1 and IOSTAR datasets, and the target domain is DRIVE. We evaluate all methods via four common metrics for segmentation task including AUC, accuracy (Acc), specificity (Spe), and sensitivity (Sen). The AUC represents the overall performance which is more appropriate to judge whether an algorithm is robust or not under an unbalanced circumstance. The higher the value of Acc, the higher the correct recognition rate of the algorithm. The Spe and Sen metrics indicate the recognition capacity of background class and vessel class, respectively.

4.2. Ablation experiments

We perform the ablation experiments to validate our proposed modules are effective or not on the DRIVE dataset. Comprehensive results are summarized in Table 1. The baseline method does not use any modules. It uses multiple source models to predict separately and then obtain pseudo labels for the prediction results of each source model directly through a hard threshold mechanism. The pseudo labels will be used to monitor the prediction results of target domain sample after the integration of the predicted results of multiple source domain models. This approach also utilizes knowledge from multiple source domains, similar to the idea of integrated learning. This method is also used as a strong baseline under multi-source scenarios in our comparison study. The method-a adds inter-domain consistency constraint module (inter-domain CC) based on the baseline. The method-b adds the intra-domain CC constraint module (Intra-domain CC) based on method-a. The method-c adds progressive entropy loss (PEL) based on method-b.

4.2.1. The impact of inter-domain consistency constraint

Inter-domain consistency constraint can filter out pixels with inconsistent categories in the predicted results under different augmentation operations, improving the stability and consistency of the model. Such constraint can explore more

TABLE 1 Ablation study on DRIVE dataset.

Method	Inter-domain CC	Intra-domain CC	PEL	AUC	Acc	Spe	Sen
Baseline	–	–	–	0.9737	**0.9660**	**0.9830**	0.7793
Method-a	✓	–	–	0.9755	0.9652	0.9819	0.7916
Method-b	✓	✓	–	0.9762	0.9623	0.9741	0.8384
Method-c	✓	✓	✓	**0.9764**	0.9611	0.9718	**0.8493**

The bold values indicate the highest performance metrics in each column.

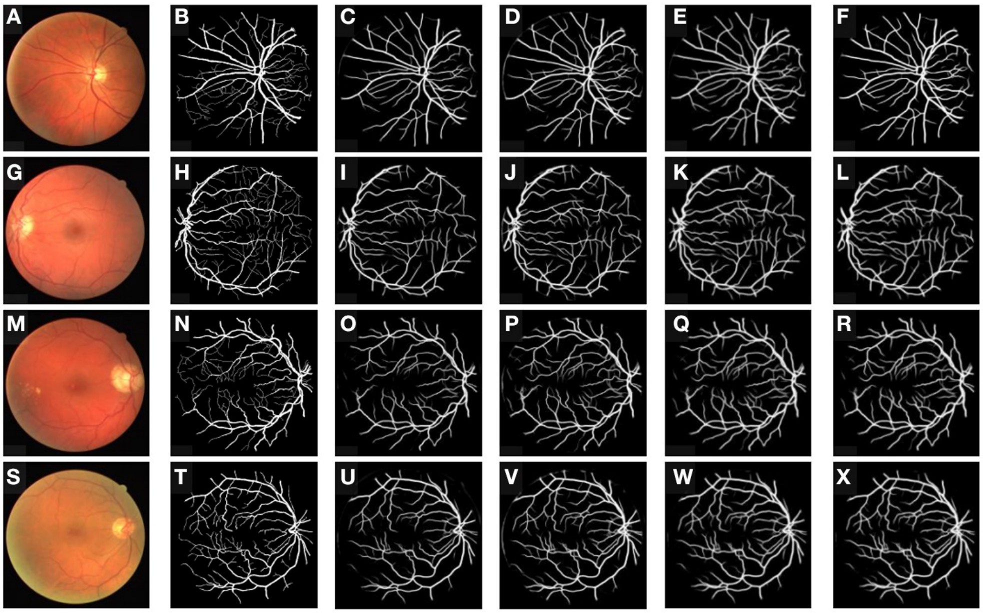


FIGURE 3
Retinal vessel segmentation results for our proposed method on the DRIVE dataset. From the column in left to right **(A–X)**, the retinal original images, the ground truths, the segmentation results of the baseline, the segmentation results of method-a, and the segmentation results of method-b, the segmentation results of method-c.

valuable vessel samples than background samples. Therefore, it has a strengthening effect on the learning of vessel regions, and the sensitivity of the method-a is improved compared to the baseline.

4.2.2. The impact of intra-domain consistency constraint

When there is only the inter-domain consistency constraint module, the knowledge of each source domain model is mixed, which is not beneficial to the learning of knowledge in the target domain. Accordingly, when introducing the intra-domain consistency constraint module, more effective vessel pixels are identified during the model training for supervised learning, making full use of the inherent knowledge of multiple source domains. Therefore, the AUC and sensitivity metrics of method-b are increased compared with method-a, especially the increase in sensitivity. However, the specificity decreases from 0.9819 to 0.9741 due to such constraint, because it filters some samples of background class when the model is able to identify more vessels.

4.2.3. The impact of progressive entropy loss

On the one hand, unsupervised progressive entropy loss enhances the compactness of intra-class features. The high probability regions obtained through supervised learning with pseudo labels guide the model to extract more discriminative features for background and vessel classes. On the other hand, because the generation of pseudo label is based on the dynamic threshold mechanism, it gradually strengthens the recognition capability of vessels during the adaptation process. Therefore, compared with the other experiment group, method-c has a significant improvement in sensitivity, with the highest AUC and sensitivity. Although the accuracy and specificity of the final model have slightly decreased, it has brought about significant improvements in sensitivity, which is more practical for medical image segmentation and can detect more foreground objects to assist in medical diagnosis.

We also provide the visualization result of our proposed method in different ablation experiment groups in Figure 3. It can be seen that for method-a group with only the inter-domain consistency constraint module, it is easy to predict the outer

TABLE 2 Comparison experiments on three target domains.

Method	AUC	Acc	Spe	Sen
Source domains: drive and CHASEDB1, target Domain:IOSTAR				
Oracle	0.9850	0.9658	0.9831	0.8426
AdaptSegNet (DRIVE/CHASEDB1)	0.9361/0.9680	0.9534/0.9569	0.9802/**0.9812**	0.6943/0.6703
DPL (DRIVE/CHASEDB1)	0.9569/0.9631	0.9458/0.9499	0.9771/0.9783	0.7027/0.7124
TT_SFUDA (DRIVE/CHASEDB1)	0.9104/0.9344	0.9390/0.9358	0.9799/0.9774	0.6218/0.6496
Multi-Source	**0.9834**	**0.9651**	0.9789	0.8180
Ours	0.9824	0.9625	0.9743	**0.8357**
Source domains: drive and IOSTAR, target Domain:CHASEDB1				
Oracle	0.9883	0.9711	0.9776	0.8769
AdaptSegNet (DRIVE/IOSTAR)	0.9659/0.9388	0.9561/0.9533	0.9853/0.9839	0.7824/0.6930
DPL (DRIVE/IOSTAR)	0.9513/0.9652	0.9511/**0.9630**	0.9842/**0.9861**	0.6397/0.7442
TT_SFUDA (DRIVE/IOSTAR)	0.9517/0.9556	0.9396/0.9393	0.9714/0.9710	0.7793/0.7871
Multi-Source	**0.9819**	0.9616	0.9688	0.8546
Ours	0.9816	0.9606	0.9674	**0.8601**
Source domain: CHASEDB1 and IOSTAR, target Domain:DRIVE				
Oracle	0.9833	0.9631	0.9738	0.8516
AdaptSegNet (CHASEDB1/IOSTAR)	0.9638/0.9470	0.9591/0.9512	**0.9877**/0.9843	0.6634/0.6789
DPL (CHASEDB1/IOSTAR)	0.9511/0.9553	0.9501/0.9528	0.9816/0.9845	0.6332/0.6351
TT_SFUDA (CHASEDB1/IOSTAR)	0.9314/0.9407	0.9336/0.9389	0.9759/0.9801	0.7768/0.7598
Multi-Source	0.9737	**0.9660**	0.9830	0.7793
Ours	**0.9764**	0.9611	0.9718	**0.8493**

The bold values indicate the highest performance metrics in each column.

circle of the eyeball as a blood vessel, indicating that the pseudo labels for blood vessels are not accurate enough, and the features extracted from the model are not clean. The introducing of the intra-domain consistency constraint module greatly improves this problem because it can filter out pixels that are prone to false segmentation by using the knowledge of multiple source models. Since progressive entropy loss can be beneficial to obtain more discriminative features, it can be found that the method-c recognizes more difficult samples correctly.

4.3. Comparison experiments

We perform experiments on DRIVE, CHASEDB1 and IOSTAR three datasets, where we choose two datasets as the source domain and the remaining one as the target domain. We compare our proposed method with three methods with three different multi-source and source-free single condition settings in Table 2.

Compared to the original unsupervised domain adaptation method such as AdaptSegNet (Tsai et al., 2018), adversarial learning at the output result level is clearly desirable due to the similar spatial location and target sizes in cityscape dataset. However, there are significant differences and complex distribution in different vessels, and it failed to capture the effective knowledge of vessel distribution. On the other hand, our proposed method achieves better performance than two source-free domain adaptation methods including DPL and TT_SFUDA (Chen et al., 2019; VS et al., 2022b). These two methods do not essentially solve the domain shift problem because of the significant differences in experimental results across the different target domains. Due to the existence of multiple source models, our proposed method can alleviate the domain shift on the target domain through dual consistency constraints and sufficiently explore the essential knowledge of multiple source domains. Therefore, it is minimally affected by the magnitude of the domain shift, and has gained relatively ideal performance in different target domains. The performance of the multi-source algorithm is familiar with our method, but when the target domain is DRIVE with a large number of thin vessels, its performance drops a lot. Such a defect can be attributed to the lack of effective treatment of the pseudo label.

Our proposed method achieves better performance than unsupervised domain adaptation methods including source-free and multi-source single scene settings on different target domains. It can sufficiently explore the knowledge fusion in multiple source models while retaining the advantage of source pretrained model of high AUC and sensitivity metrics.

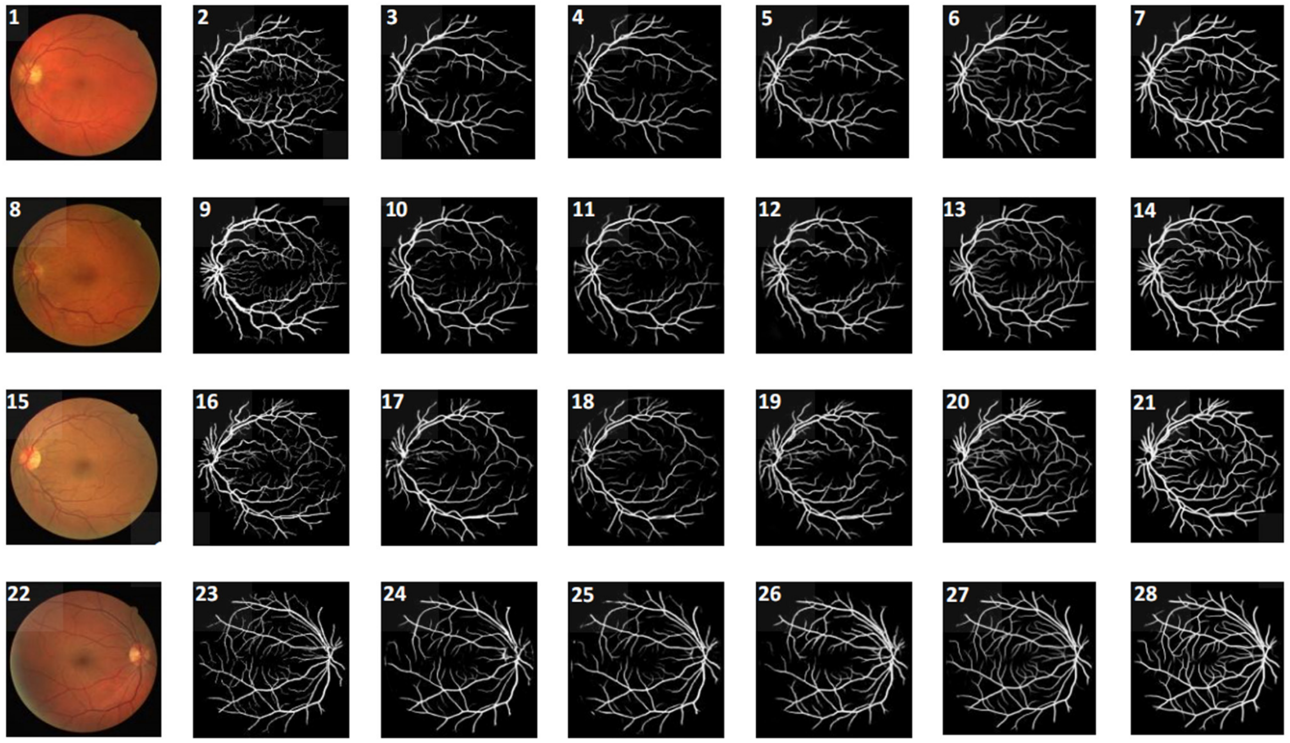


FIGURE 4
Visualization for different methods on DRIVE target domain. From the **left** to **right** columns, they are original image, ground truth, and the segmentation results of AdaptSegNet, DPL, TT_SFUDA, multi-source, and our proposed MSSF algorithm, respectively.

Such advantage makes our approach more meaningful and practical that more vessels can be identified as correctly as possible especially under the unsupervised domain adaptation scenario.

We also present a visual comparison of the segmentation results of several methods as shown in Figure 4. Compared with other methods, our approach has fewer false segmentation cases, which effectively avoids the occurrence of mistakenly identifying the outer circle of the eye as vessel class. It also has the best recognition performance for a large number of capillaries in the middle area of a fundus image, preventing the fracture of vessel occurring.

5. Conclusion

This study designs a brand-new unsupervised domain adaptation framework, which expands the single unsupervised domain adaptation scene including source-free and multi-source settings. Our proposed dual consistency constraint can filter out noisy pseudo labels based on the knowledge in each source models and the fusion between them. To effectively promote the feature clustering, progressive entropy loss can not only compress the distance within each class but also can benefit the generation of pseudo label in turn. The proposed MSSF framework combines the advantages of source-free and multi-source adaptation. We hope this paradigm can inspire future studies about unsupervised domain adaptation.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: https://drive.grand-challenge.org/, http://www.retinacheck.org/download-iostar-retinalvessel-segmentation-dataset, and https://blogs.kingston.ac.uk/retinal/chasedb1/.

Ethics statement

Written informed consent was obtained from the individual(s), and minor(s)' legal guardian/next of kin, for the publication of any potentially identifiable images or data included in this article.

Author contributions

BC: conceptualization, methodology, software, investigation, formal analysis, and writing-original draft. LM: data curation, methodology, resources, software, supervision, writing-original draft. YS: visualization and investigation. All authors contributed to the article and approved the submitted version.

Funding

This study was supported in part by the National Key R&D Program of China (No. 2021YFA1003004) and in part by the Shanghai Municipal Natural Science Foundation under Grant 21ZR1423300.

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

Abbasi-Sureshjani, S., Smit-Ockeloen, I., Zhang, J., and Ter Haar Romeny, B. (2015). "Biologically-inspired supervised vasculature segmentation in SLO retinal fundus images," in *Image Analysis and Recognition: 12th International Conference, ICIAR 2015* (Niagara Falls, ON: Springer), 325–334. doi: 10.1007/978-3-319-20801-5_35

Ahmed, S. M., Raychaudhuri, D. S., Paul, S., Oymak, S., and Roy-Chowdhury, A. K. (2021). "Unsupervised multi-source domain adaptation without access to source data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Kuala Lumpur), 10103–10112. doi: 10.1109/CVPR46437.2021.00997

Bateson, M., Kervadec, H., Dolz, J., Lombaert, H., and Ayed, I. B. (2022). Source-free domain adaptation for image segmentation. *Med. Image Anal.* 82:102617. doi: 10.1016/j.media.2022.102617

Chen, C., Liu, Q., Jin, Y., Dou, Q., and Heng, P. A. (2021). "Source-free domain adaptive fundus image segmentation with denoised pseudo-labeling," in *Medical Image Computing and Computer Assisted Intervention-MICCAI 2021* (Strasbourg), 225–235. doi: 10.1007/978-3-030-87240-3_22

Chen, M., Xue, H., and Cai, D. (2019). "Domain adaptation for semantic segmentation with maximum squares loss," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Seoul), 2090–2099. doi: 10.1109/ICCV.2019.00218

Feng, W., Ju, L., Wang, L., Song, K., Wang, X., Zhao, X., et al. (2021). Unsupervised domain adaptation for retinal vessel segmentation with adversarial learning and transfer normalization. *arXiv preprint arXiv:2108.01821*.

Guo, H., Li, J., Gao, G., Li, Z., and Zeng, T. (2023). "PFT-SSR: Parallax fusion transformer for stereo image super-resolution," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Rhode Island), 1–5. doi: 10.1109/ICASSP49357.2023.10096174

He, J., Jia, X., Chen, S., and Liu, J. (2021). "Multi-source domain adaptation with collaborative learning for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Kuala Lumpur), 11008–11017. doi: 10.1109/CVPR46437.2021.01086

Hoyer, L., Dai, D., Wang, H., and Van Gool, L. (2022). MIC: masked image consistency for context-enhanced domain adaptation. *arXiv preprint arXiv:2212.01322.*

Kang, G., Jiang, L., Wei, Y., Yang, Y., and Hauptmann, A. (2020). Contrastive adaptation network for single-and multi-source domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 1793–1804. doi: 10.1109/TPAMI.2020.3029948

Li, Q., Feng, B., Xie, L., Liang, P., Zhang, H., and Wang, T. (2015). A cross-modality learning approach for vessel segmentation in retinal images. *IEEE Trans. Med. Imaging* 35, 109–118. doi: 10.1109/TMI.2015.2457891

Li, Y., Yuan, L., Chen, Y., Wang, P., and Vasconcelos, N. (2021). "Dynamic transfer for multi-source domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Kuala Lumpur), 10998–11007. doi: 10.1109/CVPR46437.2021.01085

Orlando, J. I., Prokofyeva, E., and Blaschko, M. B. (2016). A discriminatively trained fully connected conditional random field model for blood vessel segmentation in fundus images. *IEEE Trans. Biomed. Eng.* 64, 16–27. doi: 10.1109/TBME.2016.2535311

Prabhu, V., Khare, S., Kartik, D., and Hoffman, J. (2021a). Augco: augmentation consistency-guided self-training for source-free domain adaptive semantic segmentation. *arXiv preprint arXiv:2107.10140.*

Prabhu, V., Khare, S., Kartik, D., and Hoffman, J. (2021b). "Sentry: selective entropy optimization via committee consistency for unsupervised domain adaptation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Montreal, QC), 8558–8567. doi: 10.1109/ICCV48922.2021.00844

Tsai, Y.-H., Hung, W.-C., Schulter, S., Sohn, K., Yang, M.-H., and Chandraker, M. (2018). "Learning to adapt structured output space for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT), 7472–7481. doi: 10.1109/CVPR.2018.00780

VS, V., Valanarasu, J. M. J., and Patel, V. M. (2022a). "Adaptive pseudo labeling for source-free domain adaptation in medical image segmentation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Shenzhen), 1091–1095.

VS, V., Valanarasu, J. M. J., and Patel, V. M. (2022b). Target and task specific source-free domain adaptive image segmentation. *arXiv preprint arXiv:2203.15792.*

Wu, H., Wang, W., Zhong, J., Lei, B., Wen, Z., and Qin, J. (2021). Scs-net: A scale and context sensitive network for retinal vessel segmentation. *Med. Image Anal.* 70:102025. doi: 10.1016/j.media.2021.102025

Yang, C., Guo, X., Chen, Z., and Yuan, Y. (2022). Source free domain adaptation for medical image segmentation with Fourier style mining. *Med. Image Anal.* 79:102457. doi: 10.1016/j.media.2022.102457

Zhang, J., Dashtbozorg, B., Bekkers, E., Pluim, J. P., Duits, R., and ter Haar Romeny, B. M. (2016). Robust retinal vessel segmentation via locally adaptive derivative frames in orientation scores. *IEEE Trans. Med. Imaging* 35, 2631–2644. doi: 10.1109/TMI.2016.2587062

Zou, Y., Yu, Z., Liu, X., Kumar, B., and Wang, J. (2019). "Confidence regularized self-training," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Seoul), 5982–5991. doi: 10.1109/ICCV.2019.00608

TYPE Methods
PUBLISHED 22 August 2023
DOI 10.3389/fnins.2023.1246995

OPEN ACCESS

EDITED BY
Zhengwang Wu,
University of North Carolina at Chapel Hill,
United States

REVIEWED BY
Lin Liu,
Peking University, China
Yu Qi,
Zhejiang University, China

*CORRESPONDENCE
Hong Ji
✉ itsjihong@gmail.com
Wen Jiang
✉ jiangwen@fmmu.edu.cn

RECEIVED 25 June 2023
ACCEPTED 24 July 2023
PUBLISHED 22 August 2023

CITATION
Ji H, Xu T, Xue T, Xu T, Yan Z, Liu Y, Chen B and Jiang W (2023) An effective fusion model for seizure prediction: GAMRNN. *Front. Neurosci.* 17:1246995. doi: 10.3389/fnins.2023.1246995

An effective fusion model for seizure prediction: GAMRNN

Hong Ji[1]*, Ting Xu[1], Tao Xue[1], Tao Xu[2], Zhiqiang Yan[3], Yonghong Liu[3], Badong Chen[4] and Wen Jiang[3]*

[1]Shaanxi Provincial Key Laboratory of Fashion Design Intelligence, Xi'an Polytechnic University, Xi'an, China, [2]School of Software, Northwestern Polytechnical University, Xi'an, China, [3]Xijing Hospital, Fourth Military Medical University, Xi'an, China, [4]Institute of Artistic Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, China

The early prediction of epileptic seizures holds paramount significance in patient care and medical research. Extracting useful spatial-temporal features to facilitate seizure prediction represents a primary challenge in this field. This study proposes GAMRNN, a novel methodology integrating a dual-layer gated recurrent unit (GRU) model with a convolutional attention module. GAMRNN aims to capture intricate spatial-temporal characteristics by highlighting informative feature channels and spatial pattern dynamics. We employ the Lion optimization algorithm to enhance the model's generalization capability and predictive accuracy. Our evaluation of GAMRNN on the widely utilized CHB-MIT EEG dataset demonstrates its effectiveness in seizure prediction. The results include an impressive average classification accuracy of 91.73%, sensitivity of 88.09%, specificity of 92.09%, and a low false positive rate of 0.053/h. Notably, GAMRNN enables early seizure prediction with a lead time ranging from 5 to 35 min, exhibiting remarkable performance improvements compared to similar prediction models.

KEYWORDS

EEG, spatial temporal feature, seizure prediction, attention module, GAMRNN

1. Introduction

Epilepsy, also known as "fits" or "the falling sickness," is a chronic neurological disorder in which sudden, abnormal electrical activity in the brain causes disruptions in its normal functioning (Artameeyanant et al., 2017). It is estimated that almost 65 million people worldwide have epilepsy, which accounts for ∼1% of the global population (Bou Assi et al., 2017). The clinical manifestations of epilepsy are complex and varied, with symptoms ranging from motor, sensory, autonomic, and cognitive disturbances. While certain medications can help reduce the frequency of epileptic seizures, they are not always effective and may lead to serious side effects, threatening to the patients' daily lives and overall safety. Therefore, developing a reliable algorithmic model for predicting epileptic seizures, which can provide early warning and preventive measures, is paramount for the patients' survival.

As an epileptic seizure begins, brain activity transitions from one state to another, accompanied by significant changes in the brain's electrical signals. Electroencephalography (EEG) is an effective method for monitoring the waveform changes in brain electrical signals during epileptic seizures. The EEG during a seizure can be categorized into four main states: preictal (a period before the onset of a seizure), ictal (a period during the seizure), postictal (a period following the seizure), and interictal (a period when the brain is not experiencing a seizure; Natu et al., 2022). Experienced experts can discern distinct states of epileptic seizure electroencephalogram (EEG) signals through observation. Nonetheless, the manual segmentation process of epileptic seizure signals is often laborious and time-consuming, necessitating graphologists with a high level of technical proficiency. Hence, its practical applicability is inherently challenging. The primary task in epileptic seizure prediction is to accurately extract features from EEG signals during the seizure period and differentiate them based on distinctive characteristics to separate preictal and interictal signals. This enables the prediction of the potential timing of a seizure, providing early warnings to patients and facilitating the implementation of intervention and remedial measures to minimize the impact of seizure episodes on patients. Throughout the course of an epileptic seizure, the importance of different EEG channels in seizure prediction research varies. Thus, there are challenges in selecting informative channels to extract more valuable feature information that ultimately helps improve the performance of seizure prediction models. Based on prior work, our proposed method for epileptic seizure prediction primarily encompasses the following steps: EEG signal acquisition, EEG preprocessing, feature extraction, model learning and training, classification of interictal and preictal data segments, seizure prediction, and model evaluation. During the model training, we incorporated the Convolutional Block Attention Module to enhance the model's attention to important channels and valuable feature information. Additionally, we utilized the Lion optimization algorithm for further optimization of the model training, ultimately improving seizure prediction performance, as illustrated in Figure 1.

In order to extract features that can effectively differentiate between pre-ictal and interictal EEG, prior researchers have attempted various methods. The most commonly used features include wavelet energy, power spectral density, phase locking value, permutation entropy, and fractal dimension value (Li et al., 2013; Joshi et al., 2014; Khalid et al., 2015; Zhang et al., 2020). Fei et al. (2017) used an improved largest Lyapunov exponent algorithm to better characterize the chaotic dynamical characteristics of EEG signals during epilepsy seizures, and the results showed that the improved algorithm had higher accuracy in identifying pre-seizure signals. Raghu et al. (2019) proposed a continuous decomposition index feature, which was proven to have a significant enhancement trend during epilepsy seizures so that the epilepsy seizure could be predicted in the pre-seizure period based on its changes. In addition, some studies use methods such as CSP transformation, principal component analysis, and autoregressive models to extract frequency or spatial domain features during epilepsy seizures (Büyükçakır et al., 2020). Due to the subjective selection of feature information, which may result in feature redundancy or the absence of crucial features, some researchers have proposed feature selection algorithms to select the optimal feature information (Karthick et al., 2018). Varatharajah et al. (2017) developed a scalp electroencephalogram (EEG) processing pipeline and introduced a seizure prediction method. The research findings indicate that the performance of the proposed prediction algorithm surpasses that of the baseline algorithm on the tested feature set. Bandarabadi et al. (2015) used an amplitude distribution-based feature selection algorithm; the study showed that this algorithm could also improve the accuracy of epilepsy prediction. After feature information extraction, the next step is to perform binary classification on the EEG signals. Yang et al. (2018) proposed a data analysis modeling method, and research showed that a seizure prediction system based on support vector machines could achieve robust preictal and interictal signals prediction. Yuan et al. (2017) utilized the diffusion distance measure and employed the Bayesian linear discriminant analysis to identify the periodicity of pre-seizure EEG signals, achieving high sensitivity and low false alarm rate. In addition, various methods have been used in seizure detection tasks, such as extreme learning machines, linear discriminant analysis, decision trees, random forest, etc. (Song et al., 2012; Rasekhi et al., 2013; Hussain, 2018; Mohan et al., 2018).

With the significant advancements of deep learning techniques in fields such as computer vision, it has also started to be gradually employed in the research of epileptic seizure prediction (Yıldırım et al., 2018; Liu et al., 2019; Yu et al., 2020). Firstly, the prediction model based on Convolutional Neural Networks (CNN) can well capture the feature information of EEG data due to its characteristics of local connectivity, weight sharing, and downsampling in time and space. Shasha et al. (2021) partitioned the experiment into two phases. They computed the Pearson correlation coefficient of the EEG signals. Subsequently, they fed the resulting correlation matrix into a simplistic CNN model to perform binary classification between interictal and preictal states. This approach effectively minimized computational overhead and yielded an accuracy rate of 89.98% when evaluated on the CHB-MIT dataset. Hu et al. (2019) employed CNN as a feature extraction model and used support vector machines (SVM) as classifiers for analyzing electroencephalograms (EEG). Truong et al. (2018) used STFT to extract frequency-domain and time-domain information from EEG signals on a 30 s window and input the transformed spectrogram into the neural network for model training. The model was evaluated on the Freiburg, CHB-MIT, and American Epilepsy Society seizure prediction challenge datasets and could predict seizures from 30 to 5 min before the onset of seizures, substantiating the advantages and generalization abilities of CNN in the field of epileptic seizure prediction research for capturing EEG signal features.

Nevertheless, despite the impressive capability of CNNs in extracting spatial features from signals, they encounter significant limitations when it comes to capturing the temporal dynamics of the signals, which is crucial for identifying and predicting epileptic seizures. Recurrent neural networks (RNNs) can handle sequential data and are suitable for non-stationary time series signals such as EEG data, as they can directly learn from raw EEG data to preserve the maximum temporal feature information of the signal (Ghosh et al., 2017). However, as the depth of RNNs increases, problems such as gradient explosion or vanishing may occur, so researchers have proposed methods using improved RNNs such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU). Tsiouris et al. (2018) employed a feature extraction methodology to extract raw EEG information and employed Long Short-Term Memory (LSTM) networks to generate prediction outcomes. Furthermore, the study evaluated the influence of different preictal windows on the assessment results. Impressively high sensitivity and specificity rates of 99.28% were achieved, along with a false alarm rate of 0.107/h. This experiment also confirmed the outstanding performance of LSTM in analyzing preictal EEG signals. Varnosfaderani et al. (2021) proposed an epileptic seizure prediction model based on a two-layer LSTM and Swish activation function. This structure performs feature

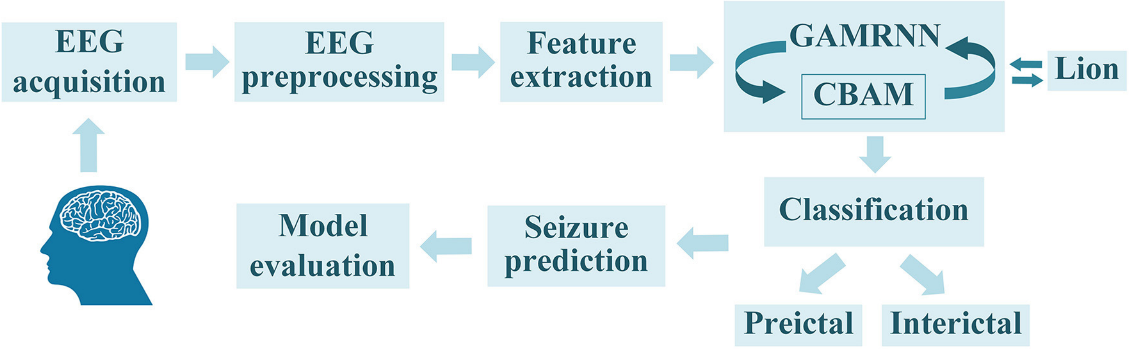


FIGURE 1
The overall seizure prediction method flow, as shown in the figure, includes EEG signal acquisition, EEG preprocessing, feature extraction, training the EEG graphs using the GAMRNN epileptic seizure prediction model augmented with the CBAM module, optimizing the trained model through the Lion optimization algorithm, partitioning the EEG signals into interictal signals and preictal signals, and ultimately predicting seizures and assessing the model's performance.

extraction based on both time and frequency domains and uses the minimum distance algorithm as a post-processing step. The model achieved a sensitivity of 86.8%, prediction accuracy of 85.1%, and a low false positive rate of 0.147/h when evaluated on the Melbourne dataset, which indicates that LSTM performs at a comparable level to CNN in the research of epileptic seizure prediction and may even have a more significant advantage in capturing the temporal features of EEG signals.

Continuous efforts of previous studies have demonstrated that integrating temporal and spatial characteristics of EEG signals is essential for enhancing the efficiency of epileptic seizure prediction. Consequently, algorithms combining Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have emerged to capture the crucial temporal and spatial feature information of EEG signals. Affes et al. (2019) proposed a Convolutional Gated Recurrent Neural Network (CGRNN) for seizure prediction and demonstrated that this model outperformed a CNN-only model in predicting seizures, achieving an average sensitivity of 89% and an average accuracy of 75.6% using a dataset from Boston Children's Hospital. Hu et al. (2020) developed a deep bidirectional long short-term memory (Bi-LSTM) network as a predictive model for epileptic seizure prediction. The experiments employed local mean decomposition (LMD) and statistical feature extraction techniques to capture essential features. The achieved sensitivity of the model was 93.61%, with a specificity of 91.85%. However, these models still face challenges in distinguishing useful signals from noise and irrelevant information, which may lead to reduced the performance of seizure prediction.

Since its introduction, the attention mechanism has been widely applied in various fields such as computer vision (Zhu et al., 2018) and natural language processing (Wu et al., 2019). This is due to its ability to allow neural network models to focus more on relevant information in the input while reducing attention to irrelevant information. Consequently, it has been applied in epileptic seizure prediction research to help models accurately capture useful temporal and spatial features in EEG signals. Concentrating on the most relevant EEG signals and disregarding noise and irrelevant information can improve the classification, and prediction performance of the models. Choi et al. (2022) proposed an ACGRU generalized prediction model that combines one-dimensional convolutional layers, gated recurrent unit layers, and attention mechanisms across patient paradigms to classify preictal and interictal data. Improved classification accuracy and predictive performance were achieved on the EEG dataset of epileptic patients from Eshan Medical Center Children's Hospital, surpassing the performance of the original model. Wang et al. (2022) proposed adding a channel attention module to their CNN-LSTM-based seizure prediction model to address the issue of equal weighting for each channel's feature map in traditional models, achieving an accuracy of 83.04% after training and improving the recognition rate during the correct seizure period. These experiments have consistently demonstrated that neural network models with incorporated attention modules exhibit superior performance in seizure prediction algorithms.

Attention modules contribute to the enhancement of predictive performance. They operate independently on either channel-specific or spatial-specific features of EEG signals. In addition, the convolutional attention module combines channel attention and spatial attention, facilitating concurrent processing of both channel and spatial information. This incorporation enables the model to comprehensively capture critical features across diverse channels and spatial dimensions, thereby elevating the accuracy and robustness of epileptic seizure prediction. The convolutional attention module (Woo et al., 2018) achieves the weighting operation on the channel and spatial information of the feature matrix through the stacking of blocks and attention modules. This process optimizes the relationship between different EEG channels and different spatial features automatically, enabling the model to focus more deeply on the essential signal features of the spatial structure of the EEG. Ultimately, it aims to optimize the performance of the model. On this basis, we propose an epileptic seizure prediction model with a graph attention module incorporating recurrent neural networks (GAMRNN) and use a novel optimization algorithm to train the model, combining multiple layers of convolution and double layers of GRU units to jointly extract the spatiotemporal features of the EEG, as shown in Figure 2. The main contributions of this research are as follows:

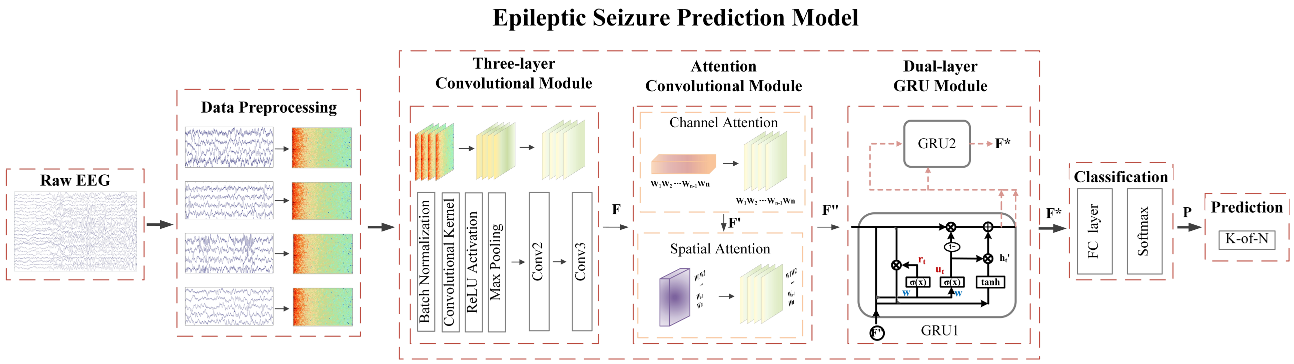


FIGURE 2
The overall workflow of the seizure prediction algorithm involves the following steps: During the data preprocessing stage, the raw EEG signals are transformed into two-dimensional time-frequency representations. These spectrograms are then fed into our proposed seizure prediction model. The initial feature extraction is performed by applying three layers of convolutional operations, resulting in the generation of a feature matrix denoted as *F*. Subsequently, the convolution attention module is applied to the obtained feature map, applying channel-wise and spatial weighting to produce F''. Next, the deep temporal feature extraction and modeling of the feature map are conducted using bidirectional gated recurrent units, resulting in F^*. Following this, the data is passed through fully connected layers and subjected to a softmax function to perform binary classification, predicting the probability *P* of the model classifying the data into interictal and preictal states. Finally, post-processing is employed to make seizure predictions.

(1) We propose a novel epileptic seizure prediction model, GAMRNN, which incorporates a convolutional attention module to focus on important channel and spatial information in EEG signals, enabling more effective capturing of spatio-temporal features.
(2) We utilized the recently introduced Lion optimizer to optimize the model, thereby expediting the convergence rate of the network model training and facilitating the performance of the proposed model in epileptic seizure prediction.
(3) Through ablation experiments on various combined models, we further validated the crucial roles of the Convolutional Block Attention Module and the Lion optimizer in epileptic seizure prediction tasks.

2. Materials and methods

2.1. Epileptic seizure prediction model

Convolutional neural networks (CNNs) have been proven to possess certain advantages in capturing spatial features in data. In contrast, recurrent neural networks (RNNs) have been demonstrated to excel in capturing temporal features of data. Previous studies have also confirmed that combining both CNNs and RNNs is conducive to identifying the temporal and spatial dependencies of epileptic seizure EEG signals. This work employed a multi-layer convolutional neural network (CNN) combined with a two-layer gated recurrent unit (GRU) as the fundamental model for epileptic seizure prediction. To extract more critical temporal and spatial feature information from important channels and spatial regions, we propose to incorporate the CBAM scheme into the base model and name it Graph Attention Module with Recurrent Neural Networks (GAMRNN) for the overall architecture of the seizure prediction model, as illustrated in Figure 3. The CNN is responsible for extracting spatial features from EEG signals, the CBAM module selectively attends to relevant information from input feature maps with larger weights in channels and spatial feature points, and the GRU layer is used to capture the temporal dynamics in the EEG feature map.

2.1.1. Convolutional feature extraction module

Given the limited size of the training dataset and for the sake of model simplicity, we employed a straightforward and shallow three-layer CNN architecture. The model consists of three-layer convolutional blocks for feature extraction. Each block comprises a batch normalization with a RELU activation function, followed by a max pooling layer. To ensure uniform input distribution across layers, batch normalization is applied between each layer, irrespective of the preceding layer's operations.The convolutional layer employs 16 kernels of size $n \times 5 \times 5$, 32 kernels of size 3×3, and 64 kernels of size 3×3, where n represents the number of channels in the EEG signal. The stride for each kernel is $1 \times 2 \times 2$, 1×1, and 1×1, respectively. In order to enhance the performance of the epilepsy seizure prediction task and mitigate the risk of overfitting, L2 regularization terms were incorporated into each convolutional layer. This regularization technique promotes weight values to be smaller and encourages a balanced distribution of weights. Consequently, it improves the convergence speed and stability of the model, thus aiding in accurate epilepsy seizure prediction. The max pooling layer has a size of 2×2, which is used to reduce the number of computations and prevent overfitting during model training. After the initial feature extraction, a feature map of size $64 \times 2 \times 5$ is obtained.

2.1.2. Attention enhancement module

In a seizure prediction system, focal epileptic EEG signals originate from one or multiple scalp electrodes, propagate and gradually spread to multiple electrodes and brain regions. They are characterized by overlapping and interfering waveforms. Some electrodes may be located in more relevant or active pathological areas, while others may be in less related or less active brain

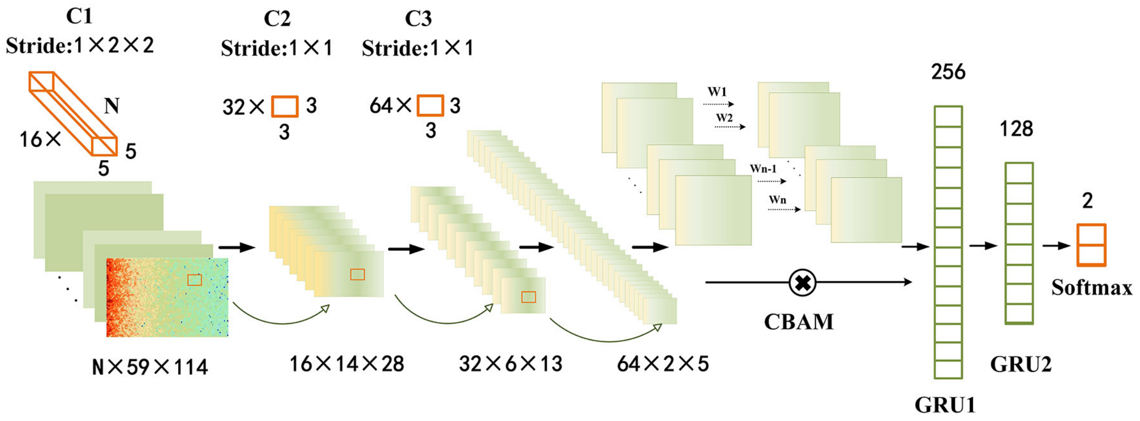


FIGURE 3
Epileptic seizure prediction model: the figure depicts the comprehensive framework of our proposed model for seizure prediction (GAMRNN). The input consists of 30-s windows of preprocessed EEG signals obtained through STFT transformation. The model begins with three convolutional blocks named C1, C2, and C3, serving as the initial feature extraction modules. Each block consists of a batch normalization layer, a convolutional layer with ReLU activation function, and a max pooling layer. C1 has 16 three-dimensional kernels of size $n \times 5 \times 5$, where n represents the number of channels in the original EEG signal, and the stride is $1 \times 2 \times 2$. After the convolution operation, the results are passed through a ReLU activation function, followed by max pooling with a shape of $1 \times 2 \times 2$ to perform downsampling. The operations in C2 and C3 are the same, with 32 and 64 convolutional kernels, respectively. The kernel size is 3×3, and the stride is 1×1. Both C2 and C3 also employ max pooling with a shape of 2×2 for downsampling.Next, the extracted feature maps are subjected to channel and spatial attention-weighted operations using the CBAM module. The input and output feature maps have the same shape of $64 \times 2 \times 5$. Subsequently, the feature maps are flattened and reshaped, and inputted into the first gated recurrent unit (GRU) layer with 256 units, followed by a fully connected layer with sigmoid activation function. The output is then fed into the second GRU layer with 128 units, and finally through two fully connected layers with 2 units and softmax activation function for classification. Two dropout layers with a dropout rate of 0.5 are placed before the two fully connected layers.

regions. Some electrodes may be located in more relevant or active pathological areas, while others may be in less relevant or less active brain regions. Therefore, the importance of signals varies among electrodes. In this case, the attention module can assign different weights to different electrodes and features, allowing the system to focus on essential electrodes or features.

A previous study has investigated using the CBAM module, integrated after batch-normalized long short-term memory (Ma et al., 2021) (BNLSTM) networks, for end-to-end seizure prediction based on raw EEG data. By introducing the attention mechanism, the system may capture the key channels and features more related to seizure events, thereby improving prediction performance. In this experiment, we placed the CBAM module after the three convolutional layers, allowing feature selection to be performed on the already processed feature maps. This approach ensures that the selected features are more accurate and representative, thereby enhancing the performance and effectiveness of the seizure prediction system.

CBAM consists of two modules, namely Channel Attention (Sun et al., 2019) and Spatial Attention (Chen et al., 2017), as shown in Figure 4. The feature map obtained after the convolution layer has the shape $F \in R^{C \times H \times W}$ (where C is the number of channels and H and W are the height and width of the feature map obtained after convolution). For each channel, we set the convolution module as a 2D convolution kernel, and the feature map obtained through channel attention is C_E, while that obtained through spatial attention is S_E.

(a) Channel attention weighting mechanism

The convolutional operation produces a feature map $F \in R^{C \times H \times W}$, where C denotes the number of channels, and H and W refer to the height and width of the feature map. This feature map is initially fed into the Channel Attention module as part of the CBAM module. Channel Attention compresses the feature map along the channel dimension and calculates weight coefficients for each channel. The output is a feature map with weight coefficients, where the dimension of the feature map remains the same as that of the input feature map. In order to improve computational efficiency, the input feature map is globally max-pooled and averaged-pooled to compress the feature map. The pooling resulted in obtaining two different feature descriptions that represent the spatial background features of the data. A channel-wise feature map of size $C_E(F) \in R^{C \times 1 \times 1}$ is obtained through a shared fully connected layer. The two obtained feature matrix mappings are added and passed through a sigmoid activation function to assign proper weights (between 0 and 1) to each input channel C. Finally, the weight matrix is multiplied by the input feature layer. Although the channel attention module assigns weights to the channels of the feature matrix obtained through convolutional operations, it represents the reorganization and integration of the original EEG electrode channels, which implies that the channel attention module assists the model in extracting more important channel feature information from the EEG signals.

In this study, the channel attention module reallocates the importance and correlation of each channel in the EEG signal by generating weight coefficients for each channel based on the convolutional operations of the EEG electrode leads. The specific operation is shown in the formula below:

$$C_E(F) = \mathrm{sigmoid}(\mathrm{Conv2D}(\mathrm{MaxPool}(F)) + \mathrm{Conv2D}(\mathrm{AvgPool}(F))) \quad (1)$$

$$F^{'} = C_E(F) \times F \quad (2)$$

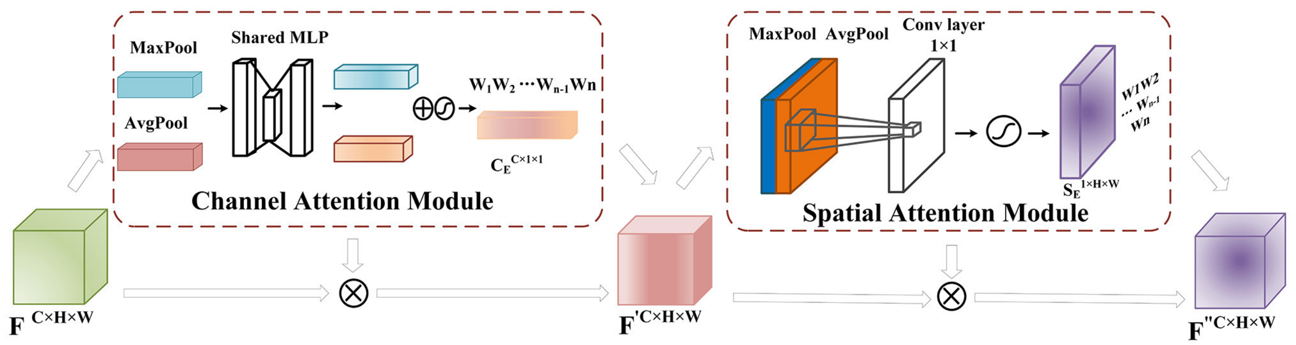


FIGURE 4
CBAM Module: After the feature extraction, the feature matrix $F \in R^{C\times H\times W}$ is obtained. In the channel attention module, global and max pooling operations are applied to F to obtain a $C \times 1 \times 1$ tensor. This tensor is then passed through two fully connected layers with a ReLU activation function, reducing the channel dimension to one-fourth of the original number and then restoring it to the original dimension to obtain the feature matrix mapping. The mapping is combined with F using a sigmoid activation function to assign weights $C_E(F) \in R^{C\times 1\times 1}$ to each input channel, representing channel importance levels. The obtained weight matrix is multiplied by the original input feature matrix to obtain $F^{'}$. In the spatial attention module, $F^{'}$ undergoes global average and max pooling operations, and the resulting tensors are stacked and processed with a 1×1 convolution. The sigmoid activation function maps the spatial feature weights to the range [0,1], obtaining the weights $S_E(F) \in R^{1\times H\times W}$ that represent the importance of each feature point. These weights are multiplied by $F^{'}$ to obtain the weighted feature map $F^{''}$.

(b) Spatial attention weighting mechanism

The spatial attention further extracts features from EEG data at the convolutional level, aiming to preserve the spatiotemporal information of EEG signals as much as possible. The input feature map $F^{'} \in R^{C\times H\times W}$ undergoes max pooling and average pooling operations at each feature point along the spatial dimensions. Then, the two results are stacked along the channel dimension. A 1×1 convolutional layer is applied to adjust the channel dimension to 1, and a sigmoid activation function is used to obtain weight values (between 0 and 1) for each feature point on the feature map. Finally, the weight matrix $S_E(F^{'}) \in R^{1\times H\times W}$is multiplied by the original feature map to obtain the feature map $F^{''}$. The convolutional layer adaptively learns features for each channel input, enabling the network to focus more on meaningful features in the signal and improve the accuracy of seizure prediction. The specific process is shown in the following formula:

$$S_E(F^{'}) = \text{sigmoid}(\text{Conv}(\text{concat}([\text{MaxPool}(F^{'}), \text{AvgPool}(F^{'})]))) \tag{3}$$

$$F^{''} = S_E(F^{'}) \times F^{'} \tag{4}$$

2.1.3. Temporal modeling and classification module

The Gated Recurrent Unit (GRU; Chung et al., 2014) is an advancement over the Long Short-Term Memory (LSTM) model, offering a more streamlined architecture. It incorporates two gate mechanisms to regulate the flow and forgetting of preceding temporal information, effectively addressing the issue of vanishing gradients encountered in recurrent neural networks. Moreover, GRU exhibits enhanced capability in capturing long-term dependencies inherent in sequential data, making it well-suited for analyzing time-series signals. In our study, we employ a dual-layer GRU network to comprehensively analyze the extracted feature matrix $F^{''}$, which allowed us to delve deeper into the temporal features of the electroencephalography (EEG) signals associated with seizure activity F^{*}, thereby facilitating a more precise and accurate classification.

Specifically, in the GRU module, the hidden state h_{t-1} represents the temporal information from the previous time step, while x_t represents the current time step's input feature matrix. This study defines the time steps based on the sequential order of the input feature matrix $F^{''}$. The influence of the previous hidden state h_{t-1} on the current time step is controlled by the reset gate r_t, as shown in the following formula:

$$r_t = \text{sigmoid}(h_{t-1}W_{rh} + x_tW_{rx} + b_r) \tag{5}$$

where W_{rh} and W_{ry} represent the weight matrices of the reset gate, and b_r is the bias matrix with a size equal to the number of hidden units nh. Moreover, the update gate $u_t \in R^{1\times nh}$ is responsible for controlling the balance between the previous hidden state and the current input at each time step, determining the extent to which the previous hidden state is retained and fused with the current input feature.

$$u_t = \text{sigmoid}(h_{t-1}W_{uh} + x_tW_{ux} + b_u) \tag{6}$$

where weight matrices W_{uh} and W_{ux} represent the weights of the update gate, and b_u is equal to the number of hidden units nh. The temporary hidden state $h_t^{'}$ at time step t is obtained by element-wise multiplication.

$$h_t^{'} = \tanh(h_{t-1}W_{hh} \times r_t + x_tW_{xh}) \tag{7}$$

where weight matrices W_{hh} and W_{xh} are used, along with the hyperbolic tangent activation function tanh, to control the flow of information through the reset gate r_t, which determines the degree to which the previous hidden state is retained. Finally, by utilizing the update gate u_t, the new hidden state h_t is computed through a linear combination of the previous hidden state h_{t-1} and the current state $h_t^{'}$, as shown in the following equation:

$$h_t = (1 - u_t) \times h_{t-1} + u_t \times h_t^{'} \tag{8}$$

In summary, the distinctive feature of the dual-layer GRU module in predicting epileptic seizures lies in its effective integration of temporal information and modulation of information flow through gating mechanisms. The dual-layer GRU structure in this study consists of 256 and 128 units, with a dropout rate of 0.5 to mitigate overfitting. The first GRU layer learns temporal dependencies and sequential relationships of neighboring feature maps from $F^{''}$. The second GRU layer captures deeper long-term dependencies and contextual information using the hidden state from the first layer, resulting in the feature matrix F^*. By modeling and synthesizing temporal features, the dual-layer GRU module effectively utilizes the features extracted by the convolutional layer and CBAM module, enhancing the classification accuracy of seizure onset and interictal data and improving the prediction model's performance.

Following the GRU layers are two fully connected layers and two Dropout layers. The first fully connected layer has 64 neurons and uses the sigmoid activation function, taking the output of the Dropout1 layer as input. The second fully connected layer consists of 2 neurons, taking the output of the Dropout2 layer as input. Finally, the Activation layer is used to pass the final softmax output to the output layer of the model, completing the classification task.

2.2. Lion optimizer

During the model training process, we employed a recently proposed optimization algorithm called the Lion optimizer (Chen et al., 2023), developed by researchers from Google and UCLA. Unlike adaptive optimizers like Adam and SGD, the Lion optimizer only requires momentum tracking and utilizes symbolic operations to compute updates, leading to fewer hyperparameters and simpler computations. It has shown superior performance to traditional optimization algorithms when applied to deep learning models in tasks like image classification while accelerating the model training process. Thus, in our research, we introduced the Lion optimizer in the context of seizure prediction models and conducted comparative experiments with the Adam optimizer to assess its impact on model performance.

3. Experiment and results

In the experimental section, we firstly performed preprocessing on the EEG dataset, including splitting the raw data into 30-second windows, removing noise and artifacts, and transforming the EEG data into time-frequency spectrograms. Secondly, the data was partitioned into training and testing sets and fed into the model for training. Meanwhile, we utilized the Lion optimization algorithm to further optimize the model's training process. Thirdly, post-processing operations were applied to the obtained classification results to predict seizure occurrences, and various metrics were used to evaluate the model's seizure prediction performance. Finally, several ablation experiments were conducted to individually assess the impact of each component on the overall model's predictive performance.

TABLE 1 Detailed information of seizure subjects in CHB-MIT dataset.

Subject no.	Age	Gender	Records	Seizure onset
Chb01	11	F	42	7
Chb02	11	M	36	3
Chb03	14	F	38	7
Chb04	22	M	42	4
Chb05	7	F	39	5
Chb06	1.5	F	18	9
Chb07	14.5	F	19	3
Chb08	3.5	M	20	5
Chb09	10	F	19	4
Chb10	3	M	25	7
Chb11	12	F	35	3
Chb12	2	F	24	21
Chb13	3	F	33	12
Chb14	9	F	26	8
Chb15	16	M	40	20
Chb16	7	F	19	10
Chb17	12	F	21	3
Chb18	18	F	36	6
Chb19	19	F	30	3
Chb20	6	F	29	8
Chb21	13	F	33	4
Chb22	9	F	31	3
Chb23	6	F	9	7

3.1. Dataset

In this experiment, we utilized the CHB-MIT dataset to validate the seizure prediction performance of the proposed model (Goldberger et al., 2000). The dataset consists of scalp electroencephalogram (EEG) recordings from 23 pediatric epilepsy patients, collected through collaboration between the Massachusetts Institute of Technology (MIT) and Boston Children's Hospital. The EEG data were sampled at a rate of 256 Hz and acquired using 22 electrodes placed according to the international 10–20 system for EEG electrode placement. The dataset spans approximately 1136 hours of continuous EEG signal activity and includes 198 epileptic seizure events. The patients' ages range from 1.5 to 22 years. Detailed information about the dataset is provided in Table 1. The CHB-MIT public dataset provides expert annotations indicating the start and end times of seizure events. In this study, we define the interictal period as a time interval of at least 4 h before and after the seizure, following the standard proposed by Truong et al. (2018) for seizure prediction research, providing a reference for comparison with their method. Additionally, we excluded the cases with more than ten seizures in the dataset, as their seizure occurrences are too close in time and the prediction results are less meaningful for these patients.

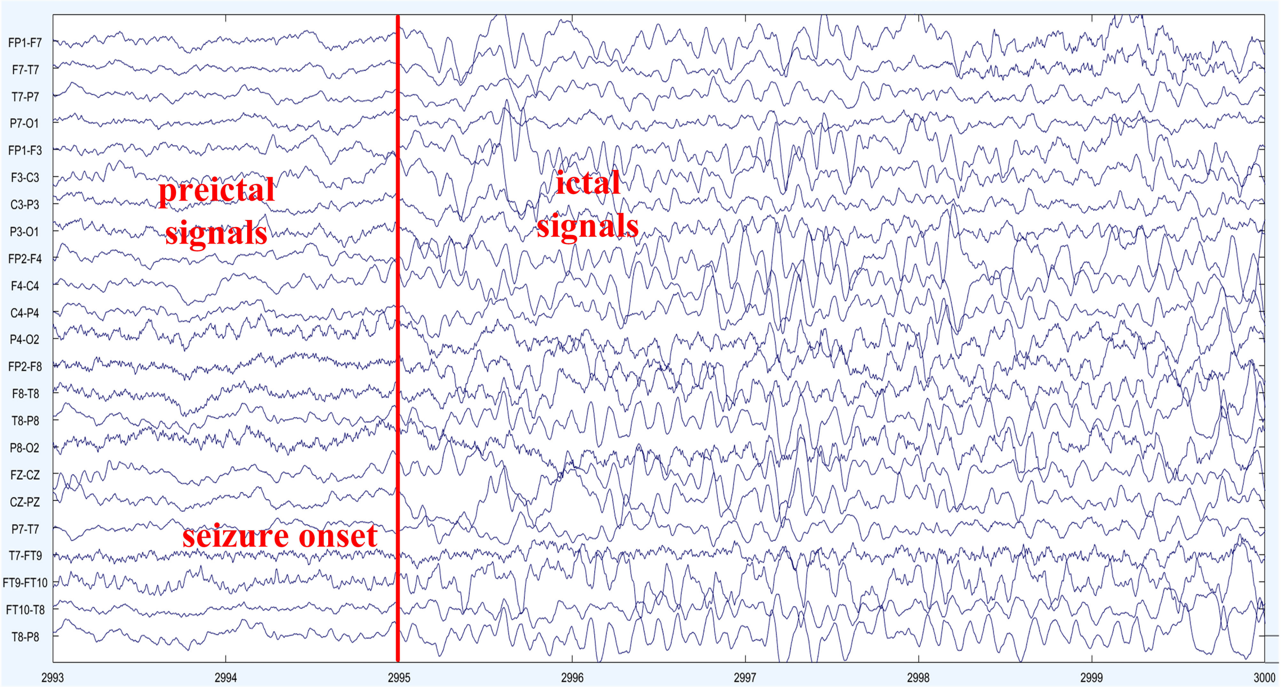


FIGURE 5
EEG signal of patient Chb01: the depicted figure showcases the scalp electroencephalographic (EEG) recordings obtained from subject Chb01 during a seizure episode. The onset of the seizure activity becomes evident at 2995 s, marked by a prominent escalation in EEG frequency and the emergence of complex waveforms characterized by irregularities and spikes. These distinct alterations in the EEG signal morphology signify the initiation of a seizure event, indicating a significant disruption in the underlying neural activity. The observed contrast in EEG patterns between the preictal and ictal periods highlights the pronounced impact of the seizure activity on the brain's electrical dynamics.

To facilitate comparison with related experiments, we evaluated the epilepsy seizure prediction model in detail using data from 13 patients. Figure 5 shows the EEG segments of a seizure event in patient chb01.

3.2. Data preprocessing

The raw EEG signals are characterized by a large quantity and continuous long duration, making them unsuitable for direct input into convolutional neural networks for feature extraction. Therefore, data preprocessing is required. Firstly, preictal and interictal data are extracted separately from the original EEG data. Subsequently, the data is splitted into 30-second windows, and the short-time Fourier transform (STFT) technique is employed to transform the raw EEG signals into two-dimensional spectrograms with frequency and time axes. The transformation helps retain crucial information from the original signals (Truong et al., 2019; Muhammad Usman et al., 2020). During data collection, the dataset is contaminated with 60 Hz power line noise. To address this issue, bandpass filtering is applied to remove frequency components between 57 Hz–63 Hz and 117 Hz–123 Hz, along with excluding the 0 Hz component.

Due to the uneven distribution of the two classes in the dataset, namely, the number of preictal data is significantly smaller than the number of interictal data in a single EEG recording of a seizure episode, it is likely that the model may not learn sufficient useful features due to the scarcity of one class during training, ultimately affecting the classification accuracy. To overcome this data imbalance issue, we employ the overlapping sampling technique along the temporal axis of the EEG signal, generating additional preictal samples using a sliding window of 30 s. After preprocessing, the spectrograms are fed into the GAMRNN model for feature extraction and classification. Through extensive training, the model learns the discriminative features of seizure EEG signals and performs sample classification into preictal and interictal states.

3.3. Experimental setting

In order to train the model and learn relevant features from the preprocessed dataset, it is necessary to partition the dataset into training and testing sets. Here, we employed the leave-one-out cross-validation method. For a subject with *N* occurrences of seizures in their data records, N-1 seizure interictal and preictal segments were concatenated as the training set, while the remaining occurrence of seizure interictal and preictal segments were used as the testing set. Furthermore, 75% of the training set data was utilized for training the model, while the remaining 25% was used as a validation set to assess the learning and training performance of the proposed model and prevent overfitting. We also incorporated an early stopping mechanism during the model training process. If the loss did not improve for ten consecutive epochs, the training was halted prematurely, and the model parameters with the best performance during training were saved. This approach aimed to minimize resource waste and training time.

The experiment was implemented on Python 3 using the Keras and TensorFlow frameworks. The training batch size was set to 64, and the number of epochs was set to 50. For Lion optimizer, the cross-entropy loss was used to compute the training loss. We set the hyperparameter β_1 for exponential decay rate to 0.95, β_2 to 0.98, learning rate η to 0.0001, and weight decay rate λ to 0.015 based on instructions of lion optimizer and our experiences.

3.4. Metrics for epileptic seizure prediction

Seizure prediction horizon (SPH) and seizure onset prediction (SOP) are two temporal periods used to evaluate the results of seizure prediction. SPH refers to the time interval from the onset of an alert to the expected seizure phase, while SOP represents the time span during which the seizure is anticipated to occur. A correct alert within the SPH serves to notify healthcare professionals and family members that a seizure is likely to happen within the subsequent SOP, enabling them to take timely measures. Consistent with Truong et al. (2018), this study sets the SPH to 5 min and the SOP to 30 min. The method for setting SPH and SOP is shown in Figure 6. The criterion for accurate prediction is the occurrence of at least one seizure event during the SOP period following the onset of the alert, while no seizures should occur within the SPH period. False alarms, on the other hand, refer to alerts issued in the absence of any seizures during the SOP period. To reduce false positives, a K-of-N post-processing method is employed (Truong et al., 2018), where an alert is triggered only when K seizure-like segments are identified within a continuous sequence of N segments. In this study, the parameters k = 8 and n = 10 are set, with predictions made every 30 s. Consequently, if more than 4 min of seizure-like segments are identified within a continuous 5 min data segment, an alert is issued.

The performance of the epilepsy seizure prediction model was evaluated using sensitivity (SEN), specificity (SPEC), accuracy (ACC), area under the curve (AUC), and false positive rate per hour (FPR/h) metrics. In typical binary classification tasks, sensitivity, specificity, and accuracy are calculated from the confusion matrix in statistics. Sample prediction can result in four possible scenarios: TP (True Positive), meaning the actual EEG signal data is preictal and the predicted result is also preictal; FP (False Positive), meaning actual interictal signal data is predicted as preictal signal data; TN (True Negative), meaning the predicted data is interictal signal data, and it is indeed interictal signal data; FN (False Negative), meaning actual preictal signal data is predicted as interictal signal data. Based on the confusion matrix, the following metrics can be calculated:

$$Sensitivity = TP/(TP + FN) \tag{9}$$

$$Specificity = TN/(TN + FP) \tag{10}$$

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \tag{11}$$

The ROC (Receiver Operating Characteristic) curve is a graphical representation where the X-axis is the false positive rate (FPR), and the Y-axis is the true positive rate (TPR). Different TPR and FPR values can be obtained and plotted as an ROC curve by changing the classifier's threshold. The AUC (Area Under the Curve) is the area under the ROC curve, with a value between 0.5 and 1. A larger AUC indicates a better performance of the classifier.

4. Results

4.1. General results

Based on the same experimental settings, we conducted a performance evaluation of the GAMRNN seizure prediction model and compared it with the GCRNN prediction model. We selected the same 13 patient data from the CHB-MIT dataset for evaluation on both models. The experimental results are shown in Table 2. We observed and compared the classification and prediction performance of the two models from the aspects of accuracy, sensitivity, and false positive rate, taking the average values of all the subjects' experimental results. From the table, we can conclude that our proposed GAMRNN prediction model demonstrates better seizure prediction performance on most subjects' data, with an average accuracy of 91.73%, which is a 6.44% improvement over the CGRNN prediction model. The GAMRNN model achieved a sensitivity of 88.09% in correctly predicting seizures, showing an ∼6% increase in sensitivity compared to the original model, which indicates that the model successfully captured 56 out of 64 seizures. After incorporating the attention convolution module and optimizing the model using the Lion optimizer in the CGRNN model, the false positive rate decreased from 0.2042 to 0.053/h. Except for patients Chb10 and Chb14, the false positive rate for seizures in other patients approached 0. The improvement in these evaluation metrics has significant practical implications for the daily life of epilepsy patients. Above results confirm that the proposed seizure prediction model can effectively distinguish between preictal and interictal EEG signal data, enabling accurate decisions on whether a seizure will occur in the later stage of the EEG signal, thereby greatly reducing the occurrence of false alarms for seizures.

However, due to various reasons, such as differences in the number of seizures, proximity to seizures, or patient-specific characteristics, the seizure prediction model may not achieve the same prediction performance for every patient. The variance calculated for various metrics of the two seizure prediction models indicates that our proposed model demonstrates greater stability in evaluating the 13 patient datasets compared to the baseline model. The comparative experiments also provide evidence that the attention modules indeed assist the seizure prediction model in focusing more on crucial regions within the feature maps. By incorporating channel and spatial dimensions, the attention modules enable the model to emphasize the essential spatiotemporal features in the EEG signal data while reducing attention to relatively less significant regions. As a result, the overall model performance for classifying two types of seizure EEG signals is enhanced, leading to

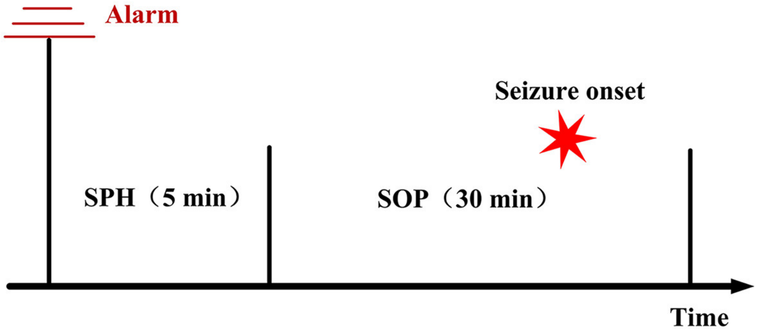


FIGURE 6
Diagram of the Seizure Prediction Period (SPH) and Seizure Onset Period (SOP): in the post-processing phase, a successful prediction of seizure onset by the seizure prediction model is defined as the absence of any seizure events during the SPH period following the onset alert, followed by the occurrence of one or more seizure events during the subsequent SOP period. When no seizure events occur during the SOP period following the onset alert, it is considered a false alarm.

TABLE 2 Seizure detection performance on the CHB-MIT dataset.

	CGRNN			GAMRNN		
Patient	Accuracy	Sensitivity	FPR (/h)	Accuracy	Sensitivity	FPR (/h)
Chb01	0.9337	0.8429	0.057	0.9455	0.9548	0
Chb02	0.9398	0.1611	0	0.9415	0.3652	0
Chb03	0.9313	0.6389	0	0.9417	0.8361	0
Chb05	0.6884	0.2867	0.3468	0.9088	0.88	0.0694
Chb09	0.9814	1	0	0.9945	0.9833	0
Chb10	0.6572	0.4056	0.566	0.7811	0.7389	0.2264
Chb13	0.8793	0.9967	0.2081	0.9141	0.9933	0.0694
Chb14	0.5178	0.8	0.7385	0.7711	0.8433	0.1846
Chb18	0.8525	0.6167	0.2041	0.904	0.9375	0.0408
Chb19	0.9827	0.7889	0	0.9893	0.9322	0
Chb20	0.8952	0.9733	0.1469	0.9544	0.99	0.098
Chb21	0.867	0.8125	0.3134	0.8903	1	0
Chb23	0.9615	0.9833	0.0752	0.9886	0.9967	0
Average	0.8529	0.7159	0.2043	0.9173	0.8809	0.053
Variance	0.0189	0.0723	0.0505	0.0047	0.0280	0.0054

significant improvements in accuracy, sensitivity, and false positive rate evaluations.

4.2. Results of ablation study

Our study conducted two sets of ablation experiments. The first set of experiments aimed to validate the performance enhancement of the GAMRNN model by adding the CBAM module and using the Lion optimizer. Specifically, the CBAM module and Lion optimizer were sequentially added to the model, and their performance on different datasets was compared and analyzed. The second set of experiments aimed to validate the individual effects of the Channel Attention Module (CAM) and Spatial Attention Module (SAM) when applied separately to the model. Additionally, we compared the combination module with the order of CAM and SAM switched to the CBAM module. The accuracy, sensitivity, and specificity results obtained from the two groups of ablation experiments are presented in Table 3.

GAMRNN (CAM only) and GAMRNN (SAM only):We incorporated Channel Attention Module (CAM) and Spatial Attention Module (SAM) separately into the model to assess the individual impacts of these attention mechanisms on model performance. Specifically, when CAM or SAM was added independently to the model, the accuracy remained similar. However, there was approximately a 6% decrease compared to the model using the combined attention mechanism CBAM. Moreover, sensitivity and specificity were lower than the Convolutional Attention Module. These results indicate that utilizing a single attention mechanism alone has a limited impact

TABLE 3 Ablation experimental results.

Methods	Accuracy	Sensitivity	Specificity
GAMRNN (CAM only)	85.71	76.32	86.27
GAMRNN (SAM only)	85.53	76.59	85.59
GAMRNN (Lion only)	87.47	76.17	88.75
GAMRNN (CBAM and Adam)	90.03	83.75	91.52
GAMRNN (CBAM and Lion)	91.73	88.09	92.09

on the performance of the seizure prediction model. However, the predictive performance was significantly enhanced when employing the Convolutional Attention Module that integrates both CAM and SAM and applies them jointly to the model. Additionally, we conducted experiments by interchanging the order of the attention modules (first applying SAM and then CAM) and combining them in the model. While there was a slight improvement in accuracy and specificity, it was not as pronounced as the original CBAM combination, suggesting that the order of combining attention modules within the Convolutional Attention Module has the most significant impact on enhancing the model's performance. In conclusion, the Convolutional Attention Module plays a more prominent role in improving the seizure prediction model than individual CAM and SAM, and the specific order of combining CAM and SAM within the Convolutional Attention Module has the most significant influence on model performance enhancement.

GAMRNN (Lion only): This model is derived from the proposed model by removing the CBAM module, allowing for the evaluation of the epileptic seizure prediction performance without the attention convolutional module. The experimental results demonstrate that the model without CBAM exhibits a significant performance decrease in accuracy, sensitivity, and specificity compared to the proposed prediction model. Specifically, the classification accuracy of interictal and preictal data decreased from 91.73 to 87.47%. The sensitivity of correctly identifying preictal data decreased from 88.09 to 76.17%, and the specificity of correctly identifying interictal data also decreased by 3.34%. These changes in results indicate the crucial role of the CBAM module in the proposed epileptic seizure prediction model, as the model without CBAM shows a significant decrease in classification performance. Therefore, we hypothesize that the inclusion of CBAM in the model allows for further attention to be given to essential channels and spatial feature points within the feature maps after the initial three-layer convolutional feature extraction, thereby aiding the model in focusing on extracting more crucial feature information and enhancing the classification and prediction performance of the model.

GAMRNN (CBAM and Adam): This model is obtained by removing the Lion optimizer from the proposed model and using the Adam optimizer, which is the same as the baseline model, to observe its impact on model training. A comparison reveals that this model also experiences a corresponding decrease in performance in various aspects, although the decrease is not particularly significant. For instance, the average accuracy of the model without the Lion optimizer is only reduced by ∼1.70%, the sensitivity is reduced by ∼4.34%, and the specificity is reduced by 0.57%. During model training, a visual inspection indicates that each epoch takes ∼1–2 s less than the Adam optimizer model. This suggests that the Lion optimizer accelerates the training process and effectively reduces the training loss of the model, thereby enhancing the stability of correct seizure prediction. In summary, the Lion optimizer plays a role in performance evaluation and training for epileptic seizure prediction research tasks. It also lays the foundation for utilizing the Lion optimization algorithm in more complex studies, offering more possibilities for training models in epileptic seizure prediction research.

The above analysis provides a detailed examination of the individual effects of the attention module and the Lion optimizer in the proposed model. The experimental results indicate that incorporating both modules into the research on epileptic seizure prediction enhances the classification and prediction performance of the model. As shown in Figure 7, the AUC results comparison represents the model's ability to accurately classify interictal and preictal data. It can be observed that regardless of whether the Channel Attention Module (CAM) or the Spatial Attention Module (SAM) is individually integrated into the seizure prediction model or if they are combined with interchanged order, the classification performance of the model on most patients data is significantly inferior to the predictive model proposed in this study, which utilizes the Convolutional Attention Module. Figure 8 illustrates the AUC comparison of the CGRNN baseline model and the GAMRNN model, which gradually incorporates both modules, using data from 13 patients. The graph shows that the models achieve good classification performance on most patient data, which becomes more pronounced as the two modules are successively integrated. Among them, the AUC performance on the Chb01, Chb09, and Chb23 data approaches 1. However, the classification performance on the Chb02, Chb10, and Chb14 patient data is relatively lower due to the imbalance in these data categories. However, significant improvements are observed after incorporating the CBAM module and using the Lion optimizer, further demonstrating that these two modules aid in accurately recognizing and classifying imbalanced data. Therefore, the results of the above ablation experiments indicate that the proposed GAMRNN model has better EEG signal classification performance. It combines the STFT spectrogram input with channel weights, simultaneously focusing on the spatial features of the signal, and uses GRU-gated units to extract important temporal information from the features, providing specific advantages in reducing false positives and improving model accuracy.

5. Discussion

With the emergence of various deep learning techniques, they have gradually been applied to predict epileptic seizures. In order to compare our proposed method with other methods on the same dataset to make the comparison more convincing, we selected several studies that evaluated models using the same dataset. Table 4 shows the comparative experimental results. There is no

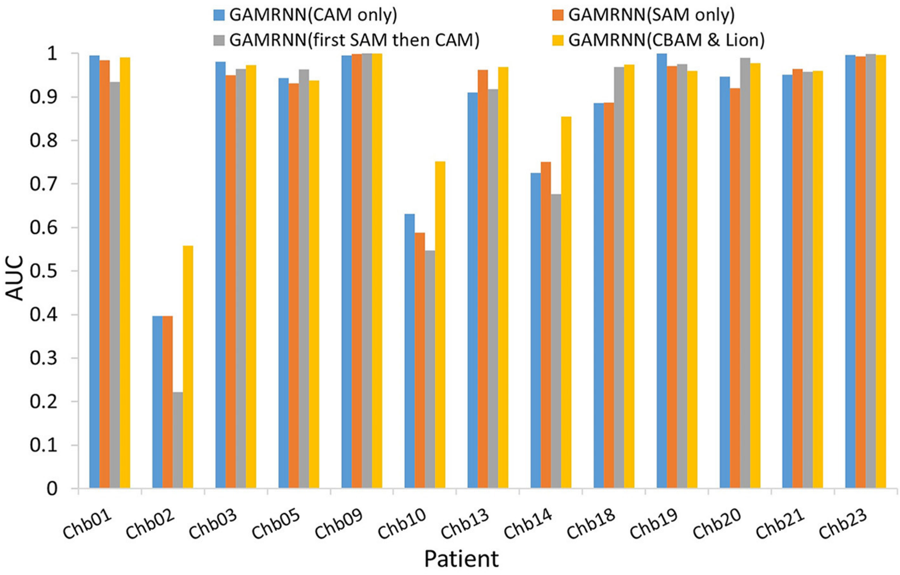


FIGURE 7
Comparison of AUCs with different attention mechanisms added to the models: the bar chart depicts the AUC evaluation results of four models, namely, GAMRNN (CAM only), GAMRNN (SAM only), GAMRNN (first SAM then CAM), and GAMRNN(CBAM and Lion), on 13 patient datasets. The comparison reveals that the proposed GAMRNN model with attention convolutional modules added in the normal sequence exhibits the most distinct and superior classification performance compared to the other three models.

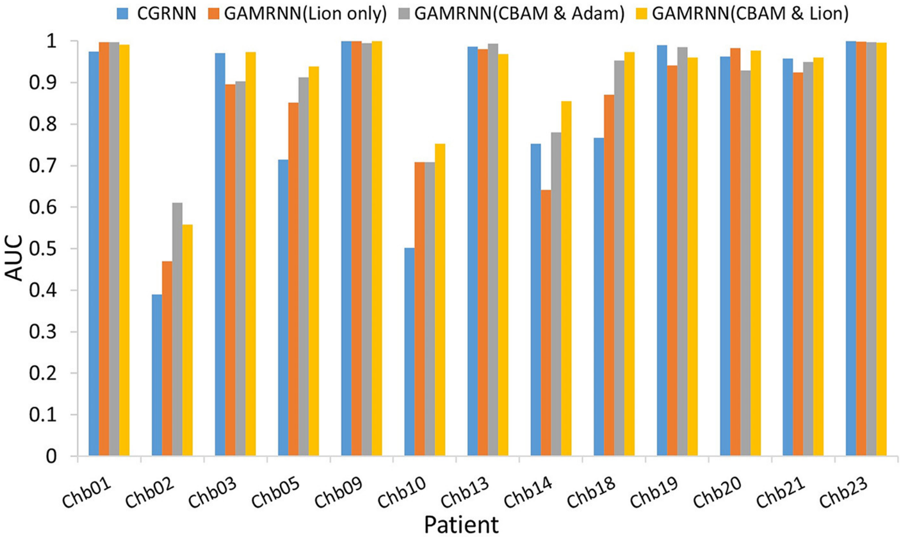


FIGURE 8
Comparison of AUC among different models: the figure depicts the AUC evaluation of four epilepsy seizure prediction models, namely CGRNN, GAMRNN (Lion only), GAMRNN (CBAM and Adam), and GAMRNN (CBAM and Lion), on the CHBMIT 13-patient dataset. AUC represents the classification performance of the prediction models. As shown in the figure, the combination of the cbam module and lion optimizer has a certain effect on the classification performance of the models.

absolute good or bad result because the models proposed by different researchers have differences, and slight changes in each step of epileptic seizure prediction may also lead to experimental differences. Our GAMRNN model is much better than the CNN model proposed by Truong et al. (2018) in all aspects. They used a three-layer convolutional model for feature extraction and achieved a prediction accuracy of 81.2% on CHB-MIT. Affes et al. (2019) proposed a CGRNN model combining three layers of convolution and two layers of gated units, achieving a classification sensitivity of 89.07%. The difference between our CGRNN model and theirs lies in the data preprocessing part, and it can be seen that the attention convolution module we introduced has a positive

TABLE 4 Comparative experimental results.

References	Methods	Accuracy	Sensitivity	FPR (/h)	AUC
Truong et al. (2018)	CNN	–	81.2	0.16	–
Affes et al. (2019)	CGRNN	75.6	89.07	1.6	–
Büyükçakır et al. (2020)	MLP	–	89.8	0.081	–
Zhang et al. (2021)	Bi-LSTM	80.09	86.67	0.26	–
Sun et al. (2021)	CADCNN	–	97.1	0.029	91.7
Proposed model	GAMRNN	91.73	88.09	0.053	91.56

effect. Büyükçakır et al. (2020) utilized the Hilbert decomposition method to decompose scalp EEG data signals from 10 patients in the CHB-MIT dataset into seven components. They achieved a sensitivity of 89.8% and a false alarm rate of 0.081/h using an MLP classifier. Although our proposed method exhibits a slightly lower sensitivity, we achieved a lower false seizure prediction rate. Zhang et al. (2021) extracted the feature of multi-scale sample entropy from 23 EEG signals from the same dataset and used a bidirectional LSTM model to predict the occurrence of epileptic seizures. The prediction accuracy achieved was 80.09%, with an FPR of 0.26/h. In comparison, our model demonstrated relatively superior performance. Sun et al. (2021) also proposed a Channel Attention Dual-input Convolutional Neural Network (CADCNN) that incorporates both time-frequency spectrograms and raw EEG signals as inputs to a convolutional neural network for feature extraction and fusion. By leveraging channel attention mechanisms, their method achieved excellent results, exhibiting superior sensitivity compared to the model proposed in this study but similar AUC performance. Therefore, we hypothesize that the different forms of dual-channel input EEG signals may help improve the accuracy of feature extraction.

Our proposed study features a relatively simple overall model architecture, resulting in lower resource overhead and computational complexity. The total number of training parameters is ∼880,000, including parameters from convolutional kernels, recurrent gating units, and fully connected layers. The experiments were conducted on a server equipped with an RTX 2080 Ti GPU (11 GB of VRAM), and the memory required for the dataset and model source code was ∼40 GB. Training the model on the CHB-MIT dataset, which includes data from 13 patients, took ∼8 h. The training time per patient varied from a few seconds to several tens of seconds per epoch, and the overall training time depended on the number of seizures and recording duration per patient. Despite its simplicity in implementation, this experiment achieved favorable performance, which highlights its relative excellence.

In the process of comparing our proposed method with others, we have reflected on potential issues that may exist. For example, the evaluation of the model on Chb02, Chb10, and Chb14 showed relatively inferior predictive performance compared to other patients. The significant inter-individual variability among patients often results in some individuals having predictable epileptic seizures while others experience unpredictable seizure occurrences. In addition to these factors, this may be closely related to the seizure condition of each patient. The Chb02 patient had only three seizures in all the records, indicating a significant imbalance in the ratio between preictal and interictal data. This imbalance adversely affected the model's ability to learn from preictal data, leading to reduced sensitivity and classification performance in identifying this data type correctly. Similarly, for the Chb10 and Chb14 patients, the relatively dense occurrence of seizure events in the recorded data files resulted in limited interictal periods available for model learning. This limitation affected the model's ability to differentiate between interictal and preictal data, leading to poorer overall classification performance. Therefore, in future research, we intend to employ data augmentation techniques to generate additional EEG data, addressing the issue of data imbalance in epileptic seizure occurrences. This endeavor aims to facilitate the epileptic seizure prediction model in achieving enhanced performance and superior outcomes.

This paper proposes a seizure prediction method based on a recurrent neural network with convolutional attention modules. Firstly, we use multiple layers of convolution to extract spatial information from multi-channel EEG recordings and apply attention mechanisms to focus on specific channels and spatial locations, mimicking the visual perception process of humans. Our model combines two channel attention modules and a spatial attention module to reassign weights to each feature channel and point in the convolution process. Two gated recurrent units are added after the attention modules to perform deep feature extraction on the temporal sequence. Experimental results show that our proposed method achieves high accuracy, sensitivity, and low false positive rate in cross-validation evaluation on the dataset, which further proves the potential of attention mechanism modules and the Lion optimization algorithm in seizure EEG prediction research, providing ideas and insights for future research in this field. In addition, we plan to explore methods for addressing imbalanced data issues and evaluate the proposed model's performance on more scalp EEG and intracranial EEG datasets to improve its generalization capability.

Data availability statement

The datasets analyzed for this study can be found in the (CHB-MIT) (http://www.physionet.org).

Ethics statement

The studies involving humans were approved by Clinical Investigations at the Beth Israel Deaconess Medical Center (BIDMC), Boston, Massachusetts, USA, and the Massachusetts Institute of Technology (MIT), Cambridge, Massachusetts, USA. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and institutional requirements.

Author contributions

WJ and HJ: conceptualization. HJ and TiX: methodology. HJ and ZY: investigation. YL: formal analysis. HJ, TiX, and TXu: writing. TXue, BC, and WJ: supervision. HJ: funding acquisition. All authors had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by National Natural Science Foundation Grant (Program No. 62106189) to HJ, and Natural Science Basic Research Plan in Shaanxi Province of China (Program No. 2021JQ-674) to HJ. The funding sources had no involvement in the study design.

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

Affes, A., Mdhaffar, A., Triki, C., Jmaiel, M., and Freisleben, B. (2019). "A convolutional gated recurrent neural network for epileptic seizure prediction," in *How AI Impacts Urban Living and Public Health*, eds J. Pagán, M. Mokhtari, H. Aloulou, B. Abdulrazak, and M. F. Cabrera (Cham: Springer International Publishing), 85–96. doi: 10.1007/978-3-030-32785-9_8

Artameeyanant, P., Sultornsanee, S., and Chamnongthai, K. (2017). Electroencephalography-based feature extraction using complex network for automated epileptic seizure detection. *Expert Syst.* 34:e12211. doi: 10.1111/exsy.12211

Bandarabadi, M., Teixeira, C. A., Rasekhi, J., and Dourado, A. (2015). Epileptic seizure prediction using relative spectral power features. *Clin. Neurophysiol.* 126, 237–248. doi: 10.1016/j.clinph.2014.05.022

Bou Assi, E., Nguyen, D. K., Rihana, S., and Sawan, M. (2017). Towards accurate prediction of epileptic seizures: a review. *Biomed. Signal Process. Control* 34, 144–157. doi: 10.1016/j.bspc.2017.02.001

Büyükçakır, B., Elmaz, F., and Mutlu, A. Y. (2020). Hilbert vibration decomposition-based epileptic seizure prediction with neural network. *Comput. Biol. Med.* 119:103665. doi: 10.1016/j.compbiomed.2020.103665

Chen, X., Liang, C., Huang, D., Real, E., Wang, K., Liu, Y., et al. (2023). Symbolic discovery of optimization algorithms. *arXiv [Preprint].* arXiv: 2302.00675. Available online at: https://arxiv.org/pdf/2302.06675.pdf

Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., et al. (2017). "SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Honolulu, HI), 6298–6306. doi: 10.1109/CVPR.2017.667

Choi, W., Kim, M.-J., Yum, M.-S., and Jeong, D.-H. (2022). Deep convolutional gated recurrent unit combined with attention mechanism to classify pre-ictal from interictal EEG with minimized number of channels. *J. Pers. Med.* 12:763. doi: 10.3390/jpm12050763

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv Preprint.*

Fei, K., Wang, W., Yang, Q., and Tang, S. (2017). Chaos feature study in fractional Fourier domain for preictal prediction of epileptic seizure. *Neurocomputing* 249, 290–298. doi: 10.1016/j.neucom.2017.04.019

Ghosh, A., Sarkar, A., Das, T., and Basak, P. (2017). "Pre-ictal epileptic seizure prediction based on ecg signal analysis," in *2017 2nd International Conference for Convergence in Technology (I2CT)*, 920–925. doi: 10.1109/I2CT.2017.8226263

Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P., Mark, R., et al. (2000). Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *Circulation* 101, E215–E220. doi: 10.1161/01.CIR.101.23.e215

Hu, W., Cao, J., Lai, X., and Liu, J. (2019). Mean amplitude spectrum based epileptic state classification for seizure prediction using convolutional neural networks. *J. Ambient Intell. Hum. Comput.* 1–11. doi: 10.1007/s12652-019-01220-6

Hu, X., Yuan, S., Xu, F., Leng, Y., Yuan, K., and Yuan, Q. (2020). Scalp eeg classification using deep BI-LSTM network for seizure detection. *Comput. Biol. Med.* 124:103919. doi: 10.1016/j.compbiomed.2020.103919

Hussain, L. (2018). Detecting epileptic seizure with different feature extracting strategies using robust machine learning classification techniques by applying advance parameter optimization approach. *Cogn. Neurodyn.* 12, 271–294. doi: 10.1007/s11571-018-9477-1

Joshi, V., Pachori, R. B., and Vijesh, V. A. (2014). Classification of ictal and seizure-free EEG signals using fractional linear prediction. *Biomed. Signal Process. Control* 9, 1–5. doi: 10.1016/j.bspc.2013.08.006

Karthick, P., Tanaka, H., Khoo, H. M., and Gotman, J. (2018). Prediction of secondary generalization from a focal onset seizure in intracerebral EEG. *Clin. Neurophysiol.* 129, 1030–1040. doi: 10.1016/j.clinph.2018.02.122

Khalid, M. I., Aldosari, S. A., Alshebeili, S. A., Alotaiby, T. N., and El-Samie, F. E. A. (2015). "Online adaptive seizure prediction algorithm for scalp EEG," in *2015 International Conference on Information and Communication Technology Research (ICTRC)* (Abu Dhabi), 44–47. doi: 10.1109/ICTRC.2015.7156417

Li, S., Zhou, W., Yuan, Q., Geng, S., and Cai, D. (2013). Feature extraction and recognition of ictal EEG using EMD and SVM. *Comput. Biol. Med.* 43, 807–816. doi: 10.1016/j.compbiomed.2013.04.002

Liu, G., Zhou, W., and Minxing, G. (2019). Automatic seizure detection based on s-transform and deep convolutional neural network. *Int. J. Neural Syst.* 30:1950024. doi: 10.1142/S0129065719500242

Ma, M., Cheng, Y., Wang, Y., Li, X., Mao, Q., Zhang, Z., et al. (2021). Early prediction of epileptic seizure based on the BNLSTM-CASA model. *IEEE Access.* 9, 79600–79610. doi: 10.1109/ACCESS.2021.3084635

Mohan, N., P.P., M. S., Sulthan, N., Khan, K. A., and S., S. (2018). "Automatic epileptic seizure prediction in scalp EEG," in *2018 International Conference on Intelligent Circuits and Systems (ICICS)* (Phagwara), 275–280. doi: 10.1109/ICICS.2018.00063

Muhammad Usman, S., Khalid, S., and Aslam, M. H. (2020). Epileptic seizures prediction using deep learning techniques. *IEEE Access* 8, 39998–40007. doi: 10.1109/ACCESS.2020.2976866

Natu, M., Bachute, M., Gite, S., Kotecha, K., and Vidyarthi, A. (2022). Review on epileptic seizure prediction: machine learning and deep learning approaches. *Comput. Math. Methods Med.* 2022:7751263. doi: 10.1155/2022/7751263

Raghu, S., Sriraam, N., Rao, S. V., Hegde, A. S., and Kubben, P. L. (2019). Automated detection of epileptic seizures using successive decomposition index and support vector machine classifier in long-term EEG. *Neural Comput. Appl.* 32, 8965–8984. doi: 10.1007/s00521-019-04389-1

Rasekhi, J., Mollaei, M. R. K., Bandarabadi, M., Teixeira, C. A., and Dourado, A. (2013). Preprocessing effects of 22 linear univariate features on the performance of seizure prediction methods. *J. Neurosci. Methods* 217, 9–16. doi: 10.1016/j.jneumeth.2013.03.019

Shasha, Z., Chen, D., Ranjan, R., Hengjin, k., Tang, Y., and Zomaya, A. (2021). A lightweight solution to epileptic seizure prediction based on EEG synchronization measurement. *J. Supercomput.* 77, 1–19. doi: 10.1007/s11227-020-03426-4

Song, Y., Crowcroft, J., and Zhang, J. (2012). Automatic epileptic seizure detection in EEGs based on optimized sample entropy and extreme learning machine. *J. Neurosci. Methods* 210, 132–146. doi: 10.1016/j.jneumeth.2012.07.003

Sun, B., Lv, J.-J., Rui, L.-G., Yang, Y.-X., Chen, Y.-G., Ma, C., et al. (2021). Seizure prediction in scalp EEG based channel attention dual-input convolutional neural network. *Phys. A Stat. Mech. Appl.* 584:126376. doi: 10.1016/j.physa.2021.126376

Sun, S., Zhao, B., Chen, X., Mateen, M., and Wen, J. (2019). Channel attention networks for image translation. *IEEE Access* 7, 95751–95761. doi: 10.1109/ACCESS.2019.2926882

Truong, N., Nguyen, A., Kuhlmann, L., Bonyadi, M., Yang, J., Ippolito, S., et al. (2018). Convolutional neural networks for seizure prediction using intracranial and scalp electroencephalogram. *Neural Netw.* 105, 104–111. doi: 10.1016/j.neunet.2018.04.018

Truong, N. D., Zhou, L., and Kavehei, O. (2019). "Semi-supervised seizure prediction with generative adversarial networks," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2369–2372. doi: 10.1109/EMBC.2019.8857755

Tsiouris, K. M., Pezoulas, V. C., Zervakis, M. E., Konitsiotis, S., Koutsouris, D. D., and Fotiadis, D. I. (2018). A long short-term memory deep learning network for the prediction of epileptic seizures using EEG signals. *Comput. Biol. Med.* 99, 24–37. doi: 10.1016/j.compbiomed.2018.05.019

Varatharajah, Y., Iyer, R. K., Berry, B. M., Worrell, G. A., and Brinkmann, B. H. (2017). Seizure forecasting and the preictal state in canine epilepsy. *Int. J. Neural Syst.* 27:1650046. doi: 10.1142/S0129065716500465

Varnosfaderani, S. M., Rahman, R., Sarhan, N. J., Kuhlmann, L., Asano, E., Luat, A., et al. (2021). "A two-layer lstm deep learning model for epileptic seizure prediction," in *2021 IEEE 3rd International Conference on Artificial Intelligence Circuits and Systems (AICAS)* (Washington, DC), 1–4. doi: 10.1109/AICAS51828.2021.9458539

Wang, J. Q., Fang, W., 2, and Sheng, V. S. (2022). Prediction of epileptic EEG signal based on SECNN-LSTM. *J. N. Media* 4, 73–84. doi: 10.32604/jnm.2022.027040

Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. (2018). "CBAM: convolutional block attention module," in *Computer Vision? ECCV 2018: 15th European Conference* (Berlin; Heidelberg: Springer-Verlag), 3–19. doi: 10.1007/978-3-030-01234-2_1

Wu, X., Du, Z., Guo, Y., and Fujita, H. (2019). Hierarchical attention based long short-term memory for Chinese lyric generation. *Appl. Intell.* 49, 1–9. doi: 10.1007/s10489-018-1206-2

Yang, Y., Zhou, M., Niu, Y., Li, C., Cao, R., Wang, B., et al. (2018). Epileptic seizure prediction based on permutation entropy. *Front. Comput. Neurosci.* 12:55. doi: 10.3389/fncom.2018.00055

Yıldırım, Ö., Baloglu, U. B., and Acharya, U. R. (2018). A deep convolutional neural network model for automated identification of abnormal EEG signals. *Neural Comput. Appl.* 32, 15857–15868. doi: 10.1007/s00521-018-3889-z

Yu, Z., Nie, W., Zhou, W., Xu, F., Yuan, S., and Leng, Y. (2020). Epileptic seizure prediction based on local mean decomposition and deep convolutional neural network. *J. Supercomput.* 76, 1–15. doi: 10.1007/s11227-018-2600-6

Yuan, S., Zhou, W., and Chen, L. (2017). Epileptic seizure prediction using diffusion distance and Bayesian linear discriminate analysis on intracranial EEG. *Int. J. Neural Syst.* 28:1750043. doi: 10.1142/S0129065717500435

Zhang, Q., Ding, J., Kong, W., Liu, Y., Wang, Q., and Jiang, T. (2021). Epilepsy prediction through optimized multidimensional sample entropy and BI-LSTM. *Biomed. Signal Process. Control* 64:102293. doi: 10.1016/j.bspc.2020.102293

Zhang, Y., Yang, R., and Zhou, W. (2020). Roughness-length-based characteristic analysis of intracranial EEG and epileptic seizure prediction. *Int. J. Neural Syst.* 2020:2050072. doi: 10.1142/S0129065720500720

Zhu, Y., Wang, J., Xie, L., and Zheng, L. (2018). "Attention-based pyramid aggregation network for visual place recognition," in *Proceedings of the 26th ACM International Conference on Multimedia* (New York, NY: Association for Computing Machinery), 99–107. doi: 10.1145/3240508.3240525

frontiers | Frontiers in Neuroscience

TYPE Methods
PUBLISHED 25 August 2023
DOI 10.3389/fnins.2023.1246769

OPEN ACCESS

EDITED BY
Zhiqiang Tian,
Xi'an Jiaotong University, China

REVIEWED BY
Bo Li,
Nanchang Hangkong University, China
Yang Yunfeng,
Northeast Petroleum University, China

*CORRESPONDENCE
Jingfei Wu
✉ afeiwu@126.com
Zhijie Wen
✉ wenzhijie@shu.edu.cn

†These authors have contributed equally to this work and share first authorship

RECEIVED 24 June 2023
ACCEPTED 11 August 2023
PUBLISHED 25 August 2023

CITATION
Fang J, Lv N, Li J, Zhang H, Wen J, Yang W, Wu J and Wen Z (2023) Decoupled learning for brain image registration. *Front. Neurosci.* 17:1246769.
doi: 10.3389/fnins.2023.1246769

Decoupled learning for brain image registration

Jinwu Fang[1,2,3†], Na Lv[4†], Jia Li[1], Hao Zhang[5], Jiayuan Wen[6], Wan Yang[5], Jingfei Wu[7*] and Zhijie Wen[5*]

[1]Institute of Infectious Disease and Biosecurity, School of Public Health, Fudan University, Shanghai, China, [2]China Academy of Information and Communication Technology, Beijing, China, [3]Industrial Internet Innovation Center (Shanghai) Co., Ltd., Shanghai, China, [4]School of Health and Social Care, Shanghai Urban Construction Vocational College, Shanghai, China, [5]Department of Mathematics, School of Science, Shanghai University, Shanghai, China, [6]College of Intelligence and Computing, Tianjin University, Tianjin, China, [7]School of Economics, Shanghai University, Shanghai, China

Image registration is one of the important parts in medical image processing and intelligent analysis. The accuracy of image registration will greatly affect the subsequent image processing and analysis. This paper focuses on the problem of brain image registration based on deep learning, and proposes the unsupervised deep learning methods based on model decoupling and regularization learning. Specifically, we first decompose the highly ill-conditioned inverse problem of brain image registration into two simpler sub-problems, to reduce the model complexity. Further, two light neural networks are constructed to approximate the solution of the two sub-problems and the training strategy of alternating iteration is used to solve the problem. The performance of algorithms utilizing model decoupling is evaluated through experiments conducted on brain MRI images from the LPBA40 dataset. The obtained experimental results demonstrate the superiority of the proposed algorithm over conventional learning methods in the context of brain image registration tasks.

KEYWORDS

unsupervised learning, data-adaptive, brain image registration, model decoupling, sub-problems

1. Introduction

Medical image registration is a vital step in the healthcare field, pivotal for diagnosing (Song et al., 2021), and planning treatments (Tan et al., 2016). It aligns multiple images, establishes spatial correlations, and assimilates varied data, thereby contributing to improved diagnostic precision and personalized treatments.

The task of image registration (Hu et al., 2018), involves identifying the optimal spatial transformation between two images, thereby establishing a unique correspondence between points in each space that are associated with the same anatomical position. This task is a high-dimensional, ill-posed optimization problem, commonly solved using a specific objective function:

$$T^{*} = \operatorname{argmin} D\left(I_f, T\left(I_m\right)\right), \tag{1}$$

where T^* represents the optimal transformation, I_f is the template (or fixed) image, and I_m is the image to be registered (or moving image). The function $D(\cdot, \cdot)$ quantifies the dissimilarity or distance between these two images.

Traditionally, medical image registration has been conducted with model-based methods. These models are typically categorized into parametric methods and global variational methods. Parametric methods approximate deformations using parameters, such as Thin-Plate Splines (TPS) (Bookstein, 1989) or B-splines (Xia and Liu, 2004), and solve an optimization problem to find optimal parameter values. Conversely, global variational methods frame the registration problem as an energy functional minimization task, often involving partial differential equations to ensure the diffeomorphism of the deformation field. Although these model-based methods offer high registration accuracy and robustness, they suffer from computational complexity and limitations in capturing complex deformations.

Recently, the rapid advancements in deep learning and the availability of extensive medical image datasets have catalyzed the emergence of learning-based registration methods. The early deep learning-based image registration models primarily utilized supervised learning methods. In this approach, output labels such as deformation vector fields or parameters are used during training to learn the mapping from input image pairs to deformation fields using neural networks. Various methods, including convolutional neural network (CNN) and fully convolutional network (FCN) (Sheikhjafari et al., 2022) architectures, have been explored to tackle single-modal or multi-modal registration tasks, rigid registration, and non-linear deformations. But these methods require a large amount of predefined ground truth deformation field labels, resulting in significant manpower costs. To overcome the limitations of supervised learning, unsupervised learning models for image registration have been developed. Rather than necessitating predefined ground truth deformation field labels, these models place reliance on the assessment of similarity between registered images and template images to guide the network learning process. Unsupervised learning models (Sideri-Lampretsa et al., 2022) have demonstrated competitive performance compared to traditional methods, surpassing them in metrics like Dice score, residual sum of squares, peak signal-to-noise ratio, and structural similarity. Despite their promising results, deep learning-based registration methods face certain challenges, including the presence of local minima during model optimization, which can impede convergence to accurate solutions.

To address these existing challenges, this paper bridges traditional model-based methods and modern learning-based deep learning methods, aiming to balance global smoothness and local data-adaptive discontinuity constraints. This combination is anticipated to enhance the accuracy and precision of brain image registration. Specifically, this paper introduces an unsupervised learning method specifically designed for medical image registration, focusing on brain images. The proposed method incorporates a regularization term to tackle the inherent complexity of the registration problem, thus splitting it into more manageable sub-problems through model decoupling techniques. These sub-problems are then addressed via deep learning networks, namely Similarity-Net and Denoiser-Net. Our main contributions include (a) the development of an innovative deep learning method: This novel method uses model decoupling to simplify the inverse problem of image registration. It accomplishes this by decomposing the problem into two less complex subproblems, (b) introduction of a deep learning algorithm based on model decoupling: This proposed algorithm addresses the highly ill-posed problem of image registration. The innovative aspect of this algorithm lies in its ability to utilize deep learning techniques to approximate the solutions to these lower complexity subproblems, and (c) The obtained experimental results demonstrate the superiority of the proposed algorithms over conventional learning methods in the context of image registration tasks.

2. Related works

2.1. Deep learning based registration methods

Supervised learning techniques in image registration utilize known deformation vector fields during training, with loss functions commonly comprising similarity and regularization terms. The creation of deformation labels can be quite challenging, prompting the use of random generation (Sun et al., 2018), or model-based generation approaches (Yang et al., 2016). While these techniques are valuable, they may encounter limitations due to the general lack of labeled data.

Unsupervised learning approaches (Liu et al., 2022), such as the VoxelMorph network (Balakrishnan et al., 2018), tackle the challenge of obtaining ground truth deformation fields by capitalizing on the similarity between registered images and template images. The VoxelMorph network incorporates the U-Net architecture (Ronneberger et al., 2015) for predicting the deformation field and the Spatial Transform Network (STN) module (Jaderberg et al., 2015) to apply the predicted deformation to the target image. This structure circumvents the need for explicit deformation labels, demonstrating the power of unsupervised learning in accurate image registration.

2.2. Regularization based methods

Diffeomorphic regularization, a widely adopted method, preserves the topological structure of images during registration (Beg et al., 2005). Approaches based on stationary velocity fields and architecture-based designs are common in this respect (Trouvé and Younes, 2005; Vercauteren et al., 2009). Recent advancements aim to predict diffeomorphic deformation fields within deep learning frameworks, with some methods, like SYMNet (Lu et al., 2019), directly outputting pairs of diffeomorphic deformation fields. These techniques aim to boost the smoothness and realism of deformation fields, thereby improving the accuracy and efficiency of registration.

Multi-scale regularization techniques, on the other hand, utilize information from multiple scales to enhance the robustness and accuracy of the process. Approaches such as multi-scale information fusion (Srivastava et al., 2022), multi-stage registration (de Vos et al., 2019; Cai et al., 2022), and coarse-to-fine registration (Zhao et al., 2020; Mok and Chung, 2022) have been developed to implement multi-scale regularization. Despite an increased demand for computational resources, these multi-scale techniques have demonstrated superior performance in various medical image registration tasks.

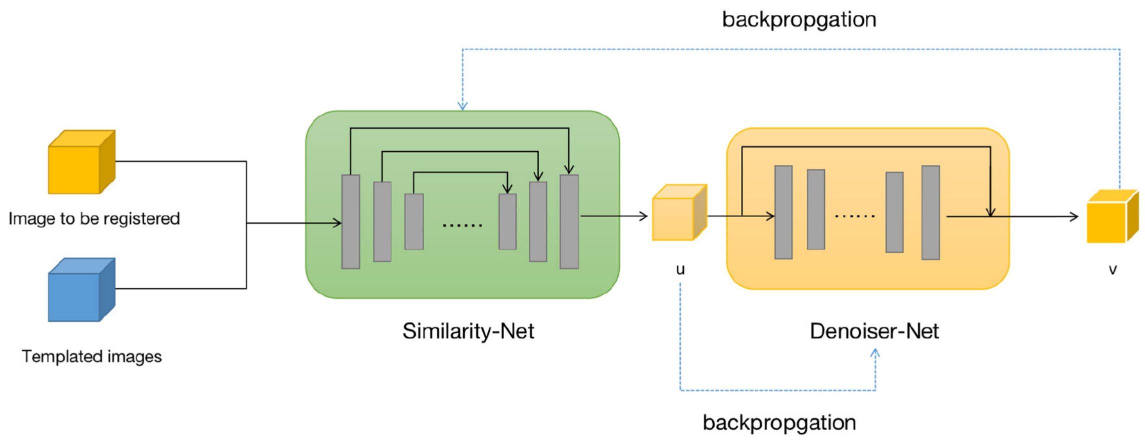


FIGURE 1
Brain image registration network structure, both *u* and *v* represents the deformation field.

3. Method

3.1. Model framework

In the context of brain magnetic resonance image registration, it is desired to maintain the topological structure of the images before and after registration. To achieve this, we consider the following optimization problem:

$$\phi^* = \arg\min \mathcal{L}_{\text{sim}} \big((I_f, I_m \circ \phi\big) + \lambda \cdot \|\nabla \phi\|_2^2 \,, \tag{2}$$

where, I_f represents the template image, I_m represents the image to be registered, ϕ denotes the predicted deformation field, and $|\nabla\phi|^2$ is the regularization term that imposes a smoothness constraint on the deformation field. The parameter λ balances the relationship between the fidelity term and the regularization term in the loss function.

Considering the complexity of image registration problems, the above optimization problem is a high-dimensional and ill-posed problem. Therefore, we propose an optimization method based on model decoupling. By introducing relaxation variables, the above optimization problem is transformed into two sub-problems:

$$u^* = \arg\min \mathcal{L}_{\text{sim}} \big((I_f, I_m \circ u) + \alpha \cdot |u - v|_2^2\big) \,, \tag{3}$$
$$v^* = \arg\min |v - u|^2 + \beta \cdot |\nabla v|_2^2 , \tag{4}$$

where, v is the relaxation variable, both u and v represents the deformation field in this problem and α and β are balancing parameters.

We design two neural networks to solve these two sub-problems. The first sub-problem is primarily addressed by using the Similarity-Net as the registration network, while for the nature of the second sub-problem, we design a denoising network, Denoiser-Net, to approximate the solution. By iteratively alternating between these two networks, a deformation field with smoothness properties is predicted. The model framework is illustrated in Figure 1. Detailed information will be discussed in Sections 3.2 and 3.3.

We provide the specific steps of the model decoupling-based method for solving the registration problem.

Model-decoupling-based brain image registration method.

```
Require: Image pairs $(I_f^n, I_m^n)$, parameters $\alpha, \beta > 0$,
iterations $k$, learning rate $lr$, batch size $B$
Ensure: Optimal solutions $u^*, v^*$.
  Input: $\left(I_f^n, I_m^n\right), n = 1, \cdots, N$.
  Initialization: Network parameters of Similarity-Net
  and Denoiser-Net, at this point $i = 0$.
  for $i \leq k$ do
  Randomly select a batch of data $\left(I_f^j, I_m^j\right), j = 1, \cdots, B$.
  Fix the network parameters of the Denoiser-Net,
  calculate $u$ and $v$.
  Compute loss (3), update Similarity-Net via
  backpropagation.
  Fix the network parameters of the Similarity-Net,
  calculate $u$ and $v$.
  Compute loss (4), update Denoiser-Net via
  backpropagation.
  $i = i + 1$.
  end for
  Output: $u^* = u, v^* = v$.
```

3.2. Similarity-Net

For the first sub-problem, we employ a similar optimization method as VoxelMorph, using a network called Similarity-Net. We adopt a network architecture similar to UNet, but with reduced network parameters and model complexity. In the encoding part, instead of performing a convolution operation with a stride of 1 after downsampling the image size, we introduce a convolution operation with a stride of 2. Additionally, the number of channels in the feature maps is reduced. In the decoding part, we restore the image size gradually using direct interpolation instead of using transposed convolution, aiming to reduce network parameters. In the encoding part of the network, we perform four convolution operations with a stride of 2 and save the corresponding feature maps. In the decoding part, we restore the image size using

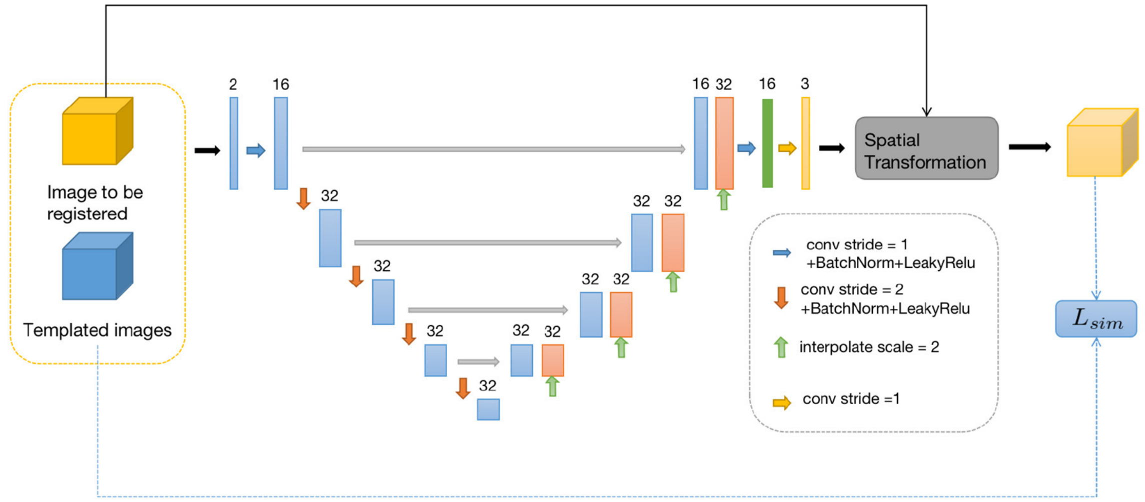


FIGURE 2
Similarity-Net network framework.

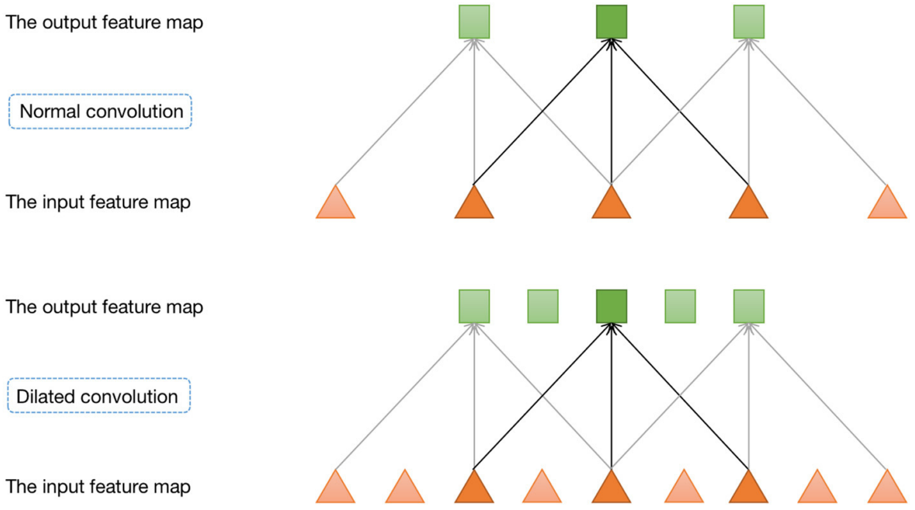


FIGURE 3
One-dimensional dilated convolution operation.

nearest-neighbor interpolation, and before each interpolation step, we connect the feature maps saved in the encoding part at the corresponding scale. Finally, after two convolution operations, the predicted deformation field is obtained.

Once the predicted deformation field is obtained, we not only use the spatial transformation layer to register the moving image but also evaluate the distance between the deformed moving image and the template image using local cross-correlation. The deformation field is then fed into the Denoiser-Net network to adjust the deformation field to satisfy the corresponding regularization constraints. The difference between the input and output of the Denoiser-Net is computed as the loss function, which guides the parameter updates of the Similarity-Net. The specific network structure is shown in Figure 2.

3.3. Denoiser-Net

The second sub-problem aims to obtain an output that is similar to the input but possesses certain desired properties. This is a common task in image denoising. To address this, we design a small denoising network called Denoiser-Net to solve the second sub-problem. Inspired by DnCNN (Huang et al., 2021) and ResNet (Zhang et al., 2017), we adopt a residual learning approach, where instead of directly mapping the input to the output, we learn the residual between the output and the input. In this design, the relationship between u and v can be expressed as:

$$v = u + \text{Residual}(u). \tag{5}$$

Furthermore, we incorporate a pyramid structure inspired by

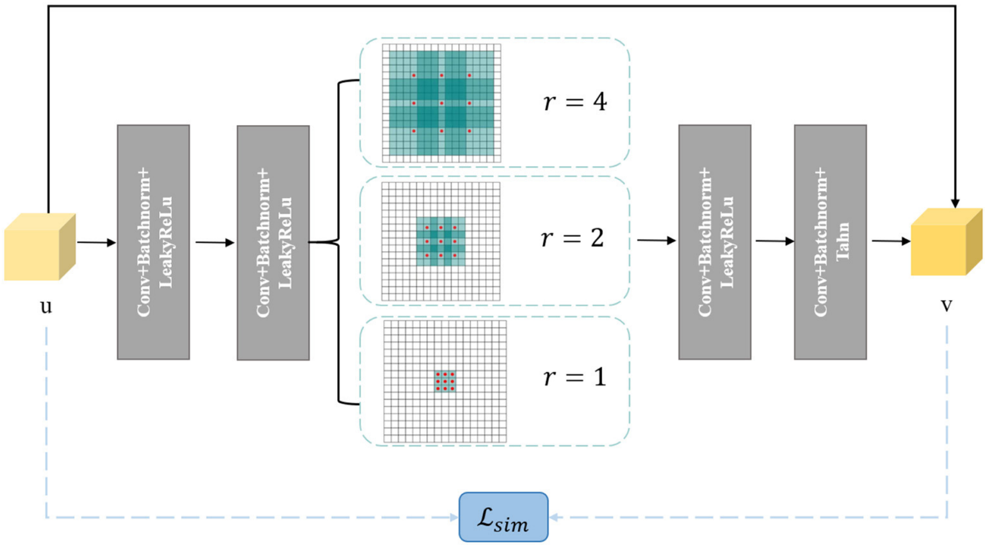


FIGURE 4
Denoiser-Net network framework.

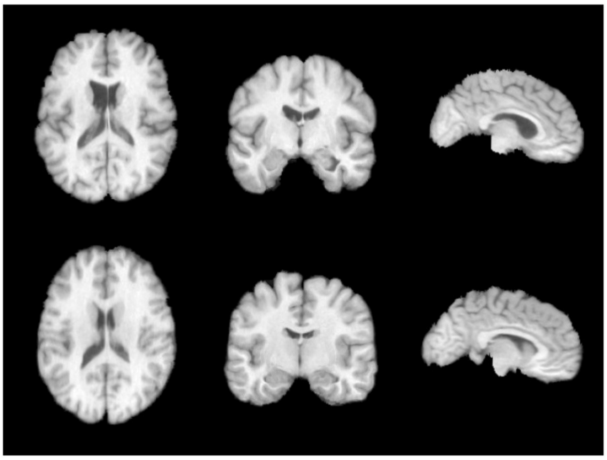


FIGURE 5
LPBA40 dataset.

TABLE 1 DSC of different methods.

Method	S-Net	SS-Net	Ours	VoxelMorph	Ours$^+$
DSC	0.6780	0.7027	0.7043	0.7053	0.7061

SPPNet (He et al., 2016) into the network construction, utilizing parallel dilated convolution operations with multiple dilation rates to achieve multi-scale information fusion. Dilated convolution, also known as atrous convolution, enables explicit control over the resolution of the computed feature maps in convolutional neural networks and allows adjustment of the filter's receptive field to capture multi-scale feature information. It is a generalization of conventional convolution operations. In the case of 1D signals, the dilated convolution applied to the input feature map x with the output feature map y and convolution filter w can be expressed as:

$$y[i] = \sum_{k} x[i + r \cdot k]w[k], \tag{6}$$

where $y[i]$ represents the value at the i-th coordinate position of the output feature map y, r denotes the dilation rate, and k represents the k-th position of the filter. Figure 3 provides a visualization of dilated convolution in 1D signals. In conclusion, the specific structure of Denoiser-Net is illustrated in Figure 4.

4. Experiments

4.1. Data preparation

The brain image dataset used in this study is the publicly available LPBA40 dataset. The LPBA40 dataset was collected at the North Shore Long Island Jewish Health System (NSLIJHS) and is maintained at the University of California, Los Angeles (UCLA). The dataset consists of 40 brain magnetic resonance imaging (MRI) scans from volunteers, with voxel sizes of 0.86 × 0.86 × 1.5 mm^3. The volunteers include 20 males and 20 females, all free of any brain disorders, psychiatric history, or intellectual developmental delay. The average age of the volunteers is 29.20 ± 6.30 years, with the youngest volunteer being 19.3 years old and the oldest being 39.5 years old. The UCLA Laboratory of Neuro Imaging (LONI) manually labeled 56 brain regions for each image in the LPBA40 dataset. The specific definitions of the brain regions can be found in Zhang and Ghanem (2018). We performed a series of standardization processes on the brain MRI images. Firstly, we used the FreeSurfer software (Shattuck et al., 2008) for skull stripping and resampled the images to a voxel size of 1 × 1 × 1 mm^3. To avoid computational

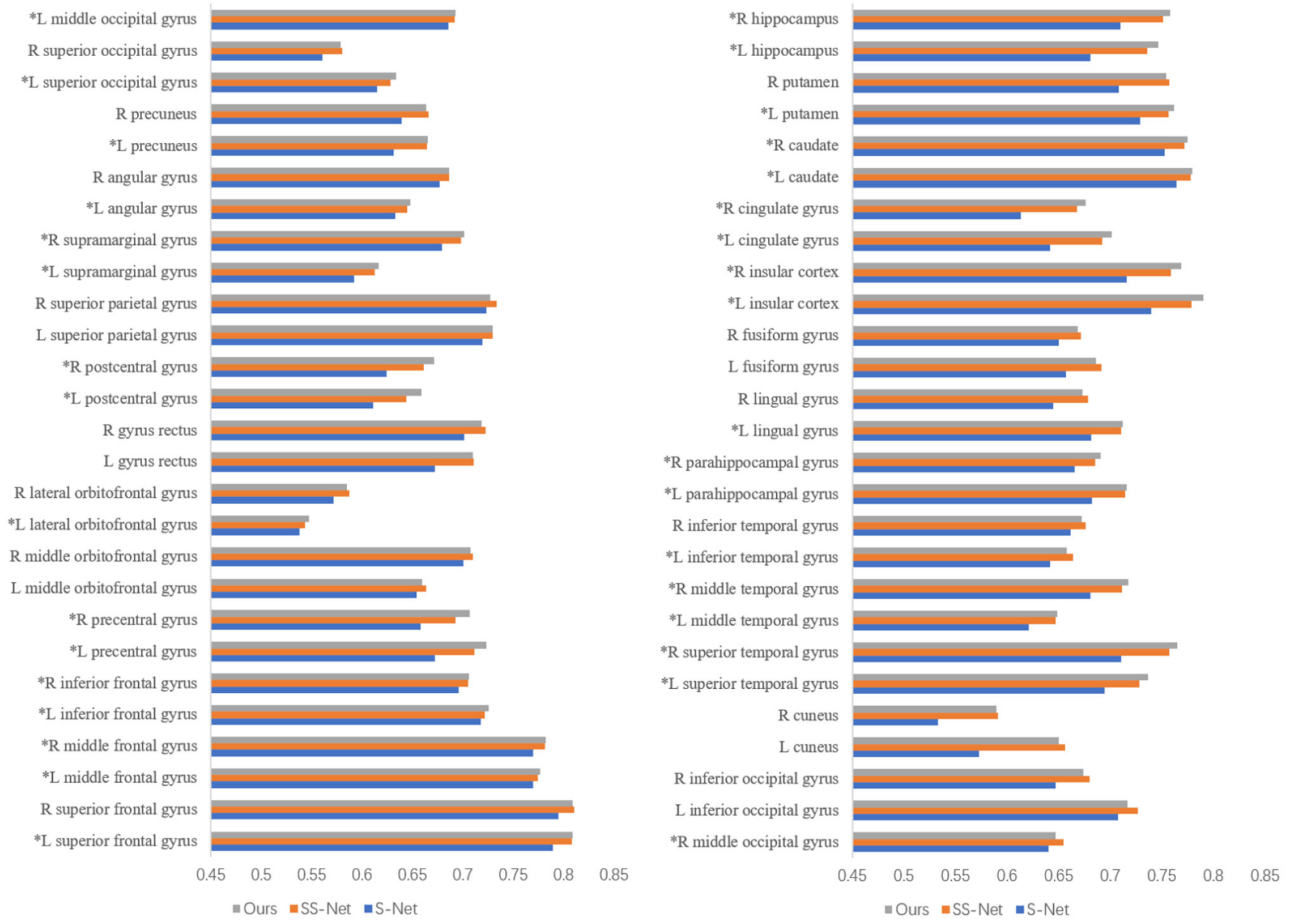


FIGURE 6
The DSC for different methods in the regions of interest.

redundancy caused by blank regions in the images, we cropped the images to a size of 144 × 192 × 160 mm^3. To eliminate the impact of grayscale value magnitude and distribution on the experiments, we normalized and histogram-equalized the cropped images. Finally, we applied affine alignment to all the images to ensure the center of study in the non-linear transformations across the brain images. Illustrations of the preprocessed images in three directions on the same slice are shown in Figure 5.

4.2. Experimental setup

The experiments were conducted on a Linux operating system, specifically Ubuntu 18.04. The network was built using the PyTorch deep learning framework. The training and testing were performed on an NVIDIA GeForce RTX 3090 GPU with 24GB of memory. To demonstrate the effectiveness of our proposed model-decoupled method on brain data, we compared it with the following methods:

(1) Similarity-Net: The network architecture is Similarity-Net without the regularization term in the loss function and without the inclusion of the denoiser network, which serves as our baseline method. For convenience, we refer to this method as S-Net.
(2) Similarity-Net with Smoothness Regularization (SS-Net): The network architecture is Similarity-Net, and the loss function includes smoothness regularization constraints but does not include the denoiser network.
(3) VoxelMorph: The network architecture is U-Net, which has more parameters than Similarity-Net, and the loss function includes smoothness regularization constraints.

4.3. Evaluation metrics

In this study, we used the Dice similarity coefficient (DSC) as a commonly used evaluation metric for quantitatively analyzing the registration performance in brain image registration. The DSC is defined as follows:

$$\mathrm{DSC}(A, B) = 2\frac{|A \cap B|}{|A| + |B|}, \tag{7}$$

where $\mathrm{DSC}(A, B)$ represents the degree of overlap between two corresponding brain regions A and B, where A and B denote the

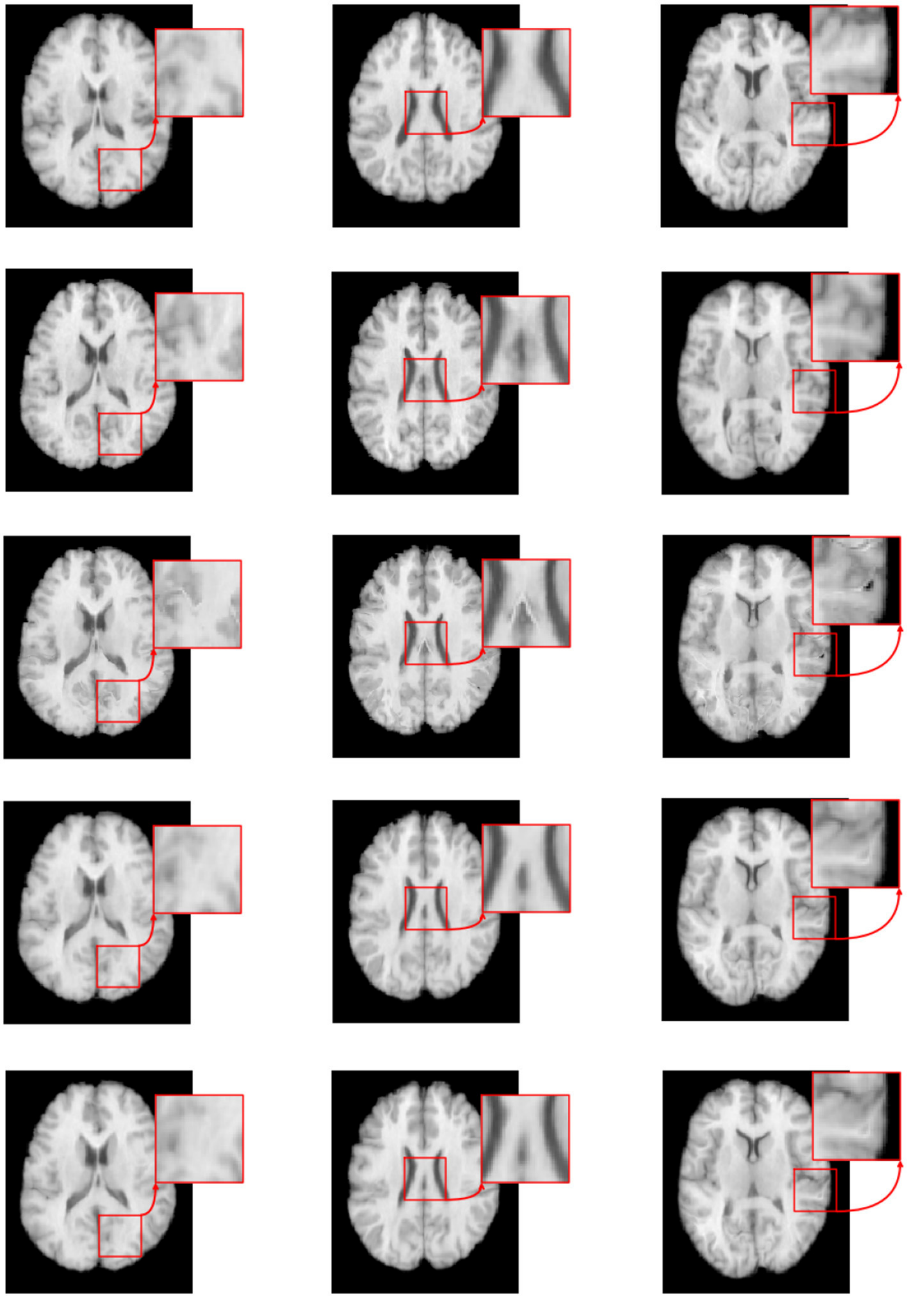


FIGURE 7
The registered images obtained through different methods.

brain regions of the template image and the registered image, respectively. The DSC value ranges from 0 to 1, with a higher value indicating a higher degree of overlap and similarity between the two brain structures.

4.4. Experimental results

For the experiment, 30 randomly selected images were used as the training set, 2 images as the validation set, and 8 images as the test set for inter-subject brain image registration. This resulted in a total of 870 image pairs available for training. The network was trained with a learning rate of 0.0005, 50,000 iterations, and a batch size of 1.

Table 1 records the Dice Similarity Coefficient (DSC) obtained under different methods. Here, "Ours+" refers to our proposed method, where we replaced the sub-network in the first step with VoxelMorph instead of Similarity-Net and performed alternating iterations with Denoiser-Net. Observing the table, we can draw the following two conclusions: (1) Compared to the method SNet, which only uses Similarity-Net, our proposed method shows a

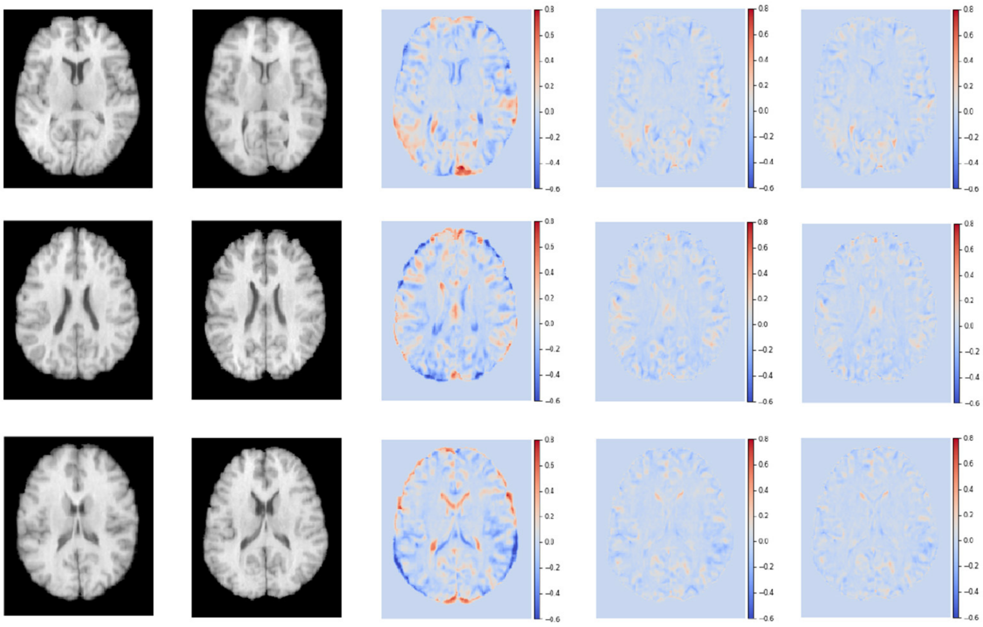


FIGURE 8
The registered images obtained through different methods.

significant improvement in the DSC metric. This indicates that our proposed method effectively imposes regularization constraints on the deformation field, thereby enhancing the registration accuracy. (2) Compared to the method SS-Net, which directly incorporates regularization terms into the loss function, our proposed method also exhibits a slight improvement in the DSC metric. Furthermore, even after replacing Similarity-Net with VoxelMorph, our proposed method still outperforms VoxelMorph, suggesting that our model-based method can further narrow the solution space and reduce the occurrence of local minima to a certain extent.

Figure 6 presents the DSC (Dice Similarity Coefficient) metrics for S-Net, SS-Net, and our proposed method across 54 regions of interest (ROIs) of interest. The parts marked with asterisks (*) indicate that our method achieved higher DSC values in those brain regions compared to the other two methods. Upon statistical analysis, our proposed method demonstrated superior registration performance in 33 brain regions. This suggests that the improvement in the DSC metric achieved by our method is not limited to specific brain regions but rather reflects an overall enhancement in registration accuracy.

Figure 7 illustrates the visual results of S-Net, SS-Net, and our proposed method on the LPBA40 dataset. The three columns represent the visualization results for three slices. The top row shows the target (moving) image, the middle row displays the template (fixed) image, the third row depicts the image registered using the S-Net method, the fourth row shows the image registered using the SS-Net method, and the fifth row displays the image registered using our proposed method. By observing the results, it is evident that the image registered using the S-Net method exhibits local discontinuities, connections, and holes that are inconsistent with the actual data. On the other hand, the images registered using the SS-Net method and our proposed method appear smoother and closer to the real data.

Figure 8 shows the residual maps of S-Net, SS-Net, and our proposed method on the LPBA40 dataset. The three rows represent the visualization results of three slices. The first column corresponds to the target image, the second column is the template image, the third row shows the difference between the two images without registration, the fourth row shows the difference between the image registered using the SS-Net method and the template image, and the fifth row shows the difference between the image registered using our proposed method and the template image. By observation, our proposed method reduces the differences between the registered floating image and the template image, and in some regions, it performs similarly to or slightly better than SS-Net.

In conclusion, our proposed method outperforms S-Net in terms of evaluation metrics and visual effects, and slightly outperforms SS-Net. This demonstrates that the method based on model decoupling and alternate iterative training strategy effectively learns the smoothness regularization constraint, thereby improving registration accuracy. Furthermore, in the experiments with increased model complexity, i.e., the improved model based on the VoxelMorph framework proposed by us still achieves a certain degree of improvement in performance. This indicates that our method can serve as a framework to be combined with other more sophisticated networks, enhancing registration accuracy on top of the existing network.

5. Conclusion

In our study, we propose a novel deep learning method that employs model decoupling to augment the precision of registration tasks in medical imaging. By constructing separate networks for fidelity and regularization terms, we achieve effective constraint of the solution space, thereby reducing the occurrence of local minima that might compromise result quality. Our method's superior performance was demonstrated through its application to image registration tasks on brain magnetic resonance imaging (MRI), enhancing the accuracy of image processing and analysis.

Although our research has made considerable strides in the domain of image registration, there remain potential areas for future exploration. One such aspect pertains to the performance of the two subnetworks within our model. Given the dependency of our unsupervised learning method's registration accuracy on the first network's output, investigating the integration of potentially more efficient network architectures into our framework could be beneficial. This could pave the way for elevated overall registration accuracy.

In terms of regularization, while our work leverages the common differential diffeomorphic regularization for brain MRI datasets, alternative regularization constraints could be explored to further refine the results. This offers another promising avenue for more comprehensive research in the future. By delving into these areas, we anticipate building on our existing contributions and facilitating further advancements in the field of brain image registration through deep learning.

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/s.

Author contributions

JF spearheaded the project's conceptualization, methodology, and manuscript drafting. NL majorly handled data analysis and interpretation and also assisting in manuscript writing. JL contributed to experimental design and implementation. HZ assisted with data interpretation, manuscript writing, and providing critical feedback. JWe and WY undertook experiments and data acquisition and aided in manuscript revision. JWu and ZW led project management, influenced experimental design, data interpretation, and manuscript writing. All authors approved the final version of the article.

Funding

This study was supported by National Natural Science Foundation of China (No. 11971296) and Shanghai Municipal Science and Technology Major Project.

Acknowledgments

We wish to express our sincere gratitude to the reviewers who contributed their valuable time and expertise to the evaluation of this manuscript. Their insightful comments and suggestions greatly improved the quality of our work. We would also like to extend special thanks to all those who provided critical feedback, contributing to the refinement of our research and manuscript.

Conflict of interest

JF is employed by Industrial Internet Innovation Center (Shanghai) Co., Ltd., Shanghai, China.

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

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

Balakrishnan, G., Zhao, A., Sabuncu, M. R., and Dalca, A. V. (2018). "An unsupervised learning model for deformable medical image registration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (New York, NY), 9252–9260.

Beg, M., Miller, M., Trouve, A., and Younes, L. (2005). Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *Int. J. Comput. Vision* 61, 139–157. doi: 10.1023/B:VISI.0000043755.93987.aa

Bookstein, F. (1989). Principal warps: thin-plate splines and the decomposition of deformations. *IEEE Trans. Pattern Anal. Mach. Intell.* 11, 567–585.

Cai, Y., Lin, J., Lin, Z., Wang, H., Zhang, Y., Pfister, H., et al. (2022). "MST++: multi-stage spectral-wise transformer for efficient spectral reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* (New York, NY), 745–755.

de Vos, B. D., Berendsen, F. F., Viergever, M. A., Sokooti, H., Staring, M., and Iogum, I. (2019). A deep learning framework for unsupervised affine and deformable image registration. *Med. Image Anal.* 52, 128–143. doi: 10.1016/j.media.2018.11.010

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* (New York, NY), 770–778.

Hu, Y., Modat, M., Gibson, E., Ghavami, N., Bonmati, E., Moore, C. M., et al. (2018). "Label-driven weakly-supervised learning for multimodal deformable image registration," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* (New York, NY), 1070–1074.

Huang, Y., Ahmad, S., Fan, J., Shen, D., and Yap, P.-T. (2021). Difficulty-aware hierarchical convolutional neural networks for deformable registration of brain MR images. *Med. Image Anal.* 67, 101817. doi: 10.1016/j.media.2020.101817

Jaderberg, M., Simonyan, K., Zisserman, A., and Kavukcuoglu, K. (2015). "Spatial transformer networks," in *Advances in Neural Information Processing Systems 28*, eds C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, R. Garnett (La Jolla, CA: Neural Information Processing Systems (NIPS)), 2017–2015.

Liu, Y., Zheng, Y., Zhang, D., Chen, H., Peng, H., and Pan, S. (2022). "Towards unsupervised deep graph structure learning," in *Proceedings of the ACM Web Conference* (New York, NY: Association for Computing Machinery), 1392–1403.

Lu, Z., Yang, G., Hua, T., Hu, L., Kong, Y., Tang, L., et al. (2019). "Unsupervised three-dimensional image registration using a cycle convolutional neural network," in *2019 IEEE International Conference on Image Processing (ICIP)* (New York, NY), 2174–2178.

Mok, T. C. W., and Chung, A. C. S. (2022). "Affine medical image registration with coarse-to-fine vision transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Los Alamitos, CA), 20835–20844.

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-Net: convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference* (Munich: Springer), 234–241.

Shattuck, D. W., Mirza, M., Adisetiyo, V., Hojatkashani, C., Salamon, G., Narr, K. L., et al. (2008). Construction of a 3d probabilistic atlas of human cortical structures. *Neuroimage* 39, 1064–1080. doi: 10.1016/j.neuroimage.2007.09.031

Sheikhjafari, A., Noga, M., Punithakumar, K., and Ray, N. (2022). "Unsupervised deformable image registration with fully connected generative neural network," in *Medical Imaging with Deep Learning*.

Sideri-Lampretsa, V., Kaissis, G., and Rueckert, D. (2022). "Multi-modal unsupervised brain image registration using edge maps," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)* (New York, NY), 1–5.

Song, J., Zheng, J., Li, P., Lu, X., Zhu, G., and Shen, P. (2021). An effective multimodal image fusion method using MRI and PET for Alzheimer's disease diagnosis. *Front. Digit. Health* 3, 637386. doi: 10.3389/fdgth.2021.637386

Srivastava, A., Jha, D., Chanda, S., Pal, U., Johansen, H. D., Johansen, D., et al. (2022). MSRF-Net: a multi-scale residual fusion network for biomedical image segmentation. *IEEE J. Biomed. Health Inform.* 26, 2252–2263. doi: 10.1109/JBHI.2021.3138024

Sun, Y., Moelker, A., Niessen, W. J., and van Walsum, T. (2018). "Towards robust CT-ultrasound registration using deep learning methods," in *Understanding and Interpreting Machine Learning in Medical Image Computing Applications: First International Workshops, MLCN 2018, DLF 2018, and iMIMIC 2018, Held in Conjunction with MICCAI 2018* (Granada: Springer), 43–51.

Tan, M., Li, Z., Qiu, Y., McMeekin, S. D., Thai, T. C., Ding, K., et al. (2016). A new approach to evaluate drug treatment response of ovarian cancer patients based on deformable image registration. *IEEE Trans. Med. Imaging* 35, 316–325. doi: 10.1109/TMI.2015.2473823

Trouvé, A., and Younes, L. (2005). Metamorphoses through lie group action. *Found. Comput. Math.* 5, 173–198. doi: 10.1007/s10208-004-0128-z

Vercauteren, T., Pennec, X., Perchant, A., and Ayache, N. (2009). Diffeomorphic demons: efficient non-parametric image registration. *Neuroimage* 45, S61–S72. doi: 10.1016/j.neuroimage.2008.10.040

Xia, M., and Liu, B. (2004). Image registration by "super-curves". *IEEE Trans. Image Process.* 13, 720–732.

Yang, X., Kwitt, R., and Niethammer, M. (2016). "Fast predictive image registration," in *Deep Learning and Data Labeling for Medical Applications: First International Workshop, LABELS 2016, and Second International Workshop, DLMIA 2016, Held in Conjunction with MICCAI 2016* (Athens: Springer), 48–57.

Zhang, J., and Ghanem, B. (2018). "ISTA-Net: interpretable optimization-inspired deep network for image compressive sensing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (New York, NY), 1828–1837.

Zhang, K., Zuo, W., Chen, Y., Meng, D., and Zhang, L. (2017). Beyond a gaussian denoiser: residual learning of deep CNN for image denoising. *IEEE Trans. Image Process.* 26, 3142–3155. doi: 10.1109/TIP.2017.2662206

Zhao, S., Lau, T., Luo, J., Chang, E. I.-C., and Xu, Y. (2020). Unsupervised 3D end-to-end medical image registration with volume tweening network. *IEEE J. Biomed. Health Inform.* 24, 1394–1404. doi: 10.1109/JBHI.2019.2951024

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 05 September 2023
DOI 10.3389/fnins.2023.1247227

OPEN ACCESS

EDITED BY
Shihui Ying,
Shanghai University, China

REVIEWED BY
Pengcheng Liu,
University of York, United Kingdom
Leslie Samuel Smith,
University of Stirling, United Kingdom

*CORRESPONDENCE
Jigen Peng
✉ jgpeng@gzhu.edu.cn
Haiyang Li
✉ fplihaiyang@126.com

RECEIVED 25 June 2023
ACCEPTED 21 August 2023
PUBLISHED 05 September 2023

CITATION
Zheng Y, Wang Y, Wu G, Li H and Peng J (2023) Enhancing LGMD-based model for collision prediction via binocular structure.
Front. Neurosci. 17:1247227.
doi: 10.3389/fnins.2023.1247227

Enhancing LGMD-based model for collision prediction via binocular structure

Yi Zheng[1,2], Yusi Wang[1,2], Guangrong Wu[1,2], Haiyang Li[1,2]* and Jigen Peng[1,2]*

[1]School of Mathematics and Information Science, Guangzhou University, Guangzhou, China, [2]Machine Life and Intelligence Research Center, Guangzhou University, Guangzhou, China

Introduction: Lobular giant motion detector (LGMD) neurons, renowned for their distinctive response to looming stimuli, inspire the development of visual neural network models for collision prediction. However, the existing LGMD-based models could not yet incorporate the invaluable feature of depth distance and still suffer from the following two primary drawbacks. Firstly, they struggle to effectively distinguish the three fundamental motion patterns of approaching, receding, and translating, in contrast to the natural abilities of LGMD neurons. Secondly, due to their reliance on a general determination process employing an activation function and fixed threshold for output, these models exhibit dramatic fluctuations in prediction effectiveness across different scenarios.

Methods: To address these issues, we propose a novel LGMD-based model with a binocular structure (Bi-LGMD). The depth distance of the moving object is extracted by calculating the binocular disparity facilitating a clear differentiation of the motion patterns, after obtaining the moving object's contour through the basic components of the LGMD network. In addition, we introduce a self-adaptive warning depth-distance, enhancing the model's robustness in various motion scenarios.

Results: The effectiveness of the proposed model is verified using computer-simulated and real-world videos.

Discussion: Furthermore, the experimental results demonstrate that the proposed model is robust to contrast and noise.

KEYWORDS

collision prediction, lobula giant movement detectors (LGMDs), binocular vision, disparity, depth distance

1. Introduction

In the real world, collisions often lead to some kind of danger and unexpected loss. Therefore, many modern artificial machines, such as ground vehicles and unmanned aerial vehicles (UAVs), should be equipped with the intellectual abilities of collision prediction. Current methods for collision prediction, such as laser, infrared, radar, and ultrasonic, are not very suitable for daily civilian machines because of the disadvantages of high price, large size, high power consumption, and so on. Meanwhile, vision-based sensors, with the characteristics of economy and energy saving, have gradually become one of the most mainstream methods of sensing collision in the past decades. However, in terms of effectiveness and robustness, it still needs to be further improved (Mukhtar et al., 2015).

As we know, in nature, many insects show excellent collision prediction and collision avoidance abilities based on visual information, which benefits from their millions of years of evolution (Eichler et al., 2017). Despite their minuscule and simple brains, these lowly creatures seem to hold the key to solving some of mankind's greatest problems (Franceschini, 2014; Xu et al., 2023), and bring us some inspirations to build a collision prediction neural network based on visual information (Serres and Ruffier, 2017; Fu et al., 2018a). Among these insects, locusts are the most representative. When locust plague breaks out, millions of locusts can travel hundreds of miles together free of collision (Kennedy, 1951). Researchers observe that when a collision is imminent, locusts can respond quickly and change their flight direction in a very short time (hundreds of milliseconds; Fu et al., 2019c). How do locusts achieve it?

Lobula giant movement detector (LGMD), which is a huge single neuron located on the third visual neuropile of the lobule, was found by O'Shea and Williams (1974). LGMD neuron responds vigorously to approaching objects while producing little or no response to receding ones (O'shea and Rowell, 1976; Sztarker and Rind, 2014; Wernitznig et al., 2015; Rind et al., 2016). Further, researchers conduct a lot of experiments and explorations around the reflection properties of LGMD neuron (Gabbiani and Krapp, 2006; Dewell and Gabbiani, 2018, 2019; Zhu et al., 2018), and the results show that the LGMD neuron is an ideal model for constructing collision prediction visual neural network.

Based on these biological experiments, Rind and Bramwell (1996) proposed an LGMD-based neural network model. The model is composed of four groups of cells—photoreceptor cells (*P* cells), excitatory cells (*E* cells), inhibitory cells (*I* cells), and summing cells (*S* cells), as well as two single cells—feed-forward inhibition and LGMD. Since then, Yue and Rind (2006) introduced an extra artificial layer (*G* layer) to extract the extended edge of the approaching object by enhancing the cluster output, which improved the model's performance and achieved ideal results in real-world scenarios.

Following the above two studies, a large number of LGMD-based visual neural network models have sprung up. For example, based on ON/OFF channels (Fu, 2023), Fu et al. (2019b) realized the special selectivity to darker looming objects in brighter background in the model, which simulated the response of LGMD2 neurons in the infancy of locusts. Inspired by the visual pathway of Drosophila, Li et al. (2022) added a contrast channel to the LGMD-based model, which improved the stability of the model under different contrasts. Luan et al. (2021, 2022) used a similar network model to build a visual neural network with the ability to encode spatial position information, and successfully simulated MLG1 neurons in crabs. Zhao et al. (2018, 2019, 2021) further optimized the original model by designing the temporal and spatial distribution in the model according to the latest discovery of locust anatomical synaptic connection, which was successfully applied to UAV agile flight. Some models are also be tested in ground vehicle scenarios (Hartbauer, 2017; Fu et al., 2019a), mobile robots (Hu et al., 2016; Čížek et al., 2017), and recently in UAVs (Poiesi and Cavallaro, 2016; Salt et al., 2017, 2019) and micro robots (Fu et al., 2020, 2021). In addition, it is also embodied in hardware implementation, such as the FPGA (Meng et al., 2010).

However, the current models lack the consideration of the depth distance of moving objects, which is certainly a highly valuable feature for collision prediction tasks. This absence of depth distance information in the existing models results in several shortcomings. First, existing models are not able to distinguish well between the three fundamental motion modes of approaching, receding and translating, resulting in their inability to consistently demonstrate a preference for approaching objects. Secondly, the response result of the models is heavily influenced by activation function parameters and corresponding given hard thresholds. Thirdly, the models are sensitive to various input image stream factors, including noise and contrast. While some models enhance certain aspects by designing artificial mechanisms, extracting the core feature of depth distance holds the potential to effectively address all of these issues simultaneously.

For that, a novel LGMD-based neural network model with binocular vision is proposed in this paper, named Bi-LGMD. This model requires two image stream inputs, coming from the left and right eye, respectively. For both inputs, a basic LGMD-based model is used to extract the contours of the moving object. Then, based on the principle of binocular stereo vision, the obtained contour information is used to compute the disparity of the moving object, and the moving object's depth distance at each time step is further estimated. Based on this, motion patterns can be effectively distinguished. Moreover, different from existing models, the activation function is not required in our model. Instead, the concept of warning depth-distance is introduced. Depending on the change of the estimated depth distance at each time, the warning depth-distance is dynamically and adaptively adjusted through a specific computational rule. The LGMD neuron is activated only when an approaching object reaches the warning depth-distance. Hence, the parameter setting problem for the activation function is avoided. More importantly, the model is more robust to input image streams. On the one hand, this is due to the consideration of more essential kinematic features of depth-distance. On the other hand, the computational process of disparity is mainly based on the matching of two contours from the left and right channel, rather than the pixel value itself, so the factors that seriously affect the pixel value of an image (such as noise, contrast, etc.) have a great impact on existing models, but the computational result of disparity is relatively stable.

The main innovations of this paper can be summarized as follows:

1. This paper proposes a novel LGMD-based model with binocular structure, and the essential feature of depth distance is introduced into the model for the first time. As a result, the proposed model is able to clearly distinguish motion modes such as approaching, receding and translating, with improved selectivity.
2. We design a dynamic adaptive warning depth distance related to the approaching velocity. On the one hand, the model could be adapted to more complex approaching modes. On the other hand, the model does not rely on the activation function parameters and a given hard threshold, alleviating the extreme sensitivity of the existing models to activation parameters.
3. Unlike existing models that heavily rely on the pixel values of G layer outputs, the proposed model ultimately focuses on matching the overall left and right outputs. Based on this novel perspective, the proposed model has stronger robustness to factors such as noise and contrast in the input image streams.

The rest of this paper is organized as follows. Section 2 introduces some related work, including motion pattern recognition in the model and the advantages of incorporating stereo vision. Section 3 describes the proposed Bi-LGMD visual neural network. Systematic experiments and analyses of the model results are illustrated in Section 4. Thereafter, further discussions are given in Section 5. Section 6 concludes the paper.

2. Related work

2.1. Motion pattern recognition

LGMD neuron is viewed as an ideal paradigm for constructing collision prediction models. Numerous LGMD-based models are validated to indeed respond significantly to looming stimuli, yet it is difficult to be completely unresponsive to other motion patterns. Therefore, further improvements are still needed to clearly distinguish between the basic motion patterns including approaching, receding and translating.

In the past, some models attempted further improvements in terms of the selective response of the model to motion patterns, for example, Lei et al. (2022) improved the LGMD-based model using the ON-OFF competition mechanism, enabling it to distinguish a looming object from a near and fast translatory moving object. However, it does not explore the response to receding stimuli, and the competition mechanism does not seem to be effective in distinguishing between approaching and receding. Fu et al. (2018b) designed a spike frequency adaptation (SFA) mechanism to enhance the collision selectivity to approaching objects, however, the model still has a brief and small response to the receding and translating stimuli, which may cause false alarms in situations where the model parameters are inappropriate (especially the activation parameter and spiking threshold).

In general, while some models could make partial discrimination between different motion patterns, there are still some problems, such as how to choose the spiking threshold. By contrast, the trend of depth distance is the most intuitive way to distinguish basic motion patterns. Once it is effectively estimated, the model can understand the motion patterns more "visually," knowing exactly which of the "approaching, receding, and translating" the motion pattern belongs to at the current moment. The results of the discrimination of motion modes will no longer be affected by parameters and thresholds, and its discrimination method is obviously simple, robust, and interpretable.

2.2. Binocular structure and stereo vision

Binocular vision, which allows for depth perception, is crucial for arthropods to interact with their environment. This is particularly important for behaviors such as motion navigation, prey capture, and attack avoidance (Nityananda et al., 2016a; Scarano et al., 2018). The binocular structure of arthropods is capable of processing information from both eyes to estimate depth and distance in the visual scene through a concept known as "disparity" (Parker, 2007; Nityananda et al., 2016b). Recent research on arthropods, like crabs, has shown a strong binocular coupling between their eyes indicating the use of binocular depth vision in capturing prey (Horridge and Sandeman, 1964; Scarano et al., 2018). Praying mantises, for example, use their stereoscopic vision to estimate the distance to their prey. Once it is within reach, they trigger a rapid strike of their forelegs (Rossel, 1986; Rosner et al., 2019).

Although the computational mechanisms behind binocular vision in arthropods are not yet fully understood, experimental findings indicate that different types of neurons in the Lobula region of their brains compute binocular information (Rosner et al., 2020). Rosner and colleagues have provided evidence that individual neurons in the praying mantis brain can recognize specific binocular information such as disparity and eccentricity, allowing them to determine locations in three-dimensional space. They identified the existence of disparity-sensitive neurons in the insect's brain and proved their role in the development of stereo vision (Rosner et al., 2019).

Interestingly, stereoscopic vision in insects, including mantises, differs from that of humans. Insects rely on changes in luminance rather than luminance directly to perceive depth (Rosner et al., 2019), which implies that insects pay more attention to moving and changing visual information rather than static details in the background. This unique approach allows insects like praying mantises to develop an efficient stereoscopic vision system using a visual network of neurons that is significantly smaller than the human brain (Rossel, 1983; Collett, 1996).

Therefore, the introduction of binocular structures in LGMD-based neural networks to extract depth-distance information is intuitively significant for enhancing collision prediction. Indeed, there has been some research work related to binocular LGMD modeling. For example, Yue and Rind (2009) proposed a network model with two LGMD modules for near-range path navigation. In their work, the input image will be decomposed into left and right parts for the two LGMD modules, and the two outputs will be compared in terms of strength and weakness to determine which way the robot's wheels should dodge. In addition, Fu et al. (2017) also designed similar binocular structures using LGMD1 and LGMD2 to investigate how this combined strategy performs for different visual stimuli when applied to a robot. However, it appears that there are few models based on LGMD that utilize binocular structures to develop stereo vision, extract depth-distance information, and explore the advantages of incorporating such information into LGMD-based models.

3. Formulation of the model

In this section, the proposed model and the corresponding computational methods are described in detail. Here, we first introduce the overall framework of the Bi-LGMD model, and then give a more specific description in the following sections.

As shown in Figure 1, in general, the proposed model contains two parallel channels to process the input image stream from the left and right camera, respectively. Each channel consists of five layers, including photoreceptor (P), excitation (E), inhibition (I), summation (S), and grouping (G) layers. Then, the outputs of the two parallel channels will be integrated in the disparity (DP) layer,

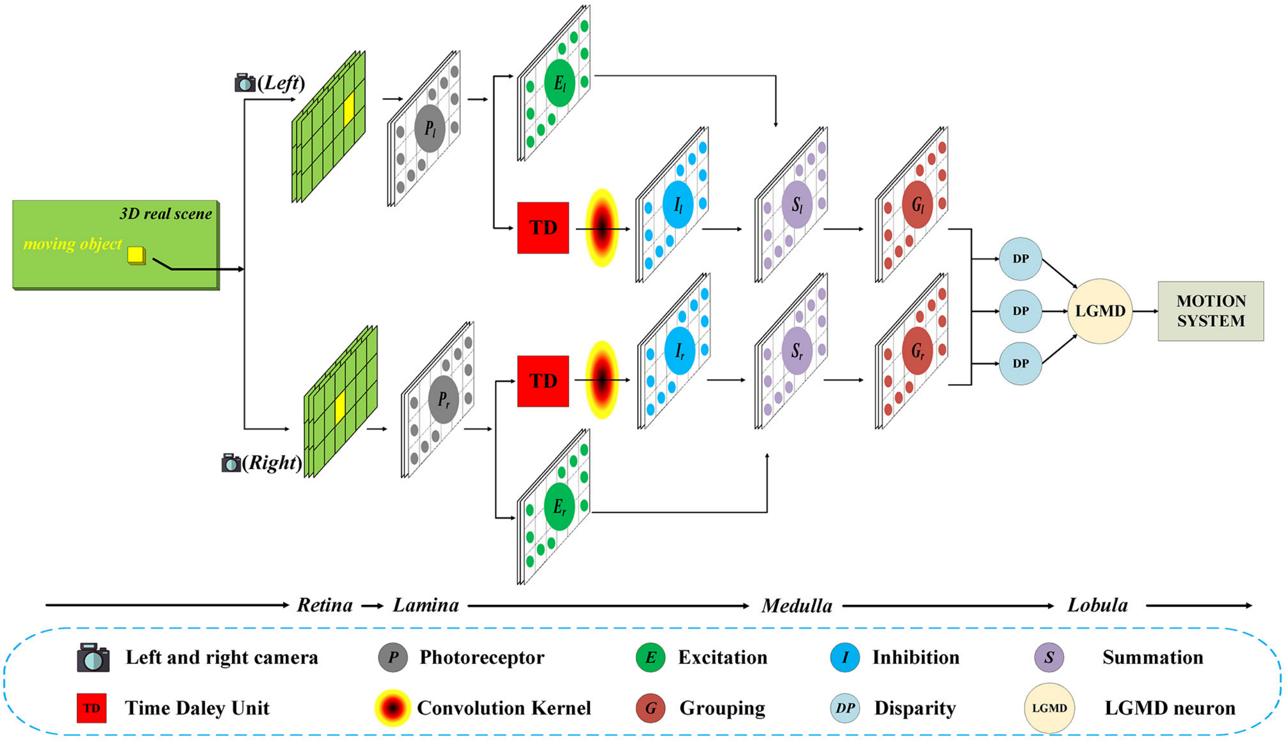


FIGURE 1
Schematic of the Bi-LGMD visual neural network. For visual stimulation in the three-dimensional world, the left and right cameras are used to shoot at the same time, and the two input image streams are processed separately in the early stage. In the last Medulla to Lobula layer, all the information is integrated through the disparity principle, so as to extract the depth distance information of the moving object. Finally, the LGMD neuron responds based on the changes in estimated depth distance.

and the information will eventually be transmitted to the LGMD layer.

In this model, inputs from both cameras are considered equally important. Therefore, the two parallel channels have exactly the same structure and the same calculation method, and the relevant parameters are set to be the same in the subsequent experiments. For convenience, in the following sections, the subscripts l and r are used to represent that the corresponding variables belong to the left and the right channel, respectively. In the following basic process, we describe the computational method in the left channel as an example, which is exactly the same as in the right channel.

3.1. Basic process

The basic process includes P, E, I, S, G layers. This classical process framework has been used in many existing models, such as Fu et al. (2019b, 2020), Luan et al. (2021), Lei et al. (2022), and Wang et al. (2023). In fact, our model does not change significantly for this part, so we will briefly review it here.

3.1.1. P layer

In this layer, the photoreceptors are arranged as a matrix. Each photoreceptor captures the grayscale luminance of the corresponding pixel in the input image stream and computes the temporal difference between the sequence frames to preliminarily extract motion information. The mathematical formula can be defined as

$$P_l(x,y,t) = L_l(x,y,t) - L_l(x,y,t-1) + \sum_{i=1}^{n_p} a_i P_l(x,y,t-i) \quad (1)$$

where $L(x,y,t)$ stands for the grayscale luminance of the pixel (x,y) at time t, and $P(x,y,t)$ represents the grayscale luminance change; n_p indicates the maximum number of frames the persistence of the luminance change could last, and a_i is a decay coefficient, which is defined by

$$a_i = (1+e^i)^{-1} \quad (2)$$

3.1.2. IE layer

The IE layer is the core of the “critical race” mentioned by Rind and Bramwell (1996). Both excitatory cells (E cells) and lateral inhibitory cells (I cells) receive the outputs of the P cells. E cells directly receive the excitation from the corresponding P cells without temporal latency, while the I cells, which pass inhibition, receive the excitation from the surrounding adjacent P cells by convolving, and there is one image frame time-delay. The mathematical formulas are defined as follows:

$$E_l(x,y,t) = P_l(x,y,t) \quad (3)$$

$$I_l(x,y,t) = \sum_{i=-1}^{1}\sum_{j=-1}^{1} P_l(x+i, y+j, t-1) w_I(i,j) \tag{4}$$

where $E(x,y,t)$ and $I(x,y,t)$ are the activity of excitatory cells and lateral inhibitory cells, respectively. w_I is the local inhibition weight that meets the following matrix, which is also used in Yue and Rind (2006), Fu et al. (2018b), Luan et al. (2021), and Li et al. (2022).

$$w_I = \begin{pmatrix} 0.125 & 0.25 & 0.125 \\ 0.25 & 0 & 0.25 \\ 0.125 & 0.25 & 0.125 \end{pmatrix}$$

3.1.3. S layer

In the S layer, the information processing results of E cells and I cells in the upper layer need to be summarized. Here, a simple linear operation is adopted (Note that inhibition has the opposite sign against excitation):

$$S_l(x,y,t) = |E_l(x,y,t)| - |I_l(x,y,t)| * W_I \tag{5}$$

where W_I is a constant which means global inhibition weight. In addition, since inhibition can reduce the activity of excitatory cells to 0 at most, it needs to be corrected here.

$$S_l(x,y,t) = [S_l(x,y,t)]^+ \tag{6}$$

where $[x]^+ = max(0,x)$.

3.1.4. G layer

To further enhance the outputs of the S layer, the G layer obtains a passing coefficient Ce through the cell's surrounding neighbors to filter out the isolated and decayed excitations, as illustrated in Figure 2. The computational formulas are as follows:

$$Ce_l(x,y,t) = \sum_{i=-1}^{1}\sum_{j=-1}^{1} S_l(x+i, y+j, t) w_e(i,j) \tag{7}$$

$$G_l(x,y,t) = S_l(x,y,t) \cdot Ce_l(x,y,t) \cdot w_l(t)^{-1} \tag{8}$$

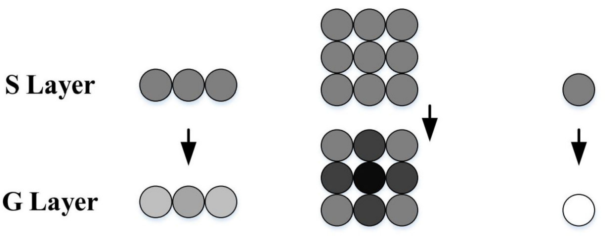


FIGURE 2
Schematic illustration of G layer processing, adapted from Yue and Rind (2006). The S cells surrounded by strong excitations obtain bigger passing coefficients, while the isolated ones gain smaller passing coefficients and may be ruled out by the threshold. The excitation strength is represented by gray levels, where the darker the color, the stronger the excitation.

$$w_e = \frac{1}{9} \times \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \tag{9}$$

$$w_l(t) = max([Ce_l]_t) \cdot C_w^{-1} + \Delta_c \tag{10}$$

where w is a scale parameter computed at every time step. C_w is a constant. $max([Ce]_t)$ stands for the largest element in matrix $[Ce]_t$. Δ_c is a small real number, which prevents the denominator from being 0 during calculation. Finally, a threshold T_{de} is introduced for the final calculation as follows.

$$\hat{G}_l(x,y,t) = \begin{cases} G_l(x,y,t), & if\ G_l(x,y,t) \geq T_{de} \\ 0, & otherwise \end{cases} \tag{11}$$

Therefore, after the processing of the G layer, the grouped excitations in the S layer representing expanding edges become stronger, while the isolated excitations caused by background details are largely filtered out.

3.2. Disparity layer (DP layer)

It is well-known that many creatures in nature have two eyes. The binocular structure can produce stereo vision, and obtain the information of depth distance through the disparity, which can not be achieved by a single eye (Ayache, 1991; Yang et al., 2017; Vienne et al., 2018). In this section, we use this principle to estimate the depth distance of moving objects at each time step. For this purpose, the information from the left and right cameras will be integrated into the DP layer.

3.2.1. Computing method of disparity

In the pictures taken by the left camera and the right camera, the imaging positions of the same object are different (see Figure 3A). More specifically, the imaging positions of closer objects are shifted considerably, while the difference is smaller for more distant objects. As shown in Figure 3B, this visual difference is called "disparity" (Ayache, 1991; Ding et al., 2021).

However, how to calculate the disparity in our model? Since the G layer mainly extracts the edge of the moving object, $\hat{G}_l$ and $\hat{G}_r$ can be used to obtain the disparity of the moving object. In the sense of the disparity described above, it can be computed by the following mathematical formula:

$$DP(t) = \arg\max_d \sum_{x=1}^{R}\sum_{y=1}^{C-d} \hat{G}_l(x, y+d, t) \cdot \hat{G}_r(x,y,t) \tag{12}$$

where DP represents the pixel-level disparity, R and C denote the rows and columns of the input image size. Note that the formulation here follows the conventions used in the matrix so that the disparity is on the component y.

In theory, the search range of disparity d should be the entire image width. However, in practice, we can reduce the computational cost of the search process based on some clear facts. For example, since a moving object is always continuously changing

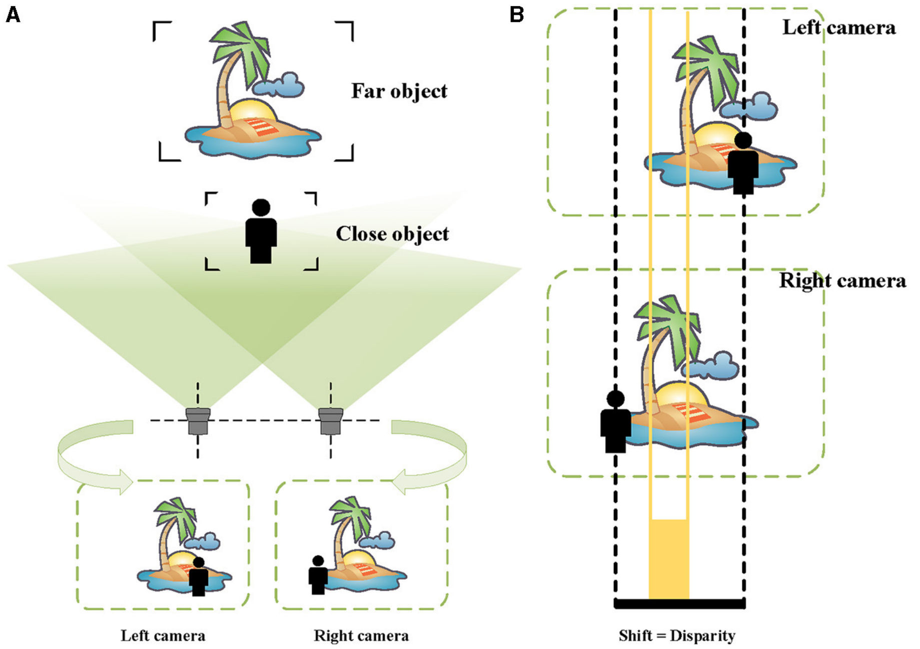


FIGURE 3
Schematic of binocular vision and disparity. **(A)** The images taken by the left and the right cameras. **(B)** For the far scenery, there is a small disparity, indicated by orange. For the person near, there is a large disparity, represented by black.

in depth distance, the results at the previous time steps can be used as a reference and searched within a reasonable range. In addition, mathematically, this optimization function usually gets a larger calculation result near its optimal disparity, so we can also quickly find the optimal disparity by jumping search.

3.2.2. Computing method of depth distance

Based on basic geometric knowledge, the depth distance between the object and the stereo cameras in the world coordinate system can be calculated using disparity. Specifically, the following relation holds when the stereo cameras with the same focal length are on the same horizontal line and the optical axes are parallel (Zhen et al., 2017; Sun et al., 2019):

$$D(t) = \frac{b \cdot f}{DP(t) \cdot pixelsize} \tag{13}$$

where D stands for the depth distance of the object. $b, f, pixelsize$ are constants, which can be obtained from the information of stereo cameras, representing the baseline length, focal length, and physical size corresponding to one pixel, respectively.

3.3. LGMD layer

After the DP layer, the proposed Bi-LGMD model is able to acquire the depth distance information. By comparing $D(t)$ and $D(t-1)$, the motion mode of the moving object at the current time t can be clearly distinguished (approaching, receding, and translating). However, to achieve a reasonable early warning response to the approaching movement, it is necessary to further judge whether the current approaching state is sufficiently dangerous. To this end, the early warning depth distance, an adaptive dynamic threshold, is introduced into our model.

3.3.1. Warning depth distance (D_W)

In fact, the proposed Bi-LGMD model also potentially extracts the approaching velocity information at each time step after DP layer. It is evident that faster moving objects require a greater warning depth-distance to ensure safety. Thus, the warning depth-distance should possess the following properties:

$$D_W(t) = F(D(t-1) - D(t)) \tag{14}$$

where $F(\cdot)$ is a strictly monotonically increasing function.

There are many functions satisfying the above basic properties. For simply, linear functions are selected for discussion in this paper. Therefore, the specific formula is as follows:

$$D_W(t) = C_T \cdot (D(t-1) - D(t)) \tag{15}$$

Although the linear function appears relatively simple, its implications are significant. The coefficient C_T holds a realistic physical interpretation, as it represents the time required for the machinery to avoid collisions, dependent upon individual machine attributes, such as flexibility in avoidance behavior. Consequently, if the moving object continues to approach at its current speed, the system will sound an early warning at the depth distance of D_W, leaving the machine C_T time to avoid collisions. It is important to note that D_W is dynamically adaptive, adjusting the warning depth-distance accordingly in response to changes in approaching speed.

By the way, as a parameter with realistic physical meaning, C_T will be set within an appropriate range. If C_T is set too large, the system may trigger an alarm prematurely. On the other hand, if C_T is set too small, the machine may not have sufficient time to complete the avoidance maneuver.

3.3.2. Activation of the LGMD neuron

In contrast to existing models that use the sigmoid function to produce activation values ranging from 0.5 to 1, our model employs a binary output: 0 and 1, representing the deactivation and activation of the LGMD neuron, respectively.

Specifically, the output of the LGMD layer is determined by two parts: one is whether the moving object is approaching, and the other is whether the moving object reaches the warning depth-distance D_W. The output of the LGMD layer is 1 only if the above two parts are both true, and 0 otherwise. In this computational mode, only approaching objects are likely to activate the LGMD neuron, while objects in other motion modes are certainly not expected to activate it. Further, even if the object is in the process of approaching, the LGMD neuron will not be activated when the object does not reach the warning depth-distance. In other words, the approaching object is in a distant position and does not pose a collision threat for the time being, so the LGMD neuron does not need to be activated.

$$LGMD(t) = \begin{cases} 1, & if\ D(t) < D_W(t)\ and\ D(t) < D(t-1) \\ 0, & otherwise \end{cases} \tag{16}$$

4. Experimental results and analysis

In this section, a series of systematic experiments will be performed from different aspects as comprehensively as possible. Also, some reasons for the experimental results will be analyzed in detail. All experiments can be divided into the following three categories: (1) Basic Synthetic Stimuli Testing, (2) Real Physical Stimuli Testing, and (3) Model Performance Testing. The state-of-the-art model (Fu et al., 2018b) will be used for comparison.

4.1. Experimental setup

For basic synthetic stimuli testing and model performance testing, all the input visual stimuli are generated using Matlab R2021b according to the projection principle (see Figure 4). The background is set to a solid color, and the pixel has a grayscale value of 0.5. For each frame, the image resolution is 600 × 600 pixels. As to real physical stimuli testing, the input visual stimuli are partly from our own recorded video (rolling ball) and partly from the publicly available KITTI dataset (vehicle scene; Geiger et al., 2013). The image resolutions are 1,280 × 720 pixels and 1,242 × 375 pixels for the videos of the rolling ball and vehicle scene, separately.

All videos are at 30 Hz, and the whole parameters are set according to this sampling rate in the experiments. We list the parameters of the proposed Bi-LGMD model in Table 1. Without special explanation, n_p is 1 and C_T is 15. For the comparative model, the parameters recommended in their literature are used. The computer is equipped with a Core i5 processor with a clock speed of 3.10 GHz, 16 GB of memory, and the operating system is Windows 10. All the experiments are conducted using MATLAB R2021b. The example video clips are shown with results in the following section.

TABLE 1 Setting parameters of the proposed Bi-LGMD model.

Parameter	Description	Value
n_p	Luminance change persistence in Equation (1)	0–2
W_I	Inhibition weight in Equation (5)	0.3
C_w	Constant to calculate w in Equation (10)	4
Δ_c	Small real number in Equation (10)	0.01
T_{de}	Decay threshold in G layer in Equation (11)	30
C_T	Time required to avoid collision in Equation (15)	10–20

4.2. Basic synthetic stimuli testing

To verify the basic validity of the proposed Bi-LGMD model, the computer-simulated stimuli are first used for testing. Common basic motion modes include the following five types: approaching, receding, translating, elongating, and shortening. In this section, all the above five types of simulated stimuli are used in the experiments. In addition, grating motion is also chosen for testing as a special phenomenon. As a collision prediction model, the most desirable result would undoubtedly be to respond only to the approaching motion, and not to any other form of movement.

Figure 4 illustrates the method of generating simulated stimulus videos required for the experiments in this section. For the proposed Bi-LGMD model, two cameras are needed to generate video data (see Figure 4A), whereas for the comparative model, only one camera is needed to generate a single video data (see Figure 4B). In addition, Figures 4C–G represent the five basic motion modes mentioned above. These data are generated by Matlab R2021b, simulated by projection transformation of the depth distance and position of the moving object. Moreover, in these experiments, the objects are all moving at a constant speed.

Figure 5 corresponds to the situation of two basic motion modes in the depth direction: approaching and receding, where Figures 5A–D show the experimental results of the proposed model and comparative model for the approaching motion, and Figures 5E–H show the experimental results of the two models for the receding motion. For each motion mode, experiments are conducted with darker and lighter objects separately to eliminate the effect of the brightness of the object relative to the background on the experimental results. However, the experimental results show that the brightness of the object has no effect on the results for either the proposed Bi-LGMD model or the comparative model, and a uniform output is given here, as shown in Figures 5C, D, G, H. For the proposed Bi-LGMD model, the computed disparity results are presented in particular, while the ground truth is also

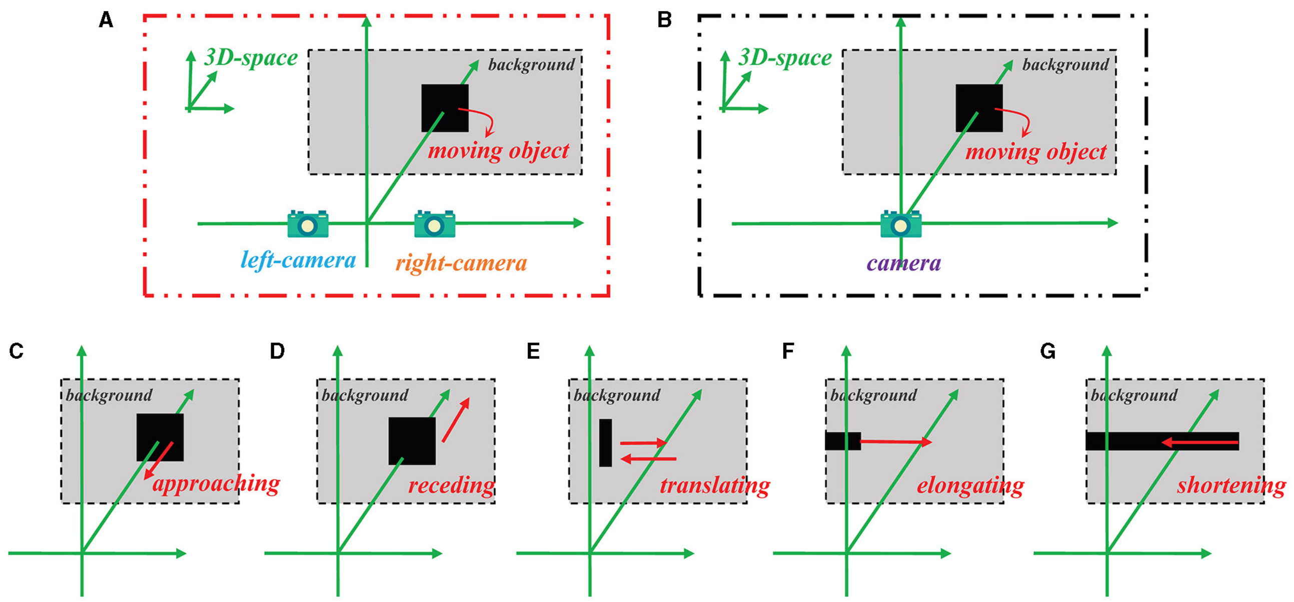


FIGURE 4
Schematic of how the stimulus videos are generated for the Bi-LGMD model and the comparative model, as well as five specific motion modes. **(A)** Binocular vision input for the Bi-LGMD model. **(B)** Monocular visual input for the comparative model. **(C)** Approaching motion. **(D)** Receding motion. **(E)** Translating motion. **(F)** Elongating motion. **(G)** Shortening motion.

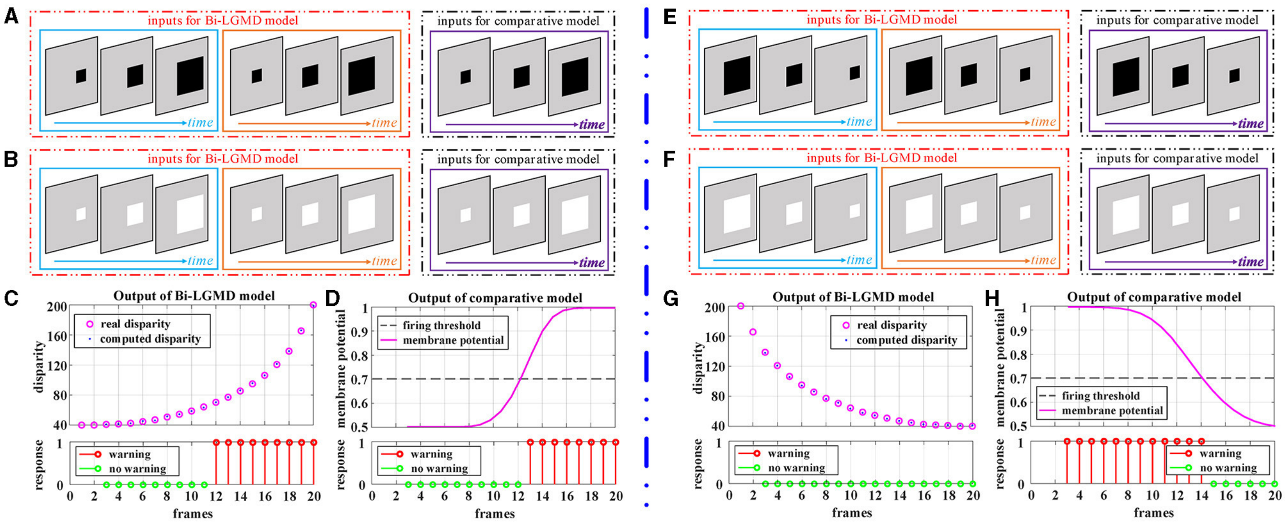


FIGURE 5
Experimental results of the proposed Bi-LGMD model and the comparative model for simulated stimuli moving in the direction of depth. **(A)** Schematic diagram of an approaching darker object. **(B)** Schematic diagram of an approaching brighter object, with the same motion process as in **(A)**. Identical experimental results for the two sets of simulated stimuli are shown in **(C, D)**. **(C)** The output of the proposed Bi-LGMD model, including the computed disparity (compared with the ground-truth), as well as the final response. **(D)** The output of the comparative model, including the sigmoid membrane potential (SMP), and its comparison to a given hard threshold (set to 0.7). Similarly, for the receding motion in the depth direction, the corresponding schematic and experimental results are presented in **(E–H)** in the same way. Specifically, **(E)** schematic diagram of a receding darker object. **(F)** Schematic diagram of a receding brighter object, with the same motion process as in **(E)**. **(G)** The output of the proposed Bi-LGMD model. **(H)** The output of the comparative model.

marked for comparison. Based on this, the Bi-LGMD model can calculate the depth distance information and obtain a final 0–1 binarized response output. For the comparative model, the sigmoid membrane potential (SMP) is shown and combined with a given hard threshold (set to 0.7), and the same form of response output is obtained for inter-model comparison.

From the experimental results, it can be seen that the disparity calculated by the Bi-LGMD model matches the

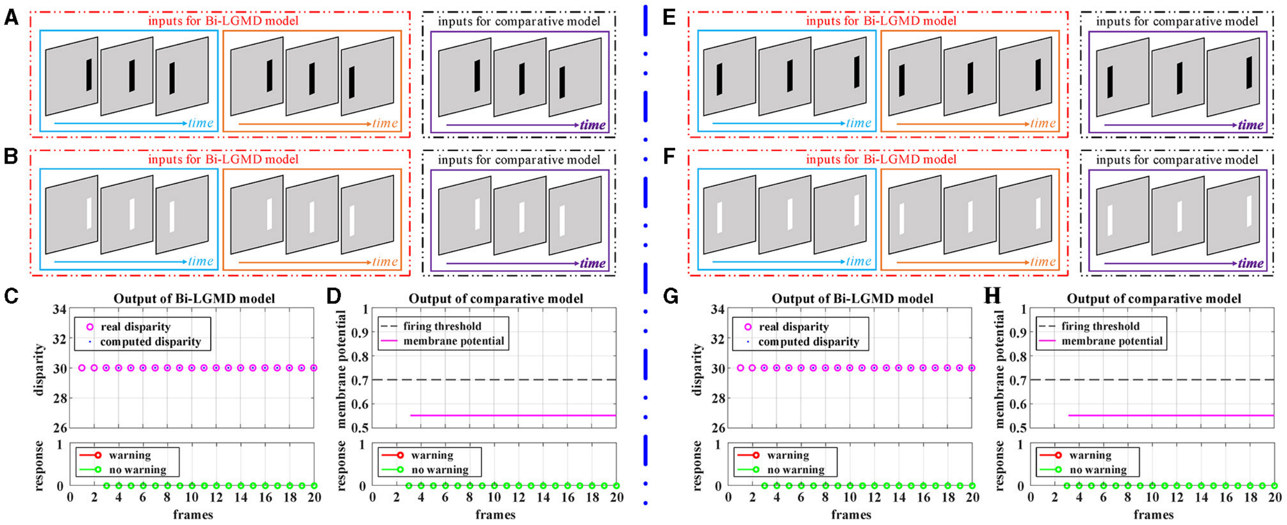


FIGURE 6
Experimental results of the proposed Bi-LGMD model and the comparative model for simulated stimuli of translating motion (no change in depth distance). **(A)** The translating leftward darker object. **(B)** The translating leftward brighter object. **(C)** The output of the proposed Bi-LGMD model. **(D)** The output of the comparative model. **(E)** The translating rightward darker object. **(F)** The translating rightward brighter object. **(G)** The output of the proposed Bi-LGMD model. **(H)** The output of the comparative model.

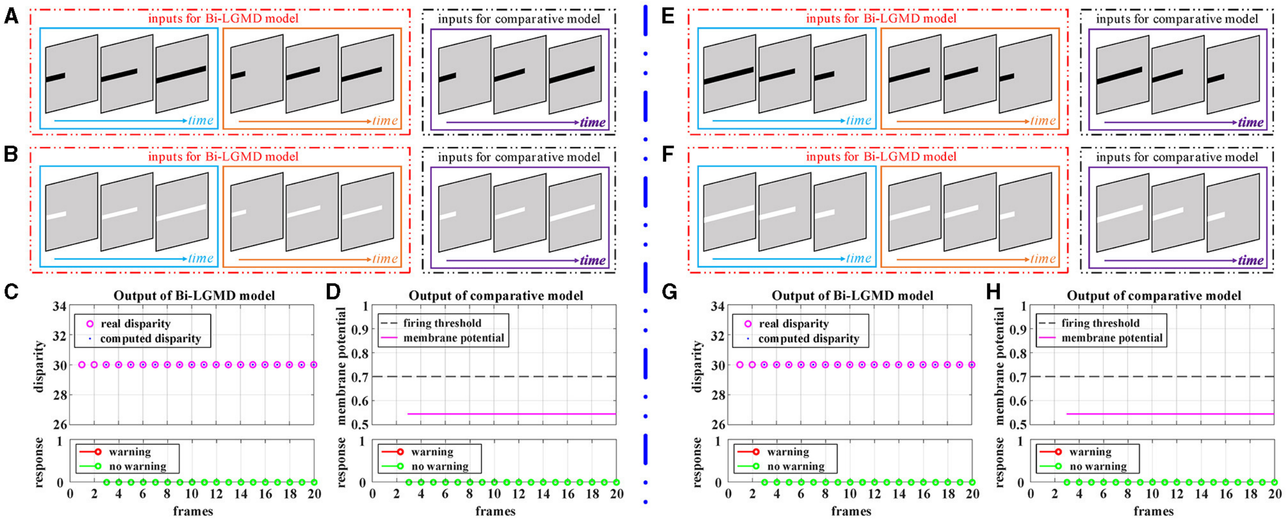


FIGURE 7
Experimental results of the proposed Bi-LGMD model and the comparative model for simulated stimuli of elongating and shortening motion (no change in depth distance). **(A)** The elongating darker object. **(B)** The elongating brighter object. **(C)** The output of the proposed Bi-LGMD model. **(D)** The output of the comparative model. **(E)** The shortening darker object. **(F)** The shortening brighter object. **(G)** The output of the proposed Bi-LGMD model. **(H)** The output of the comparative model.

actual value perfectly. Moreover, the model responds to the approaching stimulus, while it remains unresponsive to the receding process of the object. In fact, as we know, when the object recedes, the disparity of the moving object decreases gradually. Therefore, the model calculates that the depth distance of the object is getting larger, and then, the output of the LGMD layer will be 0, which makes the final result unresponsive. On the contrary, when the object is approaching, in the initial stage, the model calculates that the moving object is far away, so there is no response temporarily. However, as the object gets closer and closer, once the warning depth-distance is reached, the model quickly produces a lasting response.

Figure 6 shows the experimental results of the proposed Bi-LGMD model and the comparative model for simulated stimuli of translating motion. As can be seen, whether the direction of translation is to the left or to the right, and whether the moving object is darker or brighter, the final response is always 0 for the proposed Bi-LGMD model. In fact, in the three-dimensional real world, when a moving object is translating horizontally, it is always at the same depth distance, so the disparity of the moving object keeps unchanged. The proposed Bi-LGMD model attempts to capture exactly this core feature and, from the computational results, the model does indeed accurately extract the correct disparity results and therefore achieves satisfactory results. For the comparative model, the final response is also always 0. However, as we have seen, the result is based on the comparison of the SMP with a given threshold, so there is conceivably the possibility that the model parameters could have a serious effect on the final result. Furthermore, the fact that the comparative model is based on the summation of the pixel values output from the *G* layer means that the SMP is also affected by the translating speed of the moving object. Overall, the proposed model effectively extracts more essential depth-distance information and will therefore behave more robustly.

Elongating and shortening movements, which are special cases of translating motion, only show a single translating edge due to the limited visual field. However, especially for elongating motion, one-sided changes can easily be confused for edge expansions, which are then misinterpreted by the model as approaching movements. Figure 7 shows the experimental results of the proposed Bi-LGMD model and the comparative model for the moving object in the process of elongating and shortening. As can be seen, the proposed model still extracted the correct disparity information very well and obtained satisfactory experimental results. The experimental analysis for this group of tests is similar to that in translational motion and will not be repeated here.

Grating movement is a very common phenomenon in our daily life. For example, when the sun shines on the front windshield of a moving car, we can see the bright and dark grating moving stripes from the driver's seat. Obviously, the ideal model does not need to respond to this. However, the grating motion is always accompanied by the luminance change of the whole field, resulting in an easily observable response in the model. In order to suppress this unnecessary response, the existing LGMD-based models introduce the feedforward inhibition (FFI) mechanism. However, no evidence has been found to show how feedforward inhibition could increase the selectivity for approaching over receding objects (Keil and Rodriguez-Vazquez, 2003), and from the perspective of biological neurology, there are still some doubts about the explanation and rationality of it. Moreover, the parameter setting of the FFI mechanism itself is also a relatively complex problem. Figure 8 shows the experimental results of the proposed Bi-LGMD model and the comparative model for simulated stimuli of grating motion. As can be seen, both models achieve the desired non-response result. However, the two models do not work in the same way. The proposed Bi-LGMD model is based on the computed disparity information, and since there is no change in depth distance, it is judged that there is no collision risk. In the comparative model, the FFI mechanism is triggered by the change of pixel gray value in a large area, forcing the response of the model to be suppressed. It is worth noting that the spacing between the grating stripes, and the moving speed, moving direction as well as the brightness of the grating stripes will not affect the experimental results of the proposed Bi-LGMD model. In fact, the "disparity" and "depth distance" are always the essence in any case, and they are not affected by the above factors. Therefore, the Bi-LGMD model can easily judge the grating motion as a translating motion.

So far, in all five basic motion modes as well as the grating motion, the Bi-LGMD model only responds to the approaching motion, while remaining unresponsive to any other motion modes, which fully meet our expectations. Moreover, such response results are independent of the brightness of the moving object. These results are largely due to the fact that the Bi-LGMD model obtains the depth distance of the object by calculating the disparity, and thus further effectively distinguishing the approaching motion mode from others. Actually, according to the computed disparity, the Bi-LGMD model can clearly classify various specific motion modes into the following three categories: approaching, receding, and translating. In addition, for approaching motion, the model will further extract the approaching velocity at each time step, combined with the current depth distance information, the model only generates a collision warning if it actually perceives the threat of an imminent collision, that is, if the object reaches a dynamically adaptive warning depth-distance.

4.3. Real physical stimuli testing

In the previous section, the validity and superiority of the proposed Bi-LGMD model is initially verified by simulated stimuli. In this section, real physical stimuli are used for testing. Compared with the computer-simulated stimuli, the biggest difference is that there is more environmental noise in the real physical scenes, such as shadows, reflections, etc. In addition, the motion speed and motion state of moving objects are also relatively unstable. Therefore, visual stimulation in real physical scenes is undoubtedly a more difficult challenge for the collision prediction task, but at the same time, it is also one of the important criteria to evaluate the performance of the model.

Firstly, the videos of a small moving ball taken indoors are used for testing. Two GoPro motion cameras of the same model (Hero 8 Black) are used to capture the scene simultaneously. The optical axes are kept parallel throughout the entire shooting process. The experimental results are shown in Figure 9. In the approaching ball video, the green ball is approaching from a distance along a fixed oblique track. Due to a certain inclination of the track, under the action of gravitational potential energy, the approaching speed of the ball gradually accelerates, and the ball bounces on the table after it gets off the track in the later stage. There are obvious shadows, reflections, and so on in the video. It can be seen from the experimental results that both models produce an early warning response, in which the proposed model has an earlier warning time, while the comparative model produces the early warning response in the late stage when the ball's approaching speed is faster. Reverse the video sequence to simulate the receding process, and the Bi-LGMD model has no response to that because the computed disparity is getting smaller over time. However, the comparative model has two early warnings at the beginning. The outputs are not shown here for brevity. For the translating ball video (in fact, it is difficult to ensure that the ball moves strictly in translation, so the ball is not always at the same depth distance. The so-called translation here is just a rough visual effect.), no warning response was generated for both models.

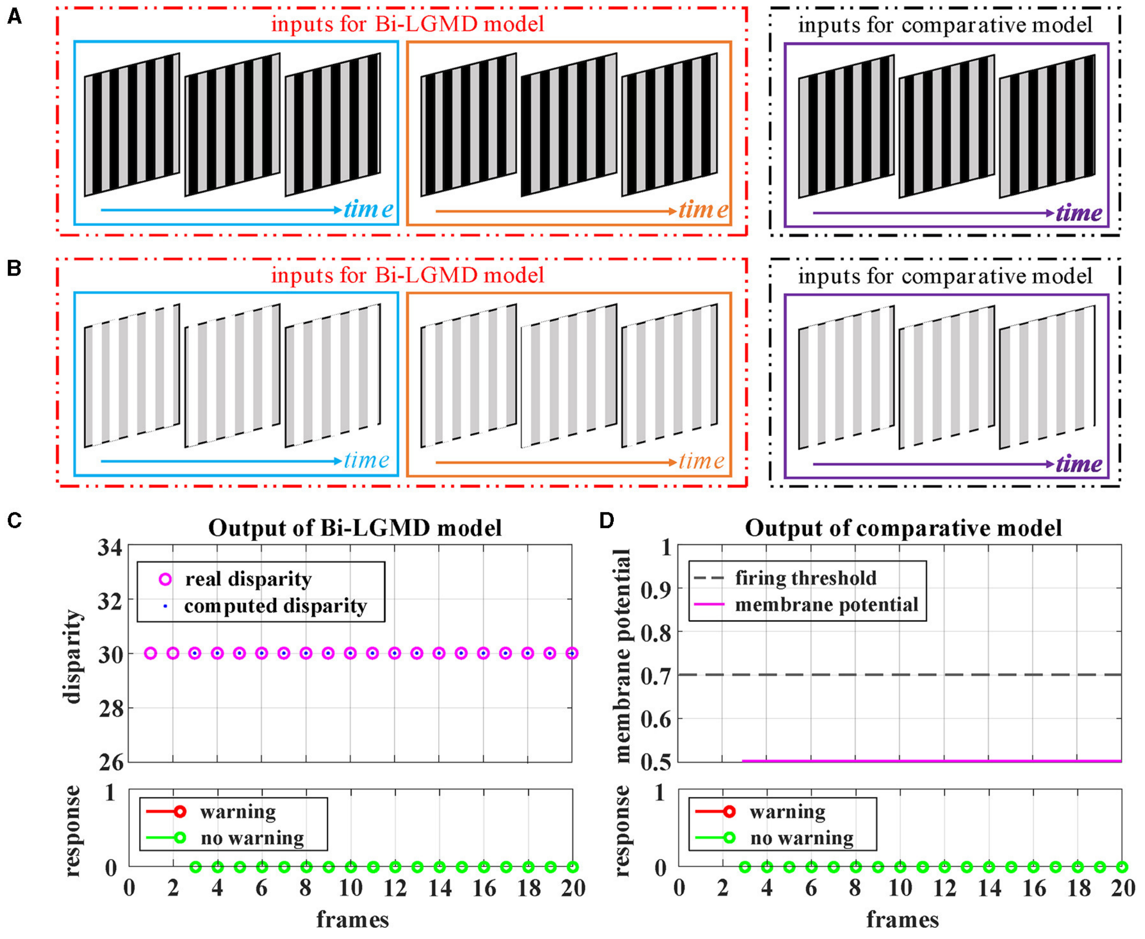


FIGURE 8
Experimental results of the proposed Bi-LGMD model and the comparative model for simulated stimuli of grating motion (no change in depth distance). **(A)** The grating motion with darker stripes. **(B)** The grating motion with brighter stripes. **(C)** The output of the proposed Bi-LGMD model. **(D)** The output of the comparative model.

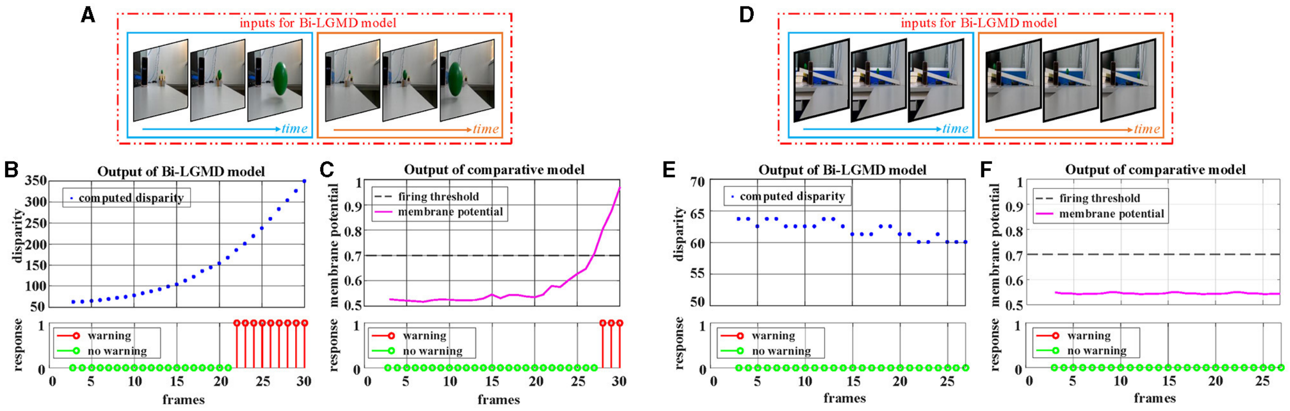


FIGURE 9
Experimental results of the proposed Bi-LGMD model and the comparative model for real scene videos of indoor moving ball. **(A)** The input image streams of a approaching ball. The blue and orange boxes indicate inputs from the left and right cameras, respectively. **(B)** The output of the proposed Bi-LGMD model, including the computed disparity, as well as the final response. **(C)** The output of the comparative model, including the sigmoid membrane potential (SMP), and its comparison to a given hard threshold (set to 0.7). Similarly, for the translating ball, the corresponding schematic and experimental results are presented in **(D–F)** in the same way. Specifically, **(D)** the input image streams of a translating ball. **(E)** The output of the proposed Bi-LGMD model. **(F)** The output of the comparative model.

Following that, the outdoor vehicle scene videos are used for testing. Here, the public KITTI data set is adopted. For each of the three basic motion modes, a video is chosen for the experiment, as shown in Figure 10. For approaching motion, a white vehicle is approaching from the front as shown in Figure 10A. For receding motion, a black vehicle drives away as shown in Figure 10D. For translating motion, a white car moves from the left to the right in the field of view as shown in Figure 10G. It can be seen that the experimental results of the proposed Bi-LGMD model are fully in line with expectations, and can effectively calculate parallax and obtain satisfactory model output based on depth and distance information. Compared with the comparative model, the explainability and robustness of the Bi-LGMD model are stronger, especially for the backward motion, the Bi-LGMD model shows better experimental results.

4.4. Model performance testing

As a binocular LGMD-based visual neural network for collision prediction, Bi-LGMD is fundamentally different from the existing models in many aspects. The estimation of the depth distance of a moving object, indeed, brings great benefits to the model. In this section, we will discuss this in detail, and analyze the advantages of Bi-LGMD by comparing it with existing models. In the following experimental comparison, since monocular and binocular stimuli need to be generated correspondingly, we use computer-simulated synthetic stimuli to carry out the experiment.

4.4.1. Sensitivity to model parameters

Parameters are undoubtedly crucial for any model and even have a direct impact on the model results. In this part, the topic of parameters of Bi-LGMD and existing models will be discussed. In fact, as we can see, the basic process (P, E, I, S, G layers) of the proposed Bi-LGMD model is consistent with existing models, therefore the parameters after the G layer will mainly be discussed.

In existing models, the following function is used to activate the summation result of G layer as the output of the LGMD layer (representing the membrane potential of the LGMD neuron). After that, a given firing threshold T_{fir} is used to determine whether the LGMD neuron is activated, such as Yue and Rind (2006), Fu et al. (2018b, 2019b, 2020), Luan et al. (2021), Lei et al. (2022), and Li et al. (2022).

$$LGMD(t) = \left(1 + exp(\frac{-\sum_{x=1}^{R}\sum_{y=1}^{C}\widehat{G}(x,y,t)}{\alpha \cdot R \cdot C})\right)^{-1} \tag{17}$$

Therefore, there are two important parameters involved: α and T_{fir}. Obviously, the existing models must fully consider the problem that the given threshold should roughly match the activation result, which is actually relatively difficult to adjust adaptively. As we know, the sigmoid function curve $y(x) = [1 + exp(-x)]^{-1}$ increases monotonically, with a range of 0.5–1. For a standard collision process that gradually approaches from a distance, the ideal sigmoid activation result should be approximately from 0.5 to nearly 1, which requires that the parameter α is very suitable so that the ratio $\sum_{x=1}^{R}\sum_{y=1}^{C}\widehat{G}(x,y,t)/(\alpha \cdot R \cdot C)$ could almost fill the interval $[0, 3]$ since $y(3) \approx 0.9526$. In other words, if the α is chosen too large so that the ratio is very small, the sigmoid activation results will be all near 0. Conversely, if the α is chosen too small, resulting in the ratio being basically >3, the sigmoid activation results will be all around 1. Obviously, in these cases, it is difficult to match the sigmoid activation results with the given thresholds. In addition, it can be seen from the formula that the value of α will also be affected by the image sizes R and C, which means that for the same collision scenario, cameras with different resolutions or different fields of view will have a serious impact on the model, which makes it more difficult to determine the parameters α. In summary, the existing models are very sensitive to the above two parameters (α and T_{fir}), making them less robust.

By contrast, in the Bi-LGMD model, there is only one parameter D_W after G layer. Furthermore, this parameter $D_W(t)$ is adaptively adjusted with the motion state of the object at each time step. In more detail, D_W is linearly determined by C_T for convenience in our case, and C_T is given a very clear realistic physical meaning, which can be used as a guide for adjusting.

In addition, Figure 11 demonstrates the impact of these parameters on the proposed model and the comparative model. The video stimuli used in the experiment were simulated approaching black blocks similar to those shown in Figure 4C. To more comprehensively illustrate the impact of parameters on the model, we set the following motion pattern: the object remains stationary for the first 15 frames, then begins to approach and stops approaching at frame 37. The speed remains constant during the approaching process.

It can be seen from the experimental results that the parameter α has a great impact on the sigmoid membrane potential results of the existing models, and if an inappropriate value α is selected, the existing models will fail (under the given firing threshold T_{fir}). Contrastingly, the influence of parameter C_T on the results of the proposed Bi-LGMD model is mainly reflected in the early warning response time. Specifically, the larger the C_T value, the earlier the early warning response time. However, the Bi-LGMD model will always produce a warning before the collision. In addition, according to the actual physical meaning of C_T, we can reasonably adjust the value range of C_T based on the system performance.

Hence, the proposed model has fewer parameters and is more robust than the existing model. In terms of parameter adjustment, the proposed model has more clear guiding significance, so it can be considered that the proposed model is superior to the existing models in this respect.

4.4.2. Adaptability to motion modes

By estimating the depth distance of a moving object at each time step, the Bi-LGMD model accurately identifies its motion modes, as seen in the previous experiment. To further fully illustrate the advantage of estimating depth distance, more detailed motion patterns are used for testing. Since the Bi-LGMD model does not respond to receding and translating motion, we mainly take the approaching motion as an example to illustrate. In particular, unlike the previous experiments in which the object is always moving at a constant velocity, we will explore other different approaching cases. Similar to the experimental setup in Figure 11, during the first 15

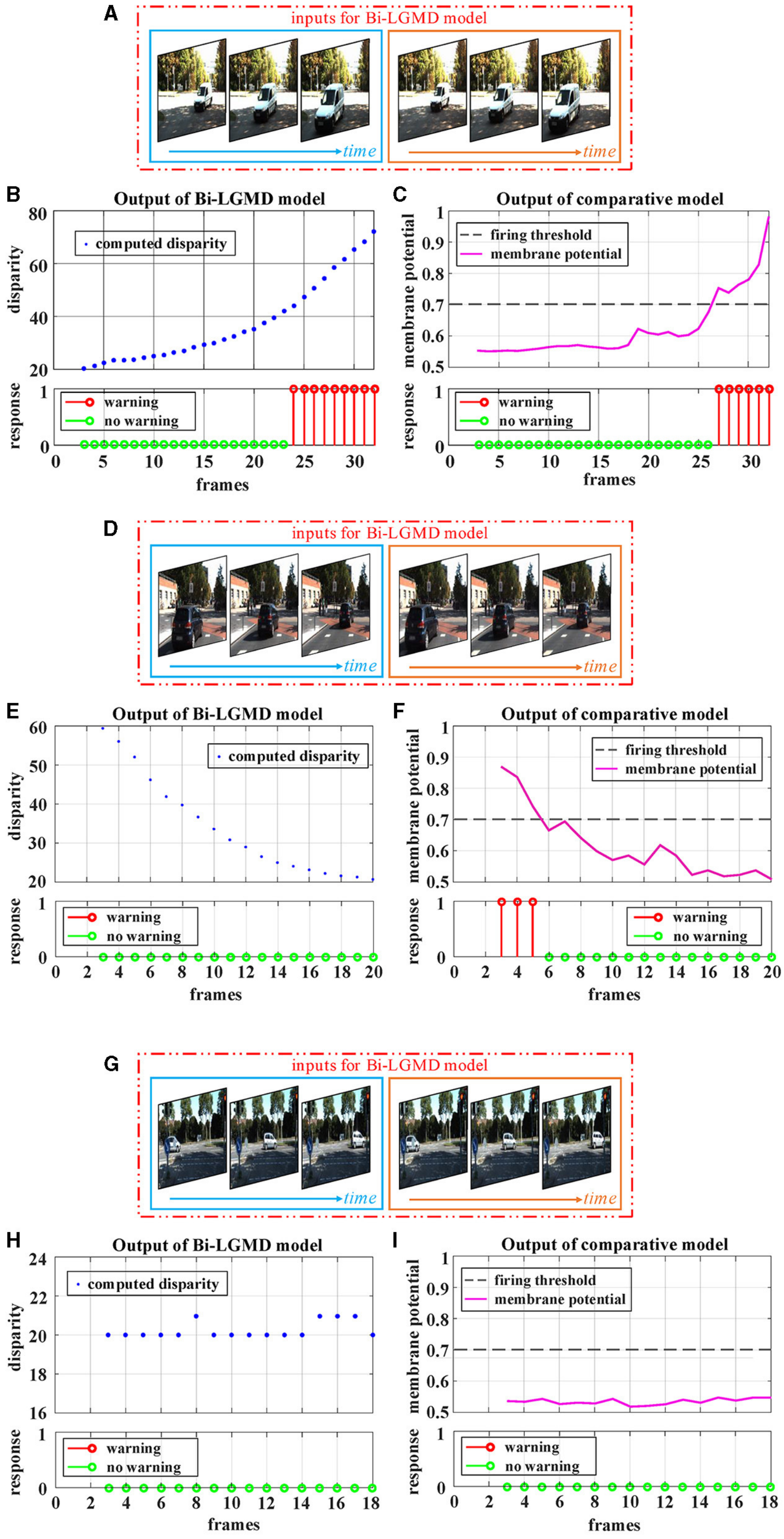


FIGURE 10
Experimental results of the proposed Bi-LGMD model and the comparative model for real scene videos of outdoor moving vehicle. There are three sets of experiments, each showing examples of the input image streams and the corresponding output of the two models. **(A)** The input of a approaching vehicle. **(D)** The input of a receding vehicle. **(G)** The input of a translating vehicle. The blue and orange boxes indicate inputs from the left and right cameras, respectively. **(B, E, H)** Are the outputs of the proposed Bi-LGMD model, including the computed disparity, as well as the final responses. **(C, F, I)** Are the outputs of the comparative model, including the sigmoid membrane potential (SMP), and its comparison to a given hard threshold (set to 0.7).

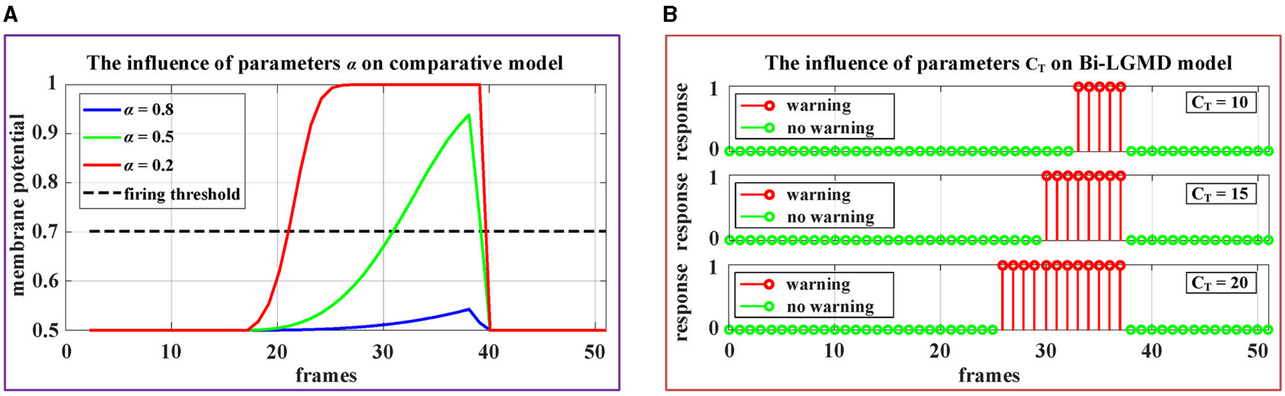


FIGURE 11
Experimental results of the effects of parameters on the comparative model and the proposed Bi-LGMD model. **(A)** The experimental results of the comparative model. **(B)** The experimental results of proposed Bi-LGMD model.

frames and the last 15 frames, the object remains stationary in the simulated stimulus video. The experimental description and results are shown in Figure 12.

Figure 12A depicts three different approaching patterns in terms of depth distance and image size over time, represented by different colors. Among them, the mode represented by the green line is approaching at a constant speed (Marked as **Approaching Pattern 1**), which is the pattern set in all previous experiments. In particular, the pattern represented by the blue line is a special deceleration approach, leading to a linear increase in imaging size (Marked as **Approaching Pattern 2**). The pattern represented by the red line is also a deceleration approach, leading to a gradual decrease in the increment of imaging size (Marked as **Approaching Pattern 3**).

Figures 12B, C shows the experimental results of the comparative model and the proposed Bi-LGMD model for the Approaching Pattern 2. As can be seen, the SMP output of the comparative model is almost a horizontal straight line, indicating that the activity of LGMD cells is always maintained at the same level. In fact, parameter α does not change the overall shape of the response, so that the model either reaches the given firing threshold at the beginning of movement or never, both of which are not the ideal results. Such experimental results are directly related to the fact that the imaging size varies linearly. By contrast, the Bi-LGMD model only outputs 1 for the first few frames when the object begins to approach, and 0 for the rest of the time. This result is actually reasonable. As can be seen, the approaching speed is very fast at the beginning, so the model needs to trigger an early warning immediately. However, when the approaching speed of the moving object gradually slows down, there is no collision threat temporarily, so the output changes to 0. The warning depth-distance D_W of each time step obtained from the model is shown in **(E)**, when the approaching speed slows down, the warning depth-distance D_W decreases accordingly, which reflects its dynamic adaptive process. Similarly, Figures 12D, E shows the experimental results of the comparative model and the proposed Bi-LGMD model for the Approaching Pattern 3.

In summary, the comparative model is not well-adapted to various approach models, while the Bi-LGMD model can achieve satisfactory results based on depth distance estimation, as well as the dynamic adaptive warning depth distance mechanism.

4.4.3. Robustness to the input image streams

Robustness is one of the important indexes for model evaluation. In the existing models, the quality of the input image streams has a certain impact on the results, which makes the model not robust enough. In this section, we select two key factors affecting image quality (contrast and noise) for testing. We make a detailed analysis based on the results, and further compare the differences between Bi-LGMD and the existing models.

4.4.3.1. Contrast

The contrast between the moving object and the background is obviously a very important factor. In this group of experiments, since both the background and the moving object are set to a solid color, the contrast ratio can be simply regarded as the gray value of the background (the gray value of the moving darker object is set to 0). Intuitively, the greater the contrast, the easier it is for the model to recognize moving objects and successfully perceive collisions. But as the contrast gradually decreases, the task of sensing collisions becomes more difficult.

Figure 13 shows the approaching motion with three different contrasts. The motion process is based on the Approaching Pattern 1 shown in Figure 12A. For the above three cases, we generated monocular data and binocular data according to the imaging principle. As can be seen, in the comparative model, the higher the contrast, the stronger the activation result of the sigmoid membrane potential of the LGMD cell. Therefore, in the case of low contrast, the activation result is far lower than the given threshold, which makes the model unable to successfully perceive collisions and generate early warnings. However, for the proposed Bi-LGMD model, even if the contrast is small enough, the output of the model is still not affected at all. It is because the Bi-LGMD model does

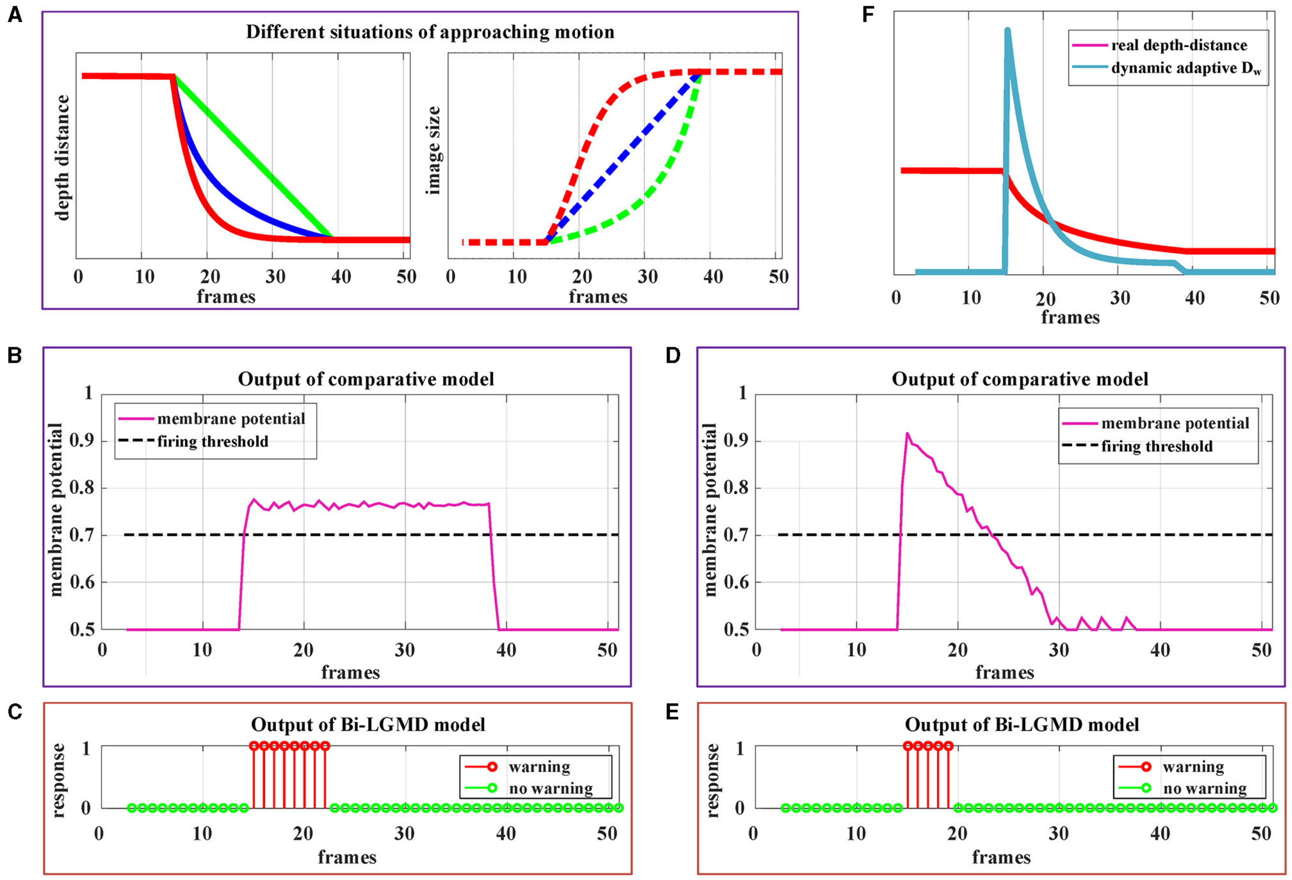


FIGURE 12
Experimental results of the proposed Bi-LGMD model and comparative model for three different approaching modes. **(A)** The specific details of three different approaches modes. Mode represented by the green line is approaching at a constant speed. Mode represented by the blue line is a special deceleration approach, leading to a linear increase in imaging size. Mode represented by the red line is also a deceleration approach, leading to a gradual decrease in the increment of imaging size. **(B)** The output of the comparative model for the approaching mode represented by the red blue in **(A)**. **(C)** The output of the proposed Bi-LGMD model for the approaching mode represented by the blue line in **(A)**. **(D)** The output of the comparative model for the approaching mode represented by the red line in **(A)**. **(E)** The output of the proposed Bi-LGMD model for the approaching mode represented by the red line in **(A)**. **(F)** The dynamic adaptive warning depth distance (D_W) and depth distance in **(E)**.

not care about the pixel value, but only needs to match the relevant position of the moving object from the left and right camera to obtain the correct disparity, so as to determine the depth distance of the moving object. As shown in Figure 13D, the model converts the focus from pixel value to corresponding position matching, which is a major difference in the Bi-LGMD model. Under this change of thinking, the model does not rely on the absolute size of the pixel value, so no matter how the contrast is, the pixel position matching is still accurate. Therefore, the contrast factor has no effect on the estimation of the depth distance of the moving object, so naturally, it does not affect the final effect at all.

4.4.3.2. Image noise

In the previous section, all synthetic stimuli used in the experiment are clean. However, the input image streams in the real world are always accompanied by different kinds and degrees of noise, which is caused by hardware equipment and other factors. In other words, noise is often an inevitable objective factor in image sampling. To test the robustness of the model to noise, different levels of White Gaussian Noise are randomly added to the synthetic stimulus.

Similar to the experiment on contrast, there are three groups of approaching processes with different levels of noise, as shown in Figure 14. Gaussian noise variances (GNV) from left to right are 0.01 (slight noise, green), 0.02 (moderate noise, yellow), and 0.05 (serious noise, blue), respectively. It can be seen that noise has a serious impact on existing models, while the Bi-LGMD model is very robust. The reasons here are the same as those mentioned above. For the existing model, the noise seriously affected the pixel value, thereby affecting the results of the model. However, for Bi-LGMD, the matching of corresponding positions is relatively stable.

5. Further discussion

As a research based on binocular LGMD visual neural network, this paper proposes a novel model with depth distance as the essential feature, and verifies the feasibility and superiority of this idea through systematic experiments. In fact, the advantages of

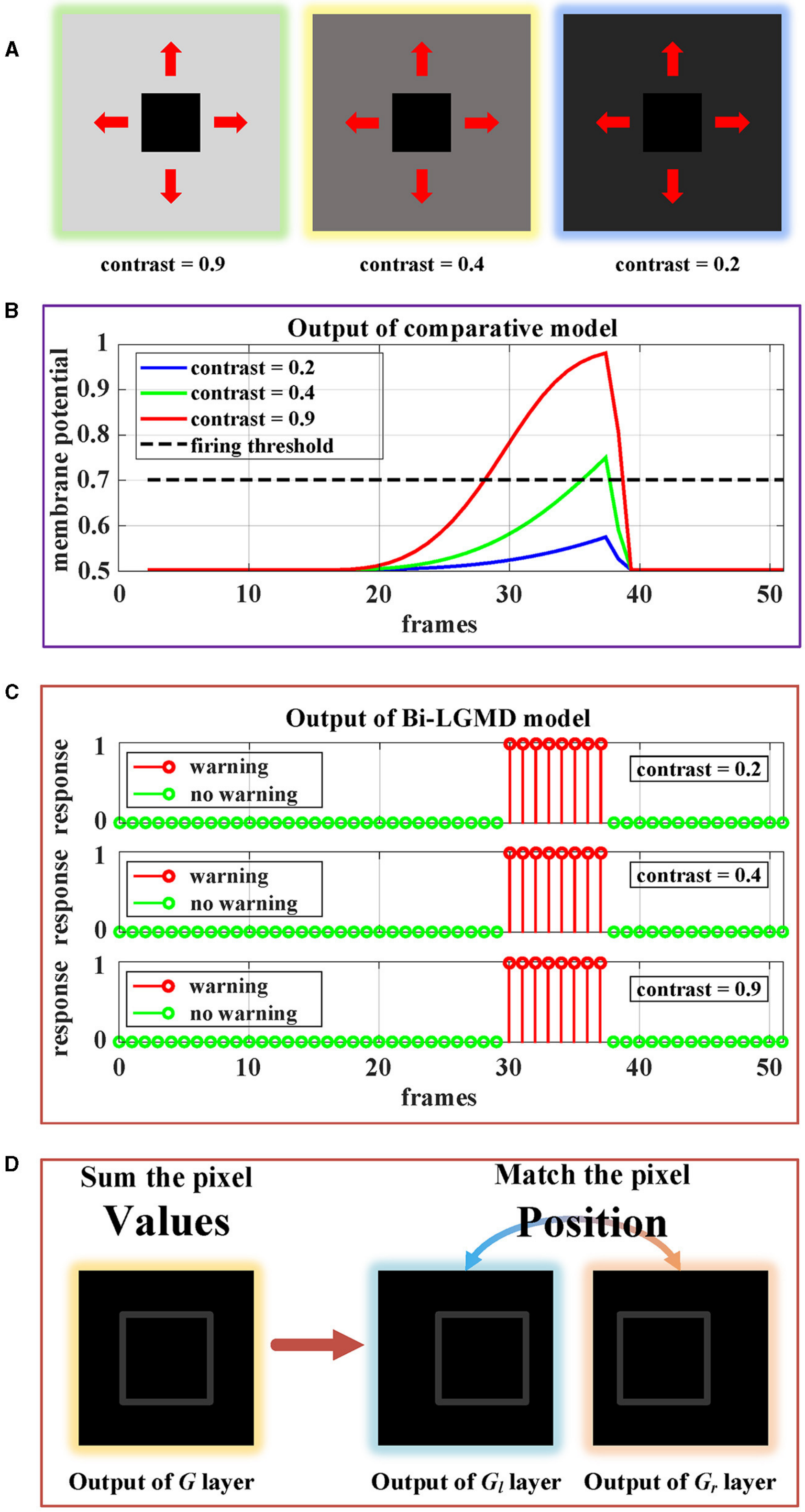


FIGURE 13
Experimental results of the proposed Bi-LGMD model and the comparative model for same approaching process with different contrast. **(A)** Visual examples in three different contrasts, decreasing from left to right. **(B)** Experimental results of the comparative model. **(C)** Experimental results of the proposed Bi-LGMD model. **(D)** Schematic of the essential differences between the two models when dealing with low contrast problems.

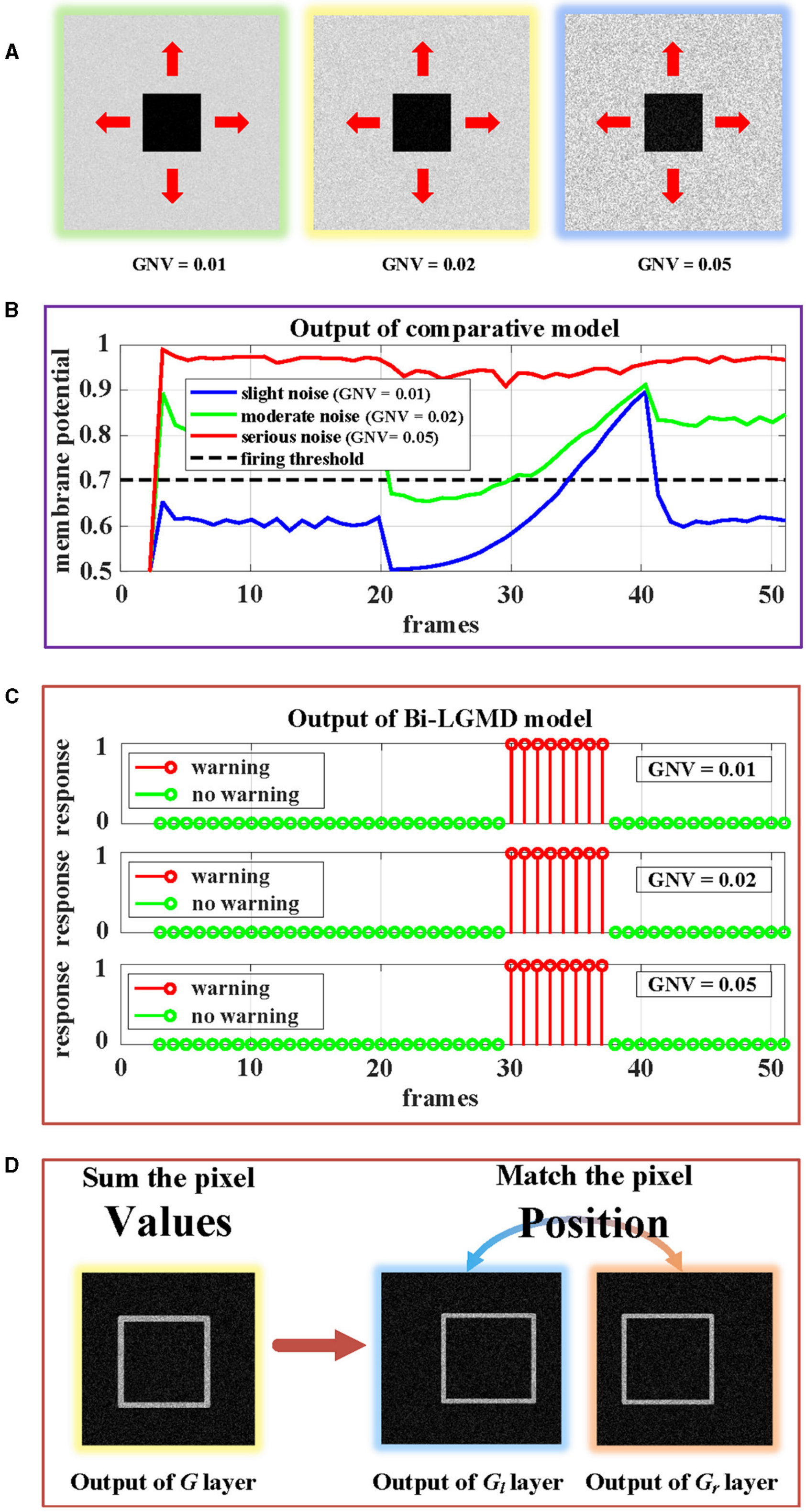


FIGURE 14
Experimental results of the proposed Bi-LGMD model and the comparative model for same approaching process with different levels of noise. **(A)** Visual examples in three different levels of noise, increasing from left to right. **(B)** Experimental results of the comparative model. **(C)** Experimental results of the proposed Bi-LGMD model. **(D)** Schematic of the essential differences between the two models when dealing with noise.

introducing depth distance into models are not limited to the work described in this paper. On the basis of the proposed Bi-LGMD model, there are more research directions worth exploring in the future.

Two points are briefly listed here: (1) For the case of multiple moving objects, the existing LGMD-based models are difficult to obtain ideal results due to the mixture of multiple stimuli. However, based on the proposed Bi-LGMD model, it is possible to distinguish moving objects at different depth distances and obtain the motion pattern of each object to achieve better model results. (2) More exploration of the approaching azimuth of the moving object can be attempted. Obviously, as a collision prediction model, it needs to respond strongly to stimuli that approach directly from the front, while it does not need to respond to the oblique approach motion such as passing-by. Based on the Bi-LGMD model and making full use of depth distance information, these ideas above will be our follow-up research.

6. Conclusion

This paper presents a LGMD-based neural network with binocular vision for collision prediction. In this model, the depth-distance information of moving objects is further taken into account, which enables the model to correctly distinguish between approaching and other modes of motion, and the model results are more interpretable. Moreover, the early warning depth-distance parameter in the proposed model is designed to be dynamically adaptive, which allows the model to generate early warnings at the most appropriate time depending on the individual performance of the system, which is a great improvement over existing LGMD-based models. The model no longer depends on the activation function and a given hard threshold, which mitigates the sensitivity to model parameters. The proposed Bi-LGMD visual neural network model is systematically tested on synthetic stimuli and real-world scene videos, showing that it is effective and robust to input quality, such as noise, low contrast, and other factors. Unlike existing LGMD-based models that rely heavily on image pixel values, the Bi-LGMD model shifts the focus to position matching, which may be a new line of research to consider in the future.

Data availability statement

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

Author contributions

YZ, HL, and JP contributed to conception and design of the study. YZ and YW organized the database. JP performed the statistical analysis. YZ wrote the first draft of the manuscript. YZ, YW, and GW wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

Funding

This work was supported in part by the National Natural Science Foundation of China under Grants 12031003 and 12271117 and in part by the Basic Research Joint Funding Project of University and Guangzhou City under Grant 202102010434.

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

Ayache, N. (1991). *Artificial Vision for Mobile Robots: Stereo Vision and Multisensory Perception*. Cambridge, MA: MIT Press.

Čížek, P., Milička, P., and Faigl, J. (2017). "Neural based obstacle avoidance with CPG controlled hexapod walking robot," in *2017 International Joint Conference on Neural Networks (IJCNN)* (Anchorage: IEEE), 650–656.

Collett, T. S. (1996). Vision: simple stereopsis. *Curr. Biol.* 6, 1392–1395. doi: 10.1016/S0960-9822(96)00739-7

Dewell, R. B., and Gabbiani, F. (2018). Biophysics of object segmentation in a collision-detecting neuron. *ELife* 7:e34238. doi: 10.7554/eLife.34238

Dewell, R. B., and Gabbiani, F. (2019). Active membrane conductances and morphology of a collision detection neuron broaden its impedance profile and improve discrimination of input synchrony. *J. Neurophysiol.* 122, 691–706. doi: 10.1152/jn.00048.2019

Ding, J., Yan, Z., and We, X. (2021). High-accuracy recognition and localization of moving targets in an indoor environment using binocular stereo vision. *ISPRS Int. J. Geo-Inform.* 10:234. doi: 10.3390/ijgi10040234

Eichler, K., Li, F., Litwin-Kumar, A., Park, Y., Andrade, I., Schneider-Mizell, C. M., et al. (2017). The complete connectome of a learning and memory centre in an insect brain. *Nature* 548, 175–182. doi: 10.1038/nature23455

Franceschini, N. (2014). Small brains, smart machines: from fly vision to robot vision and back again. *Proc. IEEE* 102, 751–781. doi: 10.1109/JPROC.2014.2312916

Fu, Q. (2023). Motion perception based on on/off channels: a survey. *Neural Netw.* 165, 1–18. doi: 10.1016/j.neunet.2023.05.031

Fu, Q., Bellotto, N., Wang, H., Claire Rind, F., Wang, H., and Yue, S. (2019a). "A visual neural network for robust collision perception in vehicle driving scenarios," in *IFIP International Conference on Artificial Intelligence Applications and Innovations* (Cham: Springer), 67–79. doi: 10.1007/978-3-030-19823-7_5

Fu, Q., Hu, C., Liu, P., and Yue, S. (2018a). "Towards computational models of insect motion detectors for robot vision," in *Towards Autonomous Robotic Systems: 19th Annual Conference, TAROS 2018, Vol. 10965* (Bristol: Springer), 465.

Fu, Q., Hu, C., Liu, T., and Yue, S. (2017). "Collision selective LGMDs neuron models research benefits from a vision-based autonomous micro robot," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Vancouver, BC: IEEE), 3996–4002. doi: 10.1109/IROS.2017.8206254

Fu, Q., Hu, C., Peng, J., Rind, F. C., and Yue, S. (2019b). A robust collision perception visual neural network with specific selectivity to darker objects. *IEEE Trans. Cybern*. 50, 5074–5088. doi: 10.1109/TCYB.2019.2946090

Fu, Q., Hu, C., Peng, J., and Yue, S. (2018b). Shaping the collision selectivity in a looming sensitive neuron model with parallel on and off pathways and spike frequency adaptation. *Neural Netw*. 106, 127–143. doi: 10.1016/j.neunet.2018.04.001

Fu, Q., Sun, X., Liu, T., Hu, C., and Yue, S. (2021). Robustness of bio-inspired visual systems for collision prediction in critical robot traffic. *Front. Robot. AI* 245:529872. doi: 10.3389/frobt.2021.529872

Fu, Q., Wang, H., Hu, C., and Yue, S. (2019c). Towards computational models and applications of insect visual systems for motion perception: a review. *Artif. Life* 25, 263–311. doi: 10.1162/artl_a_00297

Fu, Q., Wang, H., Peng, J., and Yue, S. (2020). Improved collision perception neuronal system model with adaptive inhibition mechanism and evolutionary learning. *IEEE Access* 8, 108896–108912. doi: 10.1109/ACCESS.2020.3001396

Gabbiani, F., and Krapp, H. G. (2006). Spike-frequency adaptation and intrinsic properties of an identified, looming-sensitive neuron. *J. Neurophysiol*. 96, 2951–2962. doi: 10.1152/jn.00075.2006

Geiger, A., Lenz, P., Stiller, C., and Urtasun, R. (2013). Vision meets robotics: the kitti dataset. *Int. J. Robot. Res*. 32, 1231–1237. doi: 10.1177/0278364913491297

Hartbauer, M. (2017). Simplified bionic solutions: a simple bio-inspired vehicle collision detection system. *Bioinspir. Biomimet*. 12:026007. doi: 10.1088/1748-3190/aa5993

Horridge, G. A., and Sandeman, D. (1964). Nervous control of optokinetic responses in the crab carcinus. *Proc. R. Soc. Lond. Ser. B Biol. Sci*. 161, 216–246. doi: 10.1098/rspb.1964.0091

Hu, C., Arvin, F., Xiong, C., and Yue, S. (2016). Bio-inspired embedded vision system for autonomous micro-robots: the LGMD case. *IEEE Trans. Cogn. Dev. Syst*. 9, 241–254. doi: 10.1109/TCDS.2016.2574624

Keil, M. S., and Rodriguez-Vazquez, A. (2003). "Toward a computational approach for collision avoidance with real-world scenes," in *Bioengineered and Bioinspired Systems, Vol. 5119* (Maspalomas: SPIE), 285–296. doi: 10.1117/12.499054

Kennedy, J. S. (1951). The migration of the desert locust (schistocerca gregaria forsk.). I. The behaviour of swarms. II. A theory of long-range migrations. *Philos. Trans. R. Soc. Lond. Se. B Biol. Sci*. 235, 163–290. doi: 10.1098/rstb.1951.0003

Lei, F., Peng, Z., Liu, M., Peng, J., Cutsuridis, V., and Yue, S. (2022). A robust visual system for looming cue detection against translating motion. *IEEE Trans. Neural Netw. Learn. Syst*. 1–15. doi: 10.1109/TNNLS.2022.3149832

Li, Z., Fu, Q., Li, H., Yue, S., and Peng, J. (2022). "Dynamic signal suppression increases the fidelity of looming perception against input variability," in *2022 International Joint Conference on Neural Networks (IJCNN)* (Padua: IEEE), 1–8. doi: 10.1109/IJCNN55064.2022.9892873

Luan, H., Fu, Q., Zhang, Y., Hua, M., Chen, S., and Yue, S. (2021). A looming spatial localization neural network inspired by MLG1 neurons in the crab neohelice. *Front. Neurosci*. 15:787256. doi: 10.3389/fnins.2021.787256

Luan, H., Hua, M., Peng, J., Yue, S., Chen, S., and Fu, Q. (2022). "Accelerating motion perception model mimics the visual neuronal ensemble of crab," in *2022 International Joint Conference on Neural Networks (IJCNN)* (Padua: IEEE), 1–8. doi: 10.1109/IJCNN55064.2022.9892540

Meng, H., Appiah, K., Yue, S., Hunter, A., Hobden, M., Priestley, N., et al. (2010). A modified model for the lobula giant movement detector and its Fpga implementation. *Comput. Vis. Image Understand*. 114, 1238–1247. doi: 10.1016/j.cviu.2010.03.017

Mukhtar, A., Xia, L., and Tang, T. B. (2015). Vehicle detection techniques for collision avoidance systems: a review. *IEEE Trans. Intell. Transp. Syst*. 16, 2318–2338. doi: 10.1109/TITS.2015.2409109

Nityananda, V., Bissianna, G., Tarawneh, G., and Read, J. (2016a). Small or far away? Size and distance perception in the praying mantis. *Philos. Trans. R. Soc. B Biol. Sci*. 371:20150262. doi: 10.1098/rstb.2015.0262

Nityananda, V., Tarawneh, G., Rosner, R., Nicolas, J., Crichton, S., and Read, J. (2016b). Insect stereopsis demonstrated using a 3d insect cinema. *Sci. Rep*. 6:18718. doi: 10.1038/srep18718

O'shea, M., and Rowell, C. (1976). The neuronal basis of a sensory analyser, the ACRIDID movement detector system. II. Response decrement, convergence, and the nature of the excitatory afferents to the fan-like dendrites of the LGMD. *J. Exp. Biol*. 65, 289–308. doi: 10.1242/jeb.65.2.289

O'Shea, M., and Williams, J. (1974). The anatomy and output connection of a locust visual interneurone; the lobular giant movement detector (LGMD) neurone. *J. Compar. Physiol*. 91, 257–266. doi: 10.1007/BF00698057

Parker, A. J. (2007). Binocular depth perception and the cerebral cortex. *Nat. Rev. Neurosci*. 8, 379–391. doi: 10.1038/nrn2131

Poiesi, F., and Cavallaro, A. (2016). "Detection of fast incoming objects with a moving camera," in *BMVC* (York: BMVA Press). doi: 10.5244/C.30.146

Rind, F. C., and Bramwell, D. (1996). Neural network based on the input organization of an identified neuron signaling impending collision. *J. Neurophysiol*. 75, 967–985. doi: 10.1152/jn.1996.75.3.967

Rind, F. C., Wernitznig, S., Pölt, P., Zankel, A., Gütl, D., Sztarker, J., et al. (2016). Two identified looming detectors in the locust: ubiquitous lateral connections among their inputs contribute to selective responses to looming objects. *Sci. Rep*. 6, 1–16. doi: 10.1038/srep35525

Rosner, R., Tarawneh, G., Lukyanova, V., and Read, J. C. (2020). Binocular responsiveness of projection neurons of the praying mantis optic lobe in the frontal visual field. *J. Compar. Physiol. A* 206, 165–181. doi: 10.1007/s00359-020-01405-x

Rosner, R., von Hadeln, J., Tarawneh, G., and Read, J. C. (2019). A neuronal correlate of insect stereopsis. *Nat. commun*. 10:2845. doi: 10.1038/s41467-019-10721-z

Rossel, S. (1983). Binocular stereopsis in an insect. *Nature* 302, 821–822. doi: 10.1038/302821a0

Rossel, S. (1986). Binocular spatial localization in the praying mantis. *J. Exp. Biol*. 120, 265–281. doi: 10.1242/jeb.120.1.265

Salt, L., Howard, D., Indiveri, G., and Sandamirskaya, Y. (2019). Parameter optimization and learning in a spiking neural network for UAV obstacle avoidance targeting neuromorphic processors. *IEEE Trans. Neural Netw. Learn. Syst*. 31, 3305–3318. doi: 10.1109/TNNLS.2019.2941506

Salt, L., Indiveri, G., and Sandamirskaya, Y. (2017). "Obstacle avoidance with LGMD neuron: towards a neuromorphic UAV implementation," in *2017 IEEE International Symposium on Circuits and Systems (ISCAS)* (Baltimore, MD: IEEE), 1–4. doi: 10.1109/ISCAS.2017.8050976

Scarano, F., Sztarker, J., Medan, V., de Astrada, M. B., and Tomsic, D. (2018). Binocular neuronal processing of object motion in an arthropod. *J. Neurosci*. 38, 6933–6948. doi: 10.1523/JNEUROSCI.3641-17.2018

Serres, J. R., and Ruffier, F. (2017). Optic flow-based collision-free strategies: from insects to robots. *Arthropod Struct. Dev*. 46, 703–717. doi: 10.1016/j.asd.2017.06.003

Sun, X., Jiang, Y., Ji, Y., Fu, W., Yan, S., Chen, Q., et al. (2019). "Distance measurement system based on binocular stereo vision," in *IOP Conference Series: Earth and Environmental Science, Vol. 252* (Beijing: IOP Publishing), 052051. doi: 10.1088/1755-1315/252/5/052051

Sztarker, J., and Rind, F. C. (2014). A look into the cockpit of the developing locust: looming detectors and predator avoidance. *Dev. Neurobiol*. 74, 1078–1095. doi: 10.1002/dneu.22184

Vienne, C., Plantier, J., Neveu, P., and Priot, A. E. (2018). (disparity-driven) accommodation response contributes to perceived depth. *Front. Neurosci*. 12:973. doi: 10.3389/fnins.2018.00973

Wang, Y., Li, H., Zheng, Y., and Peng, J. (2023). A directionally selective collision-sensing visual neural network based on fractional-order differential operator. *Front. Neurorobot*. 17:1149675. doi: 10.3389/fnbot.2023.1149675

Wernitznig, S., Rind, F. C., Pölt, P., Zankel, A., Pritz, E., Kolb, D., et al. (2015). Synaptic connections of first-stage visual neurons in the locust Schistocerca gregaria extend evolution of tetrad synapses back 200 million years. *J. Compar. Neurol*. 523, 298–312. doi: 10.1002/cne.23682

Xu, M., Wang, H., Chen, H., Li, H., and Peng, J. (2023). A fractional-order visual neural model for small target motion detection. *Neurocomputing* 2023:126459. doi: 10.1016/j.neucom.2023.126459

Yang, L., Wang, B., Zhang, R., Zhou, H., and Wang, R. (2017). Analysis on location accuracy for the binocular stereo vision system. *IEEE Photon. J*. 10, 1–16. doi: 10.1109/JPHOT.2017.2784958

Yue, S., and Rind, F. C. (2006). Collision detection in complex dynamic scenes using an lgmd-based visual neural network with feature enhancement. *IEEE Trans. Neural Netw*. 17, 705–716. doi: 10.1109/TNN.2006.873286

Yue, S., and Rind, F. C. (2009). "Near range path navigation using LGMD visual neural networks," in *2009 2nd IEEE International Conference on Computer Science and Information Technology* (Beijing: IEEE), 105–109.

Zhao, J., Ma, X., Fu, Q., Hu, C., and Yue, S. (2019). "An LGMD based competitive collision avoidance strategy for UAV," in *IFIP International Conference* *on Artificial Intelligence Applications and Innovations* (Hersonissos: Springer), 80–91. doi: 10.1007/978-3-030-19823-7_6

Zhao, J., Wang, H., Bellotto, N., Hu, C., Peng, J., and Yue, S. (2021). Enhancing LGMD's looming selectivity for uav with spatial-temporal distributed presynaptic connections. *IEEE Trans. Neural Netw*. Learn. Syst. 34, 2539–2553. doi: 10.1109/TNNLS.2021.3106946

Zhao, J., Hu, C., Zhang, C., Wang, Z., and Yue, S. (2018). "A bio-inspired collision detector for small quadcopter," in *2018 International Joint Conference on Neural Networks (IJCNN)* (Rio de Janeiro: IEEE), 1–7. doi: 10.1109/IJCNN.2018.8489298

Zhen, X., Shengyong, C., and Garrick, O. (2017). Event-based stereo depth estimation using belief propagation. *Front. Neurosci*. 11:535. doi: 10.3389/fnins.2017.00535

Zhu, Y., Dewell, R. B., Wang, H., and Gabbiani, F. (2018). Pre-synaptic muscarinic excitation enhances the discrimination of looming stimuli in a collision-detection neuron. *Cell Rep*. 23, 2365–2378. doi: 10.1016/j.celrep.2018.04.079

frontiers | Frontiers in Neuroscience

TYPE Methods
PUBLISHED 07 September 2023
DOI 10.3389/fnins.2023.1247315

Check for updates

OPEN ACCESS

EDITED BY
Zhengwang Wu,
University of North Carolina at Chapel Hill,
United States

REVIEWED BY
Xuanang Xu,
Rensselaer Polytechnic Institute, United States
Zhen Zhou,
University of Pennsylvania, United States

*CORRESPONDENCE
Hong Ji
✉ itsjihong@gmail.com

RECEIVED 25 June 2023
ACCEPTED 18 August 2023
PUBLISHED 07 September 2023

CITATION
Ji H, Zhang X, Chen B, Yuan Z, Zheng N and Keil A (2023) Groupwise structural sparsity for discriminative voxels identification.
Front. Neurosci. 17:1247315.
doi: 10.3389/fnins.2023.1247315

Groupwise structural sparsity for discriminative voxels identification

Hong Ji[1]*, Xiaowei Zhang[2], Badong Chen[2], Zejian Yuan[2], Nanning Zheng[2] and Andreas Keil[3]

[1]The Shaanxi Key Laboratory of Clothing Intelligence, School of Computer Science, Xi'an Polytechnic University, Xi'an, China, [2]Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, China, [3]Center for the Study of Emotion and Attention, Department of Psychology, University of Florida, Gainesville, FL, United States

This paper investigates the selection of voxels for functional Magnetic Resonance Imaging (fMRI) brain data. We aim to identify a comprehensive set of discriminative voxels associated with human learning when exposed to a neutral visual stimulus that predicts an aversive outcome. However, due to the nature of the unconditioned stimuli (typically a noxious stimulus), it is challenging to obtain sufficient sample sizes for psychological experiments, given the tolerability of the subjects and ethical considerations. We propose a stable hierarchical voting (SHV) mechanism based on stability selection to address this challenge. This mechanism enables us to evaluate the quality of spatial random sampling and minimizes the risk of false and missed detections. We assess the performance of the proposed algorithm using simulated and publicly available datasets. The experiments demonstrate that the regularization strategy choice significantly affects the results' interpretability. When applying our algorithm to our collected fMRI dataset, it successfully identifies sparse and closely related patterns across subjects and displays stable weight maps for three experimental phases under the fear conditioning paradigm. These findings strongly support the causal role of aversive conditioning in altering visual-cortical activity.

KEYWORDS

fMRI, groupwise regularization, voxel selection, stable hierarchical voting (SHV), randomized structural sparsity (RSS), effective vote ratio (EVR)

1. Introduction

Machine learning approaches have become popular in cognitive neuroscience, often in the context of using neuroimaging techniques to discriminate between brain patterns associated with different experimental conditions, emotional states, cognitive processes, and ultimately health outcomes. Variable selection and feature selection have become the focus of studies using brain-based data with tens or hundreds of thousands of variables. The objective of the research addressing this problem falls broadly into two categories: (1) brain image decoding, e.g., Haxby et al. (2001) and brain-computer interface (BCI) (Wolpaw et al., 2002; Saha et al., 2021), as well as (2) multivariate hypothesis testing (Bzdok et al., 2017; Kia et al., 2017; Wen et al., 2019) including identification of candidate biomarkers for medical diagnosis (Demirci et al., 2008). The former applications pursue the maximum predictive power of the predictors, providing faster and more cost-effective predictors, while the latter put more attention on providing a better understanding of the underlying process that reflects the spatiotemporal nature of the generated data. In the present study, we are interested in the second application, i.e., brain decoding. We specifically address the problem of identifying the brain activity patterns that are associated with specific behavior. The classic univariate analysis typically models each response channel separately, which is inconsistent with the multivariate nature of neuronal population codes and also with the observation that noise is spatially correlated. Separate modeling of each response entails low power for testing and comparing models, for two reasons: (1) Single fMRI responses may be noisy, and the evidence is not combined across locations. (2) The analyses treat the responses as independent, thus forgoing the benefit exploited by linear decoding approaches to model the noise in a multivariate manner. This is particularly important in fMRI data analysis, where nearby voxels have highly correlated noise. As spatial resolution increases, we face the combined challenge of increasing the number of individual voxels (inflating the feature space) and also increasing the noise in those individual voxels.

In order to understand the learning process of human in response to an initial neutral visual stimulus predicting an aversive outcome, we conducted a study using fMRI to observe the large-scale neurophysiological changes. In neuroimaging, a decoder is a predictive model that, given a series of brain images, fits the binary classification information regarding an experimental condition, a stimulus category, a motor behavior, or a clinical state. In the context of aversive conditioning, one of two initially harmless stimuli [referred to as conditioned stimuli (CS)] acquires motivational significance by consistently predicting the occurrence (CS+) of a negative event [known as the unconditioned stimulus (US)], while the other stimulus (CS−) predicts its absence. Since US is generally a noxious stimulus, it is difficult to obtain satisfactory sample sizes for such psychological experiments, given the tolerability of the subjects already ethical considerations. Therefore, we here focus on linear brain decoding because of its broader usage in analyzing inherently small sample size (Pereira et al., 2009). The estimated classification or regression weights can be visualized in the form of brain maps, which can aid in understanding how brain activity in space and time underlies a cognitive function (Mourao-Miranda et al., 2005). Selecting an appropriate set of voxels as the input for the classifier construction is of critical importance. The voxels corresponding to the non-zero weights are considered as the relevant features. The identification of discriminative voxels is based on the values of the weight vector, and their importance is proportional to the absolute values of the weights.

Due to the high-dimensionality of neuroimaging, high correlations among different voxels and low signal-to-noise ratios (SNRs), multiple weight maps yielding the same predictive power. In other words, different models lead to very similar generalization performance, and the recovered brain maps often suffer from lack of interpretability. Therefore, improving the interpretability of brain decoding approaches is of primary interest in many neuroimaging studies, especially in a group analysis of multi-subject data. At present, there are two main approaches proposed to enhance the interpretability of multivariate brain maps, as reviewed by Kia et al. (2017): (1) Introducing new metrics into the model selection procedure. (2) Introducing new hybrid penalty terms for regularization. The first approach to improving the interpretability looks for the best values for the hyper-parameters of a model (Lemm et al., 2011; Hoyos-Idrobo et al., 2018). The second approach involves applying regularization or prior knowledge (Zou and Hastie, 2005; Yuan and Lin, 2006; Rasmussen et al., 2012) to restrict model complexity, also known as dimension reduction. This approach is commonly used for the ill-posed nature of brain decoding problems (Geman et al., 1992).

As a representative of the second category, structured sparsity models (Chambolle, 2004; Bach et al., 2012; Micchelli et al., 2013) extend the least absolute shrinkage and selection operator (LASSO) model by promoting sparse models in some preferred way. For example, regression weights may be encouraged to be constant or vary smoothly within regions of the brain (Michel et al., 2011; Baldassarre et al., 2012; Gramfort et al., 2013). Despite the fact that sparsity has traditionally been connected with interpretability, these structured sparsity models incorporating additional spatial constraints into the predictive model, allowing for even greater ease of interpretation by further grouping the discriminative voxels into few clusters based on prior information (Yuan et al., 2011; Li et al., 2014; Shimizu et al., 2015). Besides, stability selection is applied as an effective way to control the false positives (Meinshausen and Bühlmann, 2010; Ye et al., 2012; Shah and Samworth, 2013; Cao et al., 2014; Rondina et al., 2014; Wang and Zheng, 2014). While the control of false positives can be achieved, a significant false negative rate is often expected, especially in the case of redundant and correlated voxels, this correlation prior is not explicitly taken into consideration. In Wang and Zheng (2014) and Wang et al. (2015) the authors proposed a "randomized structural sparsity", incorporating the idea of structural sparsity in the stability selection framework, together with the subsampling scheme which further help to refine and outline the exact shapes of the discriminative regions. These regions may not be the same size as the prior partitions, which is crucial for neighboring voxels belonging to the same brain area. Although they may be highly correlated, not all neighboring voxels are necessarily significant discriminative voxels (Witten et al., 2014). A similar strategy was used in Wan et al. (2014) and Yan et al. (2015) to predict cognitive outcomes via cortical surface measures. The results showed improved decoding accuracy and interpretability of brain maps.

In order to enhance the stability and reproducibility of our model during optimization, we apply group constraints and regularization across multiple subjects. This technique is commonly used in transfer learning or multitask learning (Bakker and Heskes, 2003; Raina et al., 2006; Dai et al., 2007; Pan and Yang, 2010). In our paper, we make the assumption that the regions of discriminative voxels are relevant or overlapping to a certain extent across subjects. Additionally, we assume that only a few clusters are actually discriminative for the classification problem. To achieve these goals, we propose to use a mixed $l1$ and groupwise $l2$ norm for regularization. The $l2$ norm penalizes large coefficients and yields a non-sparse weight distribution inside the group, while the $l1$ norm promotes sparsity on selected clusters. This nested mixed-norm regularization enables us to construct stable and interpretable models by pooling data from multiple subjects. It is important to note that the $l2$ norm does not imply the application of unified weights to the functionally significant clusters, which might be a too strong constraint and impractical for the real data.

Based on stability selection and the groupwise structural, we propose a stable hierarchical voting (SHV) mechanism to monitor the quality of spatial random sampling and reduce the risk of false and missed detections. When using uniform sampling, there is a possibility that many noisy and uninformative voxels will be included. To address this issue, we use multiple cross-validations of test accuracy during the voting process to select high-quality samples. In addition, small perturbations in the observations can cause instability in the model generated (Arlot et al., 2010). To mitigate this problem, we apply model averaging to aggregate the output of multiple models as suggested (Nemirovski, 2000). Furthermore, the number of selected candidate features is allowed to be much larger when incorporating group structure (Jenatton et al., 2011; Xiang et al., 2015), which allows us a more global search among brain regions.

2. Methods

2.1. Pre-segmentation

For the class of methods that use structural information for dimensionality reduction, the number of clusters to be generated is estimated based on finding a compromise between several factors: (1) To enhance area homogeneity, it tends to conduct fine segmentation for small patches. (2) To avoid the false negative selection due to spatial sparsity induced by the $l1$ norm, it tends to perform rough segmentation for large patches. (3) The number of trials is taken into consideration as the unknowns of the optimization problem is now equal to the number of clusters. From the previous study (Craddock et al., 2013), with 200 ROIs, the resulting parcellations consist of clusters with anatomic homology and thus offer increased interpretability.

In our work, we first obtain the structural information about the brain according to their strong local correlations. Here we perform a data-driven segmentation operation to partition the voxels into small clusters using the normalized cut (NCut) (Shi and Malik, 2000; Cour et al., 2005). To define the affinity between two voxels v_1 and v_2 we combine three cues: (1) the correlations of the raw BOLD time series, (2) the correlations of BOLD features for each trial, (3) a connection radius σ_d to attenuate the influence from far away voxels. Voxels in close proximity with similar BOLD waveforms are likely to be part of the same cluster. Additionally, incorporating correlations among features helps to minimize the impact of signal clutter. Furthermore, averaging the features results in a fit with lower variance compared to individual features, especially when they are positively correlated (Park et al., 2006; Wang et al., 2015). This aspect also contributes to the potential enhancement of stronger features.

The affinity matrix is computed based on finding the combined data from multiple subjects since uniform segmentation is required for group-wise regularization. Let us denote the preprocessed fMRI data matrix as $\tilde{\boldsymbol{X}} \in \mathbb{R}^{N_t \times N_V}$, where N_t is the number of scans, N_V is the number of voxels. To access the columns of a matrix, the v-th column is denoted as $(:, v)$. We construct the affinity matrix $\boldsymbol{A}$ as follows:

$$\boldsymbol{A}_{v_1,v_2} = |corr(\tilde{\boldsymbol{X}}(:, v_1),\ \tilde{\boldsymbol{X}}(:, v_2))| \cdot \exp(-dist(v_1, v_2)^2/\sigma_d^2)$$

where $|\cdot|$ gets the absolute value, $corr(\cdot, \cdot)$ is the correlation between two variables, and $dist(\cdot)$ evaluates the Euclidean distance of two voxels in 3D space.

2.2. Classification using groupwise structural sparsity

Let us denote the feature matrix from subject i as $\boldsymbol{X^i} \in \mathbb{R}^{N_T \times N_V}$, $i \in \{1...N_S\}$, where N_T is the number of trials, N_V is the number of voxels, and N_S is the number of subjects. For this study, we are interested in classifying the experimental conditions. We denote the binary labeling information as $\boldsymbol{y} \in \mathbb{R}^{N_T}$, $\boldsymbol{y}(t) \in \{1, -1\}$ that correspond to the CS+ and CS− conditioning, respectively. The stability sampling is performed in terms of the subsampling on the features, i.e., the columns of $\boldsymbol{X^i}$, as well as subsampling of the observations, i.e., the rows of $\boldsymbol{X^i}$. Then parceling information is used to average the features within a cluster. We denote the set of the clusters via the pre-segmentation as $\mathcal{G}$, and denote the number of clusters as N_C. Specifically, each cluster $g_j \in \mathcal{G}$, consists of highly correlated neighboring voxels, the sampled voxels lying in cluster j are noted as a set $g_j{}' \subset g_j \in \mathcal{G}$, for each chosen trial t, and $\boldsymbol{D}(t, j)$ is the corresponding average of $\boldsymbol{X}(t, g_j^{'})$ of cluster j. The model can be simplified to the following low dimensional problem.

$$F = \arg\min_{\mathbf{w}} \sum_{t=1}^{N_T} log\left(1 + exp\left(-\boldsymbol{y}(t)\left(\boldsymbol{D}(t, :)\mathbf{w} + b\right)\right)\right) + \lambda \sum_{j=1}^{N_C} \| \mathbf{w}(j) \| \tag{1}$$

where $\mathbf{w} \in \mathbb{R}^{N_C}$ is the weight vector. $\mathbf{w}(j)$ denotes the weight of j-th cluster, corresponding to the subset $g_j \in \mathcal{G}$. The voxels corresponding to weight with large absolute value are considered as discriminative voxels (Wang et al., 2015).

In this paper, we propose to consider a group of subjects together and constrain the model using a mixed l_1/l_2 norm. We combine the weight vectors from all subjects into a matrix $\boldsymbol{W} \in \mathbb{R}^{N_C \times N_S}$. Correspondingly, the objective of the model is below:

$$F = \arg\min_{\boldsymbol{W}} \sum_{i=1}^{N_S} \sum_{t=1}^{N_T} log\left(1 + exp\left(-\boldsymbol{y}^i(t)\left(\boldsymbol{D}^i(t, :)\boldsymbol{W}(:, i) + b^i\right)\right)\right) + \lambda \sum_{j=1}^{N_C} \| \boldsymbol{W}(j, :) \| \tag{2}$$

As shown in Figure 1, the $l2$ norm over multiple subjects for each cluster is proposed as a group constraint, i.e., the rows of $\boldsymbol{W}$ shown in the red box of Figure 1B, while the $l1$ norm on clusters further enforces structural sparsity on the solution. Using the mixed $l1$ and $l2$ norm as a joint optimization criterion allows the pooling of data from multiple subjects and enforces consistency of the selection of clusters across subjects. For the convenience of optimization, the weight matrix is vectorized, and the individual feature matrix and the label information are integrated from all subjects accordingly.

Note that the number of clusters obtained is typically much smaller than the number of voxels (N_V) and comparable to the total number of total samples. By reducing the number of unknowns and

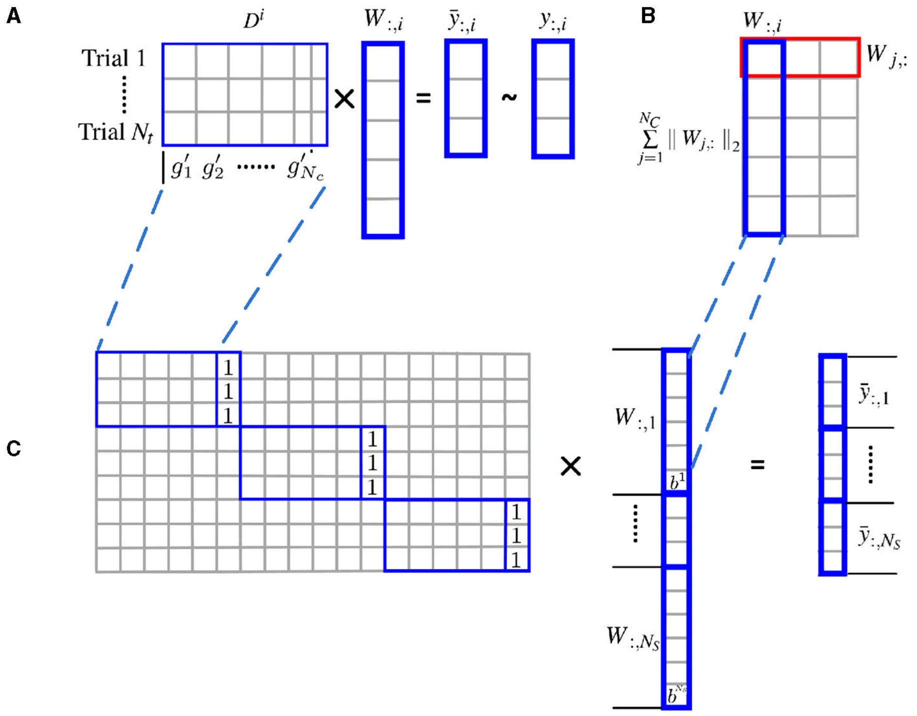


FIGURE 1
Estimate cluster weights using joint optimization of multiple subjects with $l2$ norm group constraint. **(A)** Structural sparsity model of single subject; **(B)** groupwise structural constraint using mixed l_1/l_2 norm; **(C)** for the convenience of optimization, the weight matrix $\boldsymbol{W}$ is vectorized, and the individual feature matrix $\boldsymbol{D}^i$ are incorporated to form a block diagonal matrix with an additional column of all 1. The label information is merged from all subjects accordingly.

integrating data from multiple subjects, we are able to use fewer samples to estimate the parameters.

2.3. Algorithmic framework

Unlike the general stability selection framework (Meinshausen and Bühlmann, 2010; Shah and Samworth, 2013; Wang et al., 2015), our algorithm, stable hierarchical voting (SHV), represents a step further with stricter control for model variance among subjects. The detailed description is outlined in Algorithm 1. Based on stability selection and the groupwise structural constraint, SHV employs a stable hierarchical voting mechanism to monitor the sample quality of spatial random sampling and reduce the risk of false and missed detections. The proposed method utilizes a two-level nested loop approach to construct a predictive decoding model for multi-subject data, while considering mixed regularity constraints. The outer loop randomly samples voxels and performs dimensionality reduction feature expressions on the corresponding motifs; The inner loop assesses the predictive ability of these features, by computing the average prediction correctness through cross-validation. Subsequently, the outer loop performs cumulative voting on the selected voxel samples, based on their prediction ability as evaluated by the inner loop. This structuring guarantees that only votes with high test precision are considered.

In the following, i denotes the subject index, $i = \{1, 2, \cdots, N_S\}$, j denotes the cluster index, $j = \{1, 2, \cdots, N_C\}$, and m denotes the voxel index, $m = \{1, 2, \cdots, N_V\}$. For the outside layer, we perform constrained block subsampling in terms of voxels (columns) and calculate the averaged feature matrix, the number of resamplings denotes as N_K. Let the subsampling fraction be $\alpha_{col} \in (0, 1)$ and $\mathcal{I}$ denotes the set of voxel indices randomly picked.

To avoid instabilities of the generated model caused by perturbations of the observed data, we apply model averaging to mitigate this problem (Nemirovski, 2000; Arlot et al., 2010). For loop k, the weight vector for lth cross-verification is denoted as $\boldsymbol{W}_l(:, i)$, the score vector $\boldsymbol{s}_{i,k}$ is calculated by the following equation

$$\boldsymbol{s}_{i,k}(j) = \frac{1}{N_L} \sum_{l=1}^{N_L} |\boldsymbol{W}_l(j, i)|, \quad j = \{1, 2, \cdots, N_C\} \tag{3}$$

where $| \cdot |$ get the absolute value, and N_L denotes the number of cross-verification, which is usually chosen according to the sample size and balance with the computation cost.

We hierarchically define the selectors, from cluster to voxel, respectively. Let $\pi(*, N_{sel})$ be the operation to select the top N_{sel} non-zero coefficients from vector $*$, and return the selector by marking the selected components to be unit valued (zero valued for the non-selected ones). If the actual non-zero components is less than N_{sel}, less components are selected.

Because uniform sampling is likely to include many noisy and uninformative voxels, for N_K times of spatial resampling, we only count N_i loops when the test accuracy of cross verification go above the sampling quality control factor q. The number of selected loops is determined based on a quality control ratio $\alpha_K \in (0, 1)$, only

```
Require:
    Dataset of subject $i$: $\boldsymbol{X^i} \in \mathbb{R}^{N_T \times N_V}$, $i \in \{1..N_S\}$;
    Label information $\boldsymbol{y} \in \mathbb{R}^{N_T}$, where $N_T$ is the number
    of trials, $N_V$ is
    the number of voxels;
    Predefined parcellation $\mathcal{G}$;
    Groupwise sparsity penalization parameter $\lambda$;
    Loops of spatial randomizations $N_K$; Loops of
    cross verification $N_L$;
    Subsampling ratio $\alpha_{row}, \alpha_{col}$ in terms of rows and
    columns of $\boldsymbol{X}$;
    Minimum acceptable precision $p$; Sampling
    quaility control ratio $\beta$;
    The number of clusters one wish to select $N_{sel}$;
Ensure:
    Effective vote ratio (EVR) for each voxel.

for $k = 1$ to $N_K$ do
  for $l = 1$ to $N_L$ do
    for $i = 1$ to $N_S$ do
      Perform subsampling on voxels (columns of
      $\boldsymbol{X^i}$) and calculate the averaged feature
      matrix: $\boldsymbol{D^i} \leftarrow \boldsymbol{X^i_{[:,\mathcal{I}]}} \leftarrow \boldsymbol{X^i}$, where $\mathcal{I} \subset$
      $\{1, 2, \cdots, N_V\}$, $\boldsymbol{D^i} \in \mathbb{R}^{N_T \times N_C}$.
      Perform subsampling on trials (rows of $\boldsymbol{X^i}$):
      $\boldsymbol{D^i_{[\mathcal{J},:]}} \leftarrow \boldsymbol{D^i}$ and update $\boldsymbol{y_{[\mathcal{J}]}} \leftarrow \boldsymbol{y}$, $\mathcal{J} \subset$
      $\{1, 2, \cdots, N_T\}$.
    end for
    Estimate $\boldsymbol{W}$ with Equation 2.
  end for
  for $i = 1$ to $N_S$ do
    Calculate the average test accuracy $R^{test}_{i,k}$
    across all the cross-verification loops.
  end for
end for

for $i = 1$ to $N_S$ do
  Select $N_i$ well sampled loops out of $N_K$ loops
  according to $R^{test}_{i,k}$
  for $k = 1$ to $N_i$ do
    Compute the score vector $\boldsymbol{s_{i,k}}$ with Equation 3.
    Select the $N_{sel}$ clusters with highest
    coefficients in $\boldsymbol{s_{i,k}}$.
  end for
end for
Compute the effective vote ratio $\boldsymbol{\phi}^V_i$ according to
Equation 6.
```

Algorithm 1. The algorithm framework of groupwise structural sparsity for discriminative voxel identification.

the top $[\alpha_K N_K]$ loops with the highest test accuracy are taken into consideration.

For group-level statistical inference, we compute the cluster-wise voting rates $\boldsymbol{\phi}^C$ that incorporate the votes from multiple subjects

$$\boldsymbol{\phi}^C = \frac{1}{N_S}\sum_{i=1}^{N_S}\pi\Big(\frac{1}{N_i}\sum_{k=1}^{N_i}\pi(\boldsymbol{s}_{i,k}, N_{sel}), N_{sel}\Big) \tag{4}$$

We accumulate the votes of all the qualified selectors and then normalize the value with the sampling times of the voxel. Given that a sampled voxel m that belongs to cluster j, the voting rate of $\boldsymbol{\phi}^V_i$ is defined as

$$\tilde{\boldsymbol{\phi}}^V_i(m) = \frac{\sum_{k=1}^{N_i}\delta(m \in \mathcal{I}_k \,\&\, \pi(\boldsymbol{s}_{i,k}, N_{sel})(j) == 1)}{\sum_{k=1}^{N_i}\delta(m \in \mathcal{I}_k)}, \quad m \in g_j \tag{5}$$

To ensure the stability and reliability of voting, the effective vote ratio (EVR) is defined as

$$\boldsymbol{\phi}^V_i(m) = \tilde{\boldsymbol{\phi}}^V_i(m) \cdot \boldsymbol{\phi}^C(j), \quad m \in g_j \tag{6}$$

We chose the regularization parameter λ in Equation (2) that maximize the averaged prediction accuracy below.

$$\bar{R} = \frac{1}{N_S}\sum_{i=1}^{N_S}\Big(\frac{1}{N_i}\sum_{k=1}^{N_i}R^{test}_{i,k}\Big) \tag{7}$$

2.4. Stability evaluation

We adopt the stability index defined by work Baldassarre et al. (2017) to evaluate the stability of our results on real fMRI across multiple subjects. The voxels selected by EVR for subject i are denoted as $S_i = \{m|\boldsymbol{\phi}^V_i(m) \neq 0\}$. Consider two sets of selected voxels, namely S_1 and S_2. The corrected pairwise relative overlap is calculated using the formula:

$$O(S_1, S_2) = \frac{||S_1 \cap S_2| - |S_1| * |S_2|/N_V|}{\max(|S_1|, |S_2|)} \tag{8}$$

Here, $|S_1 \cap S_2|$ is the number of voxels that are present in both sets, while $|S_1| * |S_2|/N_V$ represents the expected number of overlapping voxels between two random samples of size $|S_1|$ and $|S_2|$ respectively, where N_V is the total number of voxels. The average pairwise overlap $\overline{O}$ is obtained by taking the average of the relative overlap values of all pairs of subjects.

3. Results

3.1. Synthetic data

To test and analyze the proposed algorithm on a similar problem scale as the real fMRI data, we work on a 53 × 63 × 52 brain image that has 173,628 voxels of interest. Specifically for small-sample fMRI data, we assume only 40 training 20 CS+ trials and 20 CS− trials since fMRI datasets of this size are most commonly found in psychological paradigm validation sessions. For the simulations, we use the Automated Anatomical Labeling (AAL) atlas template that segments the brain into 116 anatomical regions (Tzourio-Mazoyer et al., 2002), commonly used for different types of functional and anatomical analysis of neuroimaging data. To test whether our algorithm has superior discriminative power, we assume that there is a linear combination of a portion of voxels with categorization ability in three brain regions that have some overlap in different individuals. Specifically, all subjects were assumed to have a functional network of three distributed discriminative brain regions $G_1 = \{32, 44, 62\}$, comprising three brain regions in the frontal, parietal and occipital lobes, each including over 300 discriminative voxels. Considering the complexity of the brain functional network and dramatic individual differences among subjects, we define 15 interference regions for each individual, and the interfering brain regions were not exactly the same for different individuals. For subject i, we define individual interference region set $G_0^i = \{t \mid (72 + i \times 3) \leq t \leq (86 + i \times 3)\}$, which are all continuous sets with 15 and three regions skipped between two sets. Each region contains roughly 300 voxels.

The base value of elements M_j^i in both discriminative regions and interference regions are generated from the standard uniform distribution $U(0, 1)$, where $j = 1, 2, \ldots, 116$ representing the index of regions, other voxels in the brain image are noise generated by a standard Gaussian distribution. For discriminative regions G_1 we simulate a spatially distributed pattern constrained by linear model $y_1^i = \sum_{j \in G_1} \tilde{W}_j^i \cdot M_j^i$, and samples of CS+ fall in the top 40% and CS- fall in the bottom 40% of the overall distribution of y_1, therefore the simulated data can be distinguished by the linear classifier. The weight $\tilde{W}_j^i$ is scaled by a personalized factor α_j^i that allows different connectivity strength $\tilde{W}_j^i = W_j^{init} \cdot \alpha_j^i$, where $W_{G_1}^{init} = \{1, 1, -2\}$ and $\alpha_j^i \sim U(0.5, 1.5)$ that uniformly distributed with minimum 0.5 and maximum 1.5. For interference regions G_0 we simulate $y_0^i = M_j^i$ and samples of CS+ fall in the top 80% and CS$-$ fall in the bottom 80% of the overall distribution. At last, gaussian noise is added to generate observations for single trials and single voxels $x_{t,v}^i = y_j^i + \epsilon_{t,v}, \epsilon_{t,v} \sim N(0, 1)$, where t denotes the index of trials and v the index of voxels.

The elements in discriminative and interference regions are both random samples from the uniform distribution; therefore, a single region should have no significant correlation with labels in absence of noise. On the contrary, the linear combination of regions in G_1 is discriminative, whereas for G_0, it is not. It is noticeable that although the discriminative areas are common for all subjects, the coefficients vary for each subject. Intentionally, we added noise to simulate the case that the interference regions may have an equal or even stronger degree of correlation by chance, which would result in false positives. Such simulation is crucial, especially for studies with few samples. In the following, we conducted several experiments on the synthetic data to examine the performance of the proposed algorithm.

3.2. Ablation study

For the ablation study, we compare experimental results with and without applying the proposed multi-subject $l2$ norm group constraint and test the effect of the algorithm on the choice of hyper parameters, including the effect of choosing different λ and N_{sel} on the results for selected discriminative clusters. In the following, we use the following notation:

- Our proposed method: estimate cluster weight using joint optimization of multiple subjects with the proposed Algorithm 1 and Equation (2);
- Alternative method: perform the same procedure of constraint block sampling and in terms of voxels and sub-sampling cross verification in terms of trials, then estimate cluster weight for each subject separately using Equation (1);

For the block bootstrap sampling methods, block size might affect the performance of the algorithm (Lahiri et al., 1999). Given the number of blocks, there are inherent trade-offs in the choice of block size. When only minimal loops of randomizations are allowed, the choice of large blocks is likely not matching the indeed supported geometry and are prone to many false positives, while the choice of small blocks may result in many false negatives due to ignorance of local correlation of adjacent voxels (Wang et al., 2015). Empirically we chose 3 × 3 × 3.

We accumulate one vote for the identified discriminative clusters corresponding to the top four weights with largest magnitude, then summing up all the votes across subjects. Although regularization helps to reduce model variance and larger regularization parameter (λ in Equations 1, 2) yields models with more degree of sparsity and fewer sets of selected variables (clusters), we tested how λ influence the outcome of selected discriminative clusters in both alternative method and our proposed method. Please note that the proposed method and the comparison method have different objective functions. Therefore, we employ two sets of lambda values, each consisting of one larger lambda and one smaller lambda. This is intended to showcase the influence of Lambda tuning on the outcomes.

The cluster scores reported in Figures 2A, C, E, G are averaged from 200 spatial subsampling steps each of which has 20 times cross validation, and the corresponding voting results are reported in Figures 2B, D, F, H. In Figure 2A we can see that for the alternative method, numerous interference clusters get higher scores than the true discriminative clusters. Larger λ, as shown in Figure 2C, helps to reduce false positives, however also increases false negatives. For the corresponding votes there is no single thresholding to distinguish discriminative clusters from the interference clusters, as can be seen in Figures 2B, D. For our proposed method, in Figure 2E as we can see from the enlarged view, scores estimated for discriminative cluster 44 are more consistent across subjects compare to the alternative method in Figure 2A, and the scores for interference clusters are relatively more sparse. As the λ increases, the score of the interference regions attenuated more significantly than the discriminative regions, as depicted in Figure 2G. Meanwhile, as shown in Figures 2F, H, there exist proper thresholds to separate all the three discriminative clusters correctly, and sparsity helps to increase the classification gap between the two.

For the synthetic data, we directly use the precision and recall curve since we know where the true discriminative features are. Precision (also called positive predictive value) is the fraction of discriminative clusters among the retrieved clusters, while recall (also known as sensitivity) is the fraction of discriminative clusters that have been retrieved over the total discriminative clusters. As shown in Figures 3A, B, when the same number of clusters is selected, our proposed method achieves both higher recall and precision score compare to the alternative approach (area under the two curves). Notice that when four clusters are selected

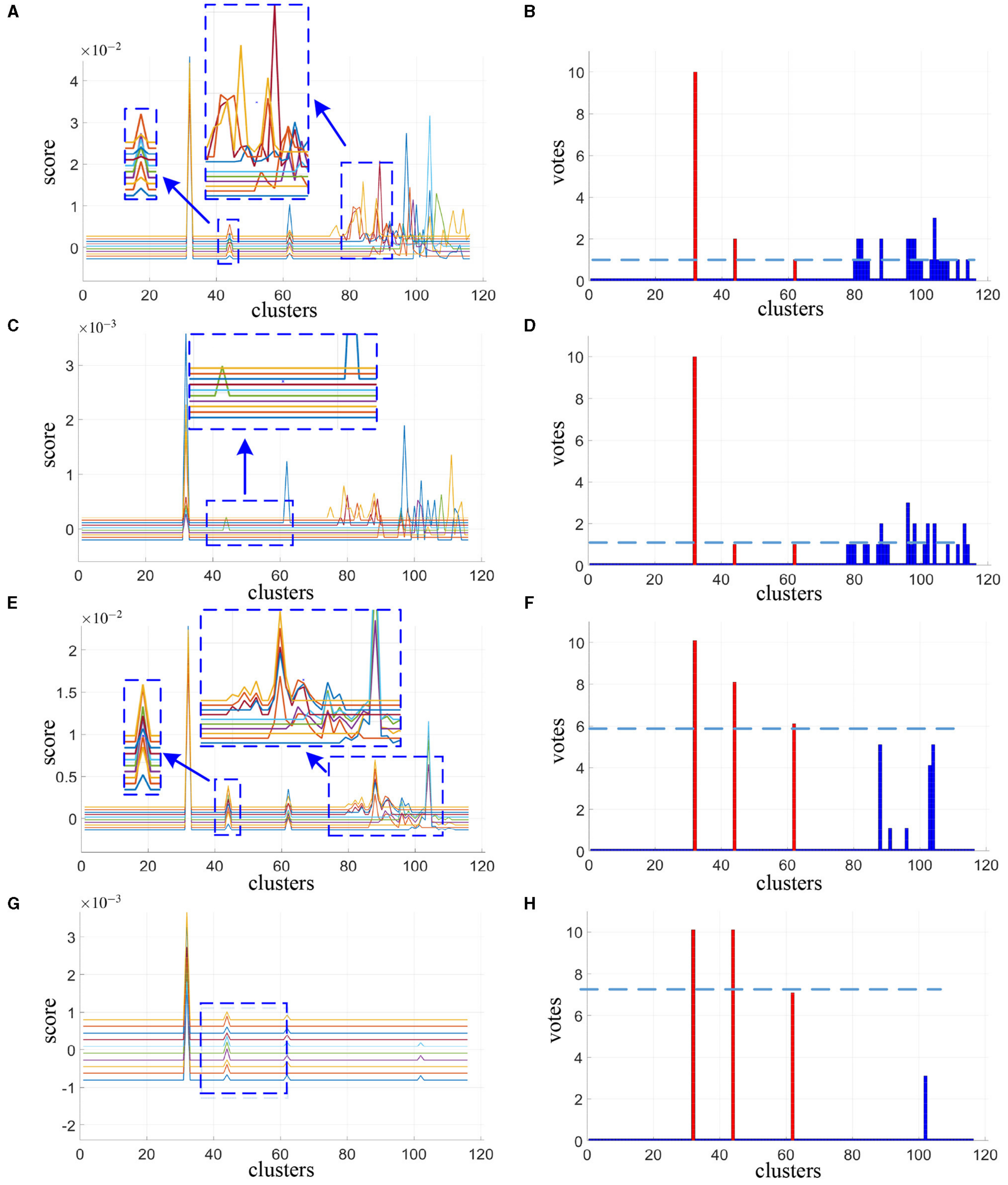


FIGURE 2
Cluster scores and vote results as estimated by alternative method at λ = 30 **(A, B)** and λ = 120 **(C, D)** and our proposed method at λ = 10 **(E, F)** and λ = 30 **(G, H)**. For cluster scores **(A, C, E, G)**, the blue arrows indicate the enlarged view of the original image, each colored line represents the result of one subject. For vote results **(B, D, F, H)**, the red lines indicate the true discriminative clusters, and the blue lines indicate the interference clusters.

(N_{sel} = 4), all the three true discriminative clusters can be detected. When increasing the number of selected clusters, our proposed method still maintained a high recall rate, while the alternative method does not seem to improve. Even when the number of clusters set to seven, the recall rate drops instead. In contrast to the alternative approach, our method is more likely to

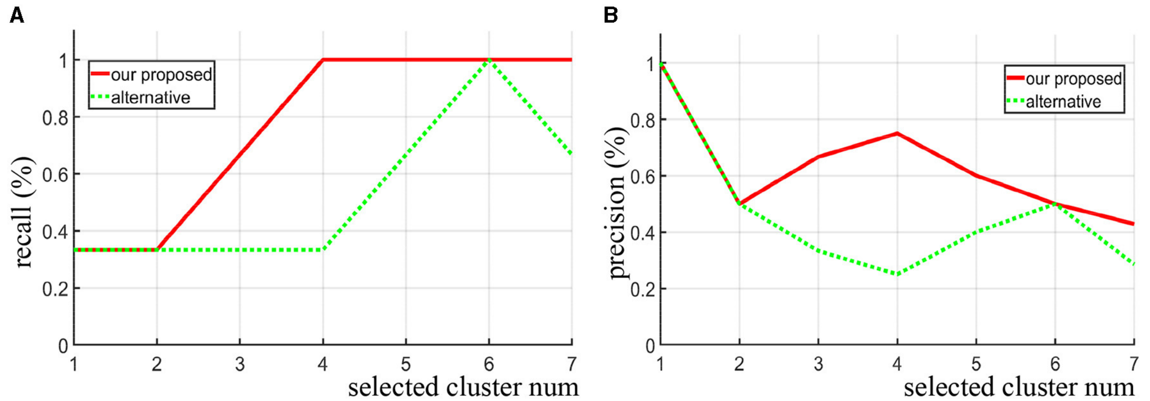


FIGURE 3
Given different selected cluster number N_{sel}, the recall curve **(A)** and precision curve **(B)** of our algorithm and the alternative method are compared on the synthetic data.

detect the real discriminative regions as increasing the number of selected clusters.

3.3. Real fMRI data I—Haxby dataset

Based on the simulation experiments, we use a well-established public dataset, Haxby, a study of face and object representation in human ventral temporal cortex (Haxby et al., 2001). The work innovatively incorporates the idea of structured sparsity into the framework of stability selection (randomized structure sparsity, RSS in short). The author compared their results with a range of classical univariate voxel selection methods and multi-voxel pattern identification methods, which showed relatively fewer false positives and confirmed the validity (higher predictive accuracy) of selected voxels. These methods include T-test, $l2$-SVM, $l2$ Logistic Regression, $l1$-SVM, $l1$ Logistic Regression, randomized $l1$ logistic regression, Smooth Lasso (Hebiri and Van de Geer, 2011) and TV-L1 (Gramfort et al., 2013) and Randomized Ward Logistic (Gramfort et al., 2012).

The Haxby dataset consists of six subjects with 12 runs per subject (dataset can be downloaded at http://data.pymvpa.org/datasets/haxby2001/). In each run, the subjects passively viewed grayscale images of eight object categories, grouped in 24s blocks separated by rest periods. Each image was shown for 500 ms and was followed by a 1,500 ms inter-stimulus interval. Full-brain fMRI data were recorded with a volume repetition time of 2.5 s. Then a stimulus block was covered by roughly nine volumes. For a complete description of the experimental design, fMRI acquisition parameters, and previously obtained results, check the reference on their website (Haxby et al., 2001; Hanson et al., 2004). In this paper, we use the fMRI data of subjects one to five and classifying the "House" and "Cat", which is a classic case for animal vs. non-animal classification. Preprocessing of the data consisted of motion correction using SPM 12, normalization and registration to the Montreal Neurological Institute (MNI) to facilitate inter-subject segmentation, removal of linear trends in each session, etc. There is no smoothing operation on the data. In the process of coregistration, the structural data is coregistered with functional data. Due to the missing of structural data, subject six is excluded from the analysis.

To have a fair comparison, we use the same parameter settings for RSS and our method: In particular, the number of clusters $N_C = 200$, the connection radius $\sigma_d = 3$, the block size $3 \times 3 \times 3$, the times of spatial randomization iterations $N_K = 200$, subsampling fraction $\alpha_{col} = 0.01$, fixed regularization parameter $\lambda = 0.3$. Several additional parameter is used in our approach for cross verification $N_{CV} = 20, \alpha_{row} = 0.9$ and sampling quality control $\alpha_K = 0.3$, $N_{sel} = 15$ is chosen for this study. This study was not interested in the activities of the cerebellum and vermis regions, therefore these regions were masked to rule out for consideration.

First, we compare the performance of our proposed method and RSS when decreasing the number of training samples. We use the first T sessions for training, which correspond to 1/2, 1/3, 1/4, and 1/6 of the data ($T = 6, 4, 3, 2$) for each subject. In Figure 4, we show the EVR maps from our method (a1–a4, not thresholded), and binominal test results of score maps across subjects (b1–b4, thresholded at 0.5). It shows that our proposed algorithm locates stable discriminative voxels at bilateral fusiform and inferior temporo-occipital even with fewer training samples (see the pattern in a3 and a4).

To evaluate the quality of the identified discriminative voxels, we conducted 4-fold cross validation using a linear l2-SVM classifier for both our proposed method and RSS. Figure 5 illustrates the changes in training and testing accuracy as the number of voxels increases. The reported curves are averaged across subjects and four times cross verification. Our method allowed for early identification of discriminative voxels. However, as more voxels were included (since the exact number of discriminative voxels is unknown), there was an increase in irrelevant voxels and noise. This led to a decline in the accuracy curve. On the other hand, the alternative method did not effectively identify discriminative voxels. With an increasing number of voxels, both irrelevant and truly relevant voxels were included,

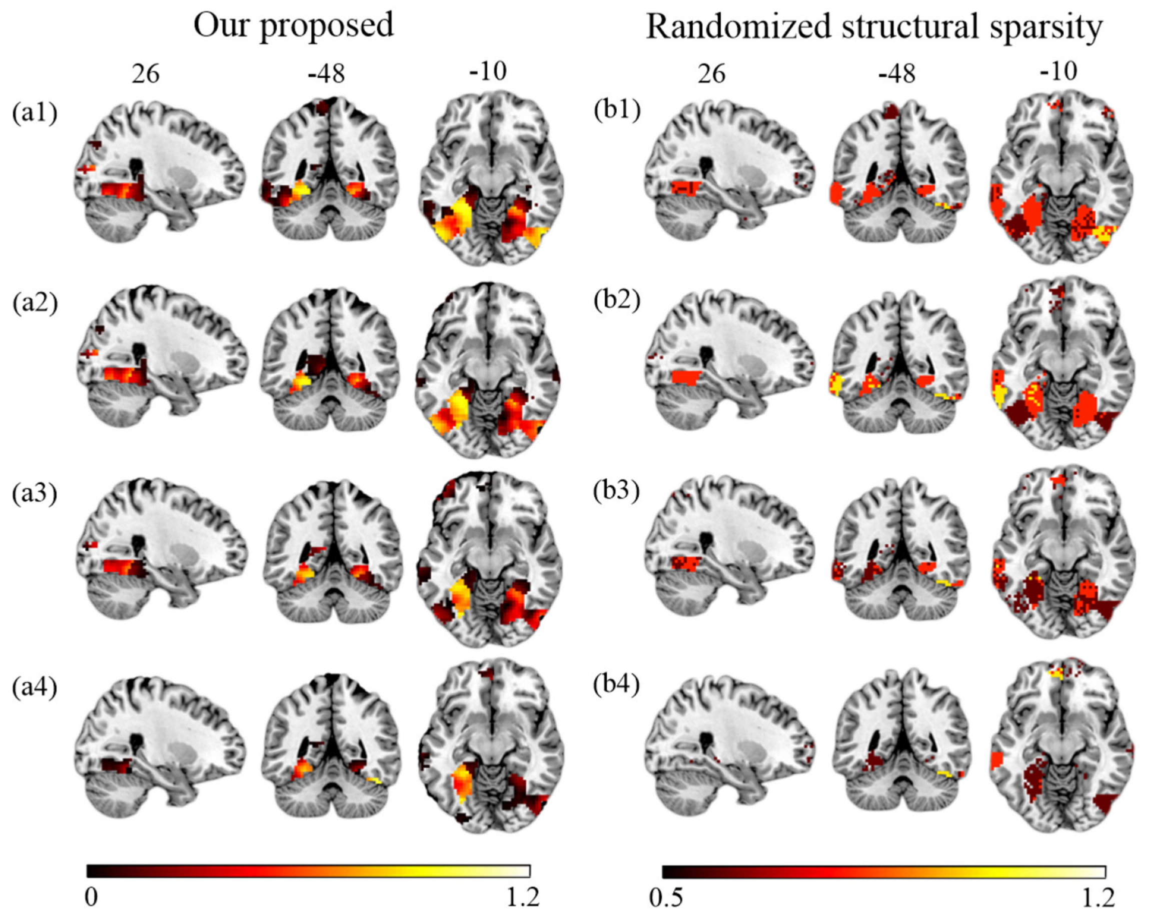


FIGURE 4
Brain maps for discriminative voxels as estimated on Haxby data (Cat vs. House). **(Left)** EVR maps (unthresholded) by our proposed approach. **(Right)** Maps of binominal test result for RSS, thresholded at 0.5. Both approaches used exactly the same amount of data for comparison (1) six sessions (the first 1/2) of five subjects; (2) four sessions (the first 1/3) of five subjects; (3) three sessions (the first 1/4) of five subjects; (4) two sessions (the first 1/6) of five subjects.

resulting in a flat curve. It is important to note that our method consistently outperforms the comparison method, as our curve consistently remains higher than the RSS curve.

3.4. Real fMRI data II—Fear conditioned dataset

After conducting experiments on synthetic data and commonly used public datasets, we initially tested and validated the robustness and sensitivity of the parameters of the proposed method. In general, our proposed approach outperforms the alternative approach in terms of its strength in recovering the discriminative pattern reliably when reducing the number of training samples, as well as keeping the sensitivity of individual specificity. Further, we exploratively conduct experiments on an earlier fMRI small sample dataset and then visualize the results. The data were recorded from a differential aversive conditioning study in which Gabors of one orientation were occasionally paired with an electric shock (see Petro et al., 2017; Ji et al., 2019, for details). For the habituation block, participants were instructed that they would not feel any shock but to fixate on the patterns. During the acquisition block, participants were informed that they would intermittently feel a cutaneous electric shock during the experiment but were not instructed as to the contingencies of the shock administration. The extinction phase was also uninstructed, such that participants were not told that no more shocks were to be given. The data reported here include 40 total trials per phase per participant. Each trial consisted of one of the two gratings being presented for $5,100ms$, during which its phase was alternated every 100 ms. An inter-trial interval (ITI) consisted of an initial gray cross (37.5 cd/m^2; 1° of visual angle) presented in the middle of the screen for a random duration between $0 - 8$ s followed by a white cross (149.0 cd/m^2) for a duration of 3 s, immediately preceding trial onset with Gabor patch presentation.

The Data were acquired during gradient-echo echo-planar imaging sequence with a 3T Philips Achieva scanner [echo time (TE), 30 ms; repetition Time (TR), 1.98 s; flip angle, 80°; slice number, 36; field of view, 224 mm; voxel size, $3.5 \times 3.5 \times 3.5$ mm^3; matrix size 64×64]. Preprocessing of BOLD fMRI data was completed using SPM12. We followed the standard preprocessing routines: slice timing correction, head movements realigning,

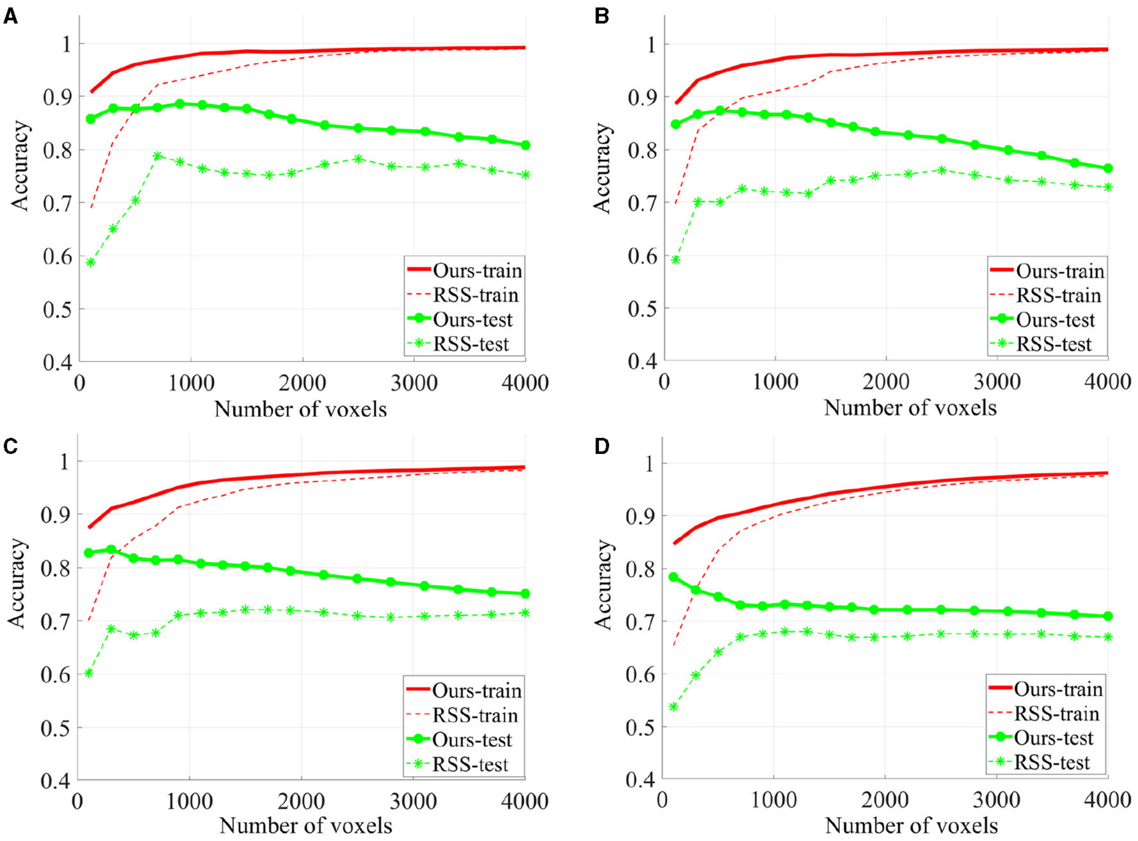


FIGURE 5
The classification accuracy based on 4-fold cross verification on House & Cat each curve is estimated on each individual and then averaged across folds and subjects. Six sessions **(A)**, four sessions **(B)**, three sessions **(C)**, and two sessions **(D)** are used for training.

normalization and resampled to a spatial resolution of 3 × 3 × 3 mm^3. Images were smoothed using a Gaussian kernel with a full-width at half-maximum of 6 mm. Low-frequency temporal drifts were removed from the BOLD data using a 1/128 Hz high-pass filter.

Following our previous work (Petro et al., 2017), the general linear models (GLMs) were constructed to extract features. The GLM aimed to model the ssVEP-BOLD coupling over the entire experiment. Thus, all trials were modeled separately using a GLM, which consisted of a sequence of boxcar functions in which the start was synchronized with the onset of each stimulus and width equal to the duration of each trial. Each boxcar function was then convolved with a canonical hemodynamic response function. Six additional regressors describing participants' head movements, as determined during preprocessing, were added to this design matrix to account for head movements during the scanning process. Excluding the motion components from the coefficient matrix, the single-trial coefficients are next used as features for decoding.

For the SHV scheme, the number of selected clusters is crucial, and as the number of of N_{sel} increases, the random overlap of clusters also increases. If N_{sel} is too large, it will reduce the sensitivity of the cluster voting rate and EVR. However, if N_{sel} is too small, it will result in more false negatives. We recommend selecting this parameter based on prior knowledge. In this study, we choose $N_{sel} = 40$ based on the previous analysis of EEG-ssVEP (Ji et al., 2018, 2019). Segmentation was performed based on the homogeneity of functional time series and feature correlations, as described in Section 2.1. Since this study did not interested in the activities of the cerebellum and vermis regions, these regions were masked out (AAL template 91-116). For the current data set, we select 200 for N_C and set the connection radius σ_d as 3 voxels. The results are reported in Figure 6. Although prediction accuracy may not be the sole criteria for selecting a model, it generally indicates that some of these voxels are truly discriminative when the prediction accuracy is high. To evaluate the quality of the discovered discriminative voxels, we employed a linear l2-SVM classifier (Hebiri and Van de Geer, 2011). Although not required, for all three experimental sessions, we pre-saved random seeds for block subsampling and cross-validation to ensure the same settings were made for all subjects to facilitate comparison. We set the times of spatial randomization iterations $N_K = 1,000$, times of cross verification $N_{CV} = 20$, subsampling fraction $\alpha_{col} = 0.015$ and $\alpha_{row} = 0.9$, sampling quality control ratio $\alpha_K = 0.3$.

We compute the EVR using Equation (6), the brain maps are shown in Figure 7 which are not thresholded for visualization purpose. Table 1 shows detail information for acquisition session, including the corresponding coverage—the ratio between the number of non-zero EVR voxels and the total number of voxels in that region—to indicate the region size of discriminative features, the “Peak-EVR" and “MNI" show the peak location and

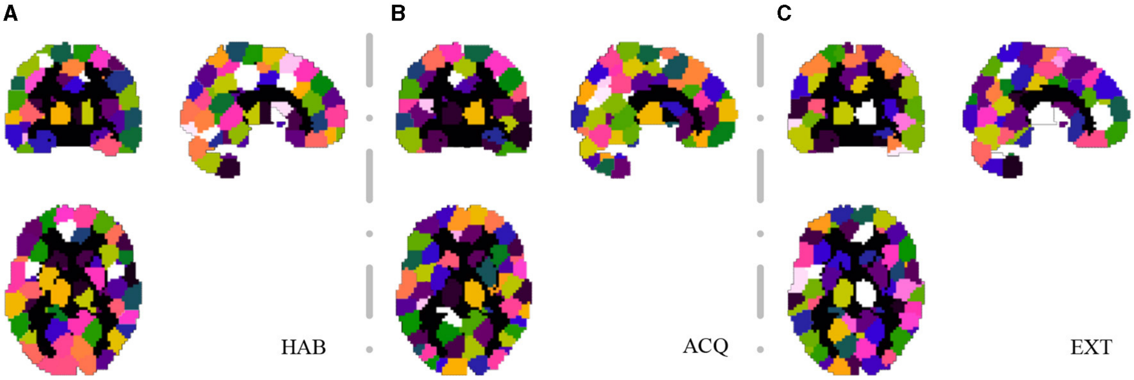


FIGURE 6
Segmentation snapshot of three experimental sessions: **(A)** habituation (HAB), **(B)** acquisition (ACQ), **(C)** extinction (EXT). Different areas are marked with different colors, for a total of 200 brain partitions.

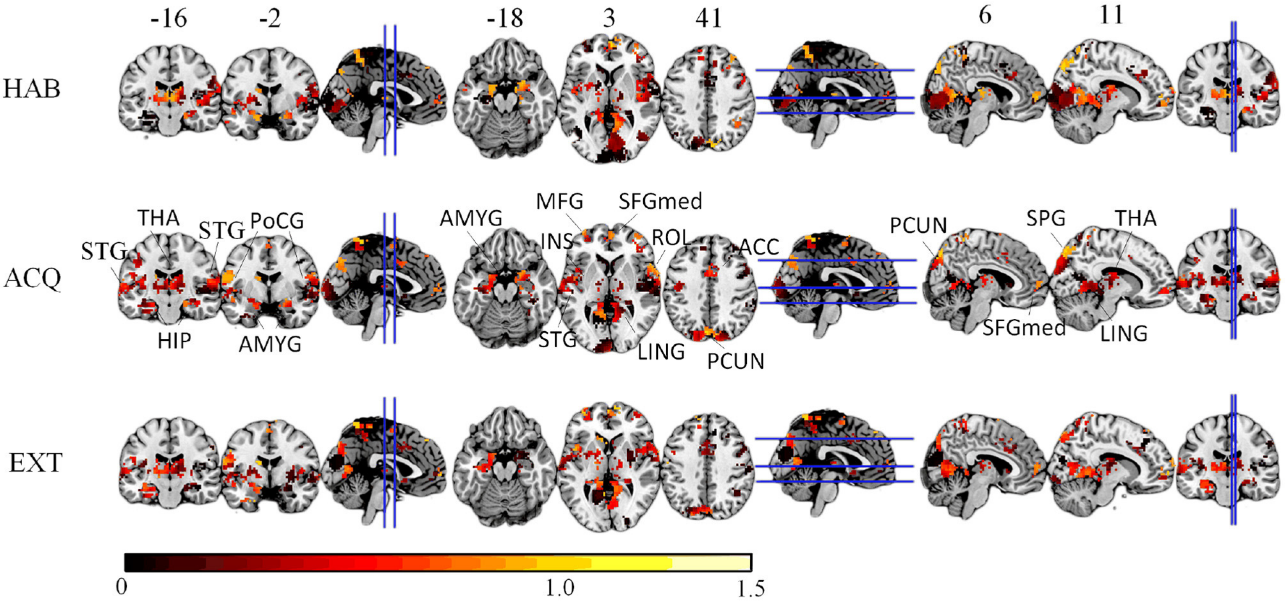


FIGURE 7
The EVR brain maps (unthreholded), which is the computed by averaging EVR across subjects.

peak intensity of each listed region. From the EVR map, the discriminative voxels across three experimental sessions largely pointed to the same regions, including the visual cortical areas such as calcarine, lingual, cuneus, occipital, and fusiform gyrus, and a set of functionally connected brain regions such as the superior frontal gyrus (orbital and medial part), postcentral, the superior temporal gyrus, the superior and middle temporal pole, precuneus and parietal gyrus, anterior cingulate cortex, insula, amygadala and thalamus. For acquisition, ROIs got the highest regional coverage are: the calcarine, lingual, superior temporal gyrus, hippocampus and parahippocampus, thalamus, as well as middle frontal gyrus, parietal, precuneus, postcentral and fusiform gyrus for their absolute number of discriminative voxels. To test the influence of N_{sel} to the results of cluster voting rates, Figure 8 is added. For most regions, increasing the number of selected clusters yield larger overlap across subjects.

To quantify the relative importance of discriminative voxels, we compute the mean effective vote ratio (EVR, see Eqution 6) across nine subjects. The resulted brain maps are shown in Figure 7, which are not thresholded for visualization purposes, meaning that the zeros displayed are actually zeros. By visual inspection, it is easy to detect the significant discriminative area. For the convenience of comparison, we also illustrate the EVR results of nine subjects in Figure 9, that only data from a single subject are used.

Lastly, we compute the stability index $\overline{O}$ and the averaged test accuracy $\bar{R}$ both for our proposed method and alternative method. The results are compared for 3 experimental sessions: habituation (HAB), acquisition (ACQ), and extinction (EXT), as

TABLE 1 The region size/coverage of discriminative features, the peak EVR value and the corresponding MNI coordinates are listed for each ROI during the acquisition session.

Location	Region size (coverage)	MNI	Peak-EVR
Calcarine	478/1,285	−6, −49, 5	0.96
Inferior occipital	7/548	−15, −100, −7	0.19
Middle occipital	113/1,592	−30, −85, 35	0.39
Superior occipital	278/840	24, −76, 47	0.98
Lingual	425/1,266	−6, −52, 2	0.90
Cuneus	204/817	6, −82, 41	1.00
Fusiform	207/1,415	−18, −43, −10	0.85
Parietal	375/2,344	9, −82, 50	1.00
Postcentral	243/2,261	−54, −4, 20	0.98
Precuneus	282/2,029	−6, −76, 41	1.00
ACC	29/390	0, 8, 41	0.71
Amygdala	57/136	24, −1, −10	0.92
Thalamus	316/663	−15, −10, 17	0.94
Insula	127/1,101	−45, 8, −7	0.84
Hippocampus	192/562	24, −16, −13	0.86
ParaHippocampus	154/634	21, 5, −25	0.92
Superior temporal	507/1,640	−51, −10, −4	0.90
Superior temporal pole	67/764	63, 14, −1	0.96
Middle temporal pole	86/2,782	−51, −61, 17	1.00
Supplementary motor	45/1,367	−6, 5, 80	0.77
Middle frontal	325/2,947	48, 50, 5	0.95
Middle frontal, orbital	58/538	21, 65, −10	0.88
Inferior frontal, triangular	50/1,435	51, 44, 5	0.83
Superior frontal	104/2,266	−36, 62, 2	0.86
Putamen	25/597	−30, −19, 8	0.28

shown in Table 2. Compared to the alternative approach, the voxels selected by our method achieves higher test correct ratio/prediction accuracy. As indicated by the stability index, our results yield solutions that more consistent and concentrated between individuals. Meanwhile, the test accuracy stably increases across experimental sessions and suggests heightened discrimination between threat and safety in visual regions in acquisition compared to habituation.

4. Discussion

We conduct numerical experiments on synthetic data and commonly used public dataset to test and cross-validate our proposed method. The results show that explicitly accounting for stability/groupwise consistency during the model optimization can mitigate some of the instability inherent in sparse methods. In particular, using the mixed $l1$ and $l2$ norm as a joint optimization criterion allows pooling data from multiple subjects and can lead to solutions that are concentrated in a few brain regions between different individuals. The number of selected candidate features is allowed to be much larger when incorporating group structure, which allows us a more global search among brain regions. Introducing groupwise regularization as an additional optimization criterion may offer promise for future methodological developments in the analysis of small-sample fMRI dataset.

These results are in line with recent predictive coding models (Rao and Ballard, 1999; Friston, 2005; Spratling, 2008), in which separate populations of neurons within a cortical region code the current estimate of sensory causes (predictions) and the mismatch between this estimate and incoming sensory signals (prediction error). Here, we did not manipulate the prior expectation of the occurrence or omission of stimuli (grating stimuli were present in all trials), but the likelihood of the stimulus having a certain

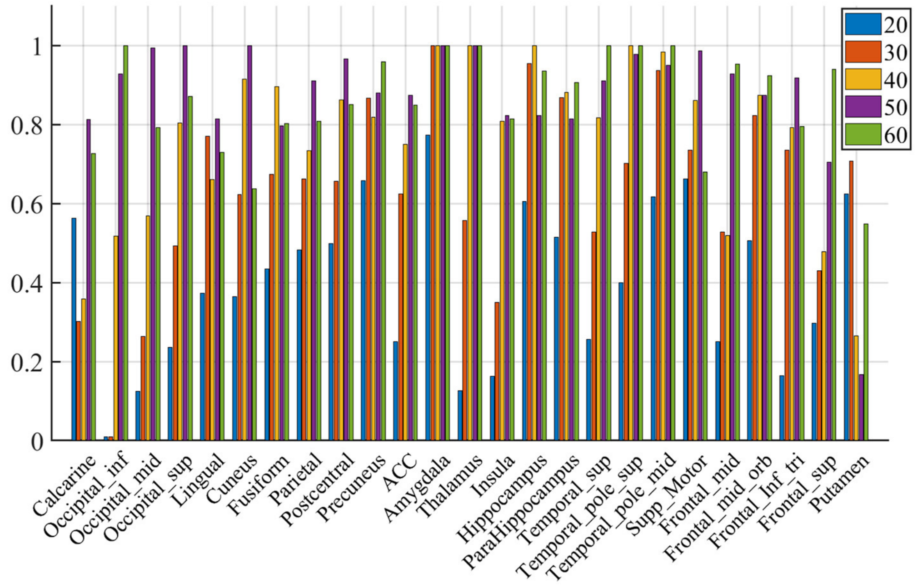


FIGURE 8
Voting rates changes with different N_{sel}, in the proportion of the vote across nine subjects. The results are for acquisition for demonstration purposes only.

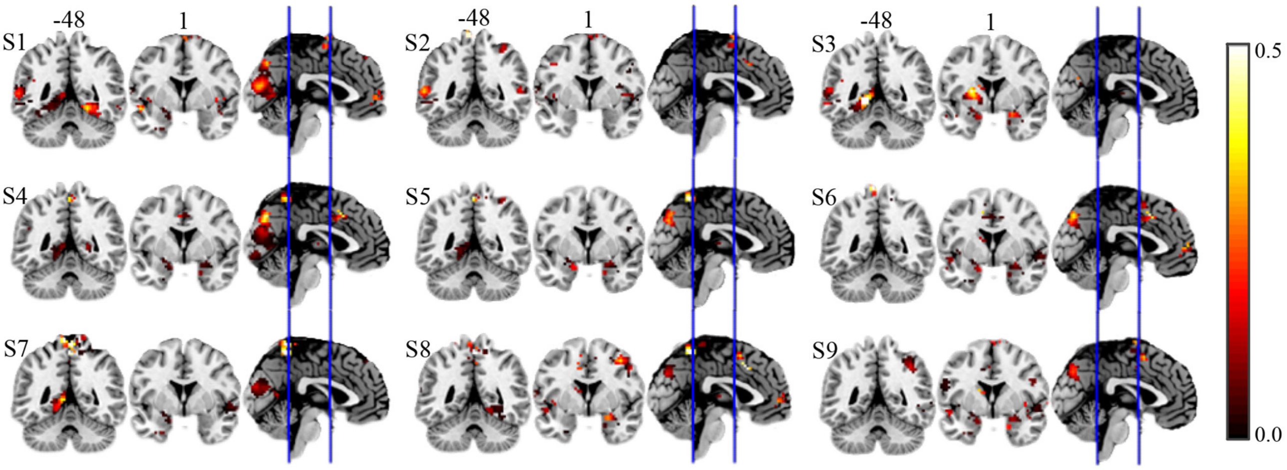


FIGURE 9
EVR results of nine subjects of alternative approach on real fMRI data.

feature (i.e., orientation) and it's followed by an electric shock. Thus, expectancy about the events during CS− (safe outcome) vs. CS+ (shock will occur after a fixed time interval) is learned as the experimental session progresses.

Finally the proposed method also resulted in findings that converge with other approaches, and with theoretical and computational models or fear conditioning and object recognition. Specifically, we found heightened discrimination between threat and safety in visual regions in acquisition compared to habituation, and we found increasing sparsification as fear learning progressed. It is worthy to note that, the prediction accuracy (the correct ratio on test set) may be significantly above chance, but far from perfect. This indicates that the code contains some linearly decodable information, but claims of linear separability may be difficult to evaluate as it would require attributing the substantial proportion of errors to limitations of the measurements (noise and subsampling), rather than to a lack of linear separability of the neuronal activity patterns. In the case of object perception, the method proposed in this thesis resulted in more robust and spatially coherent regions, illustrating its potential usefulness and applicability to a wide range of questions in cognitive neuroscience.

TABLE 2 The stability index and the averaged test accuracy of our proposed method and alternative method across three experimental sessions, habituation (HAB), acquisition (ACQ), and extinction (EXT), respectively.

Session	$\overline{O}_{alter.}$	$\overline{O}_{our}$	$\bar{R}_{alter.}$	$\bar{R}_{our}$
HAB	0.12	0.86	0.62	0.65
ACQ	0.20	0.87	0.65	0.69
EXT	0.22	0.87	0.70	0.73

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 Behavioral/NonMedical Institutional Review Board, University of Florida. 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

NZ and HJ: conceptualization. HJ and XZ: methodology. HJ and AK: investigation. HJ, XZ, and AK: writing. BC, ZY, and AK: supervision. HJ: funding acquisition. All authors had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by National Natural Science Foundation Grant (Program No. 62106189) and Natural Science Basic Research Plan in Shaanxi Province of China (Program No. 2021JQ-674) to HJ. The funding sources had no involvement in the study design.

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

Arlot, S., and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Stat. Surv.* 4, 40–79. doi: 10.1214/09-SS054

Bach, F., Jenatton, R., Mairal, J., and Obozinski, G. (2012). Structured sparsity through convex optimization. *Stat. Sci.* 27, 450–468. doi: 10.1214/12-STS394

Bakker, B., and Heskes, T. (2003). Task clustering and gating for bayesian multitask learning. *J. Mach. Learn. Res.* 4, 83–99. doi: 10.1162/153244304322765658

Baldassarre, L., Mourao-Miranda, J., and Pontil, M. (2012). "Structured sparsity models for brain decoding from fMRI data," in *2012 Second International Workshop on Pattern Recognition in NeuroImaging* (IEEE), 5–8.

Baldassarre, L., Pontil, M., and Mourão-Miranda, J. (2017). Sparsity is better with stability: combining accuracy and stability for model selection in brain decoding. *Front. Neurosci.* 11, 62. doi: 10.3389/fnins.2017.00062

Bzdok, D., Varoquaux, G., and Thirion, B. (2017). Neuroimaging research: from null-hypothesis falsification to out-of-sample generalization. *Educ. Psychol. Meas.* 77, 868–880. doi: 10.1177/0013164416667982

Cao, H., Duan, J., Lin, D., Shugart, Y. Y., Calhoun, V., and Wang, Y.-P. (2014). Sparse representation based biomarker selection for schizophrenia with integrated analysis of fMRI and snps. *Neuroimage* 102, 220–228. doi: 10.1016/j.neuroimage.2014.01.021

Chambolle, A. (2004). An algorithm for total variation minimization and applications. *J. Math. Imaging Vis.* 20, 89–97. doi: 10.1023/B:JMIV.0000011321.19549.88

Chen, G., Taylor, P. A., Shin, Y.-W., Reynolds, R. C., and Cox, R. W. (1999). Theoretical comparisons of block bootstrap methods. *Ann. Stat.* 27, 386–404. doi: 10.1214/aos/1018031117

Cour, T., Benezit, F., and Shi, J. (2005). "Spectral segmentation with multiscale graph decomposition," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), Vol. 2* (San Diego, CA: IEEE), 1124–1131.

Craddock, R. C., Jbabdi, S., Yan, C.-G., Vogelstein, J. T., Castellanos, F. X., Di Martino, A., et al. (2013). Imaging human connectomes at the macroscale. *Nat. Methods* 10, 524. doi: 10.1038/nmeth.2482

Dai, W., Xue, G.-R., Yang, Q., and Yu, Y. (2007). "Co-clustering based classification for out-of-domain documents," in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Jose, PA: ACM), 210–219.

Demirci, O., Clark, V. P., Magnotta, V. A., Andreasen, N. C., Lauriello, J., Kiehl, K. A., et al. (2008). A review of challenges in the use of fmri for disease classification/characterization and a projection pursuit application from a multi-site fMRI schizophrenia study. *Brain Imaging Behav.* 2, 207–226. doi: 10.1007/s11682-008-9028-1

Friston, K. (2005). A theory of cortical responses. *Philos. Transact. R. Soc. B Biol. Sci.* 360, 815–836. doi: 10.1098/rstb.2005.1622

Geman, S., Bienenstock, E., and Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Comput.* 4, 1–58. doi: 10.1162/neco.1992.4.1.1

Gramfort, A., Thirion, B., and Varoquaux, G. (2013). "Identifying predictive regions from fMRI with TV-L1 prior," in *2013 International Workshop on Pattern Recognition in Neuroimaging* (Philadelphia, PA: IEEE), 17–20.

Gramfort, A., Varoquaux, G., and Thirion, B. (2012). "Beyond brain reading: randomized sparsity and clustering to simultaneously predict and identify," in *Machine Learning and Interpretation in Neuroimaging* (Sierra Nevada: Springer), 9–16.

Hanson, S. J., Matsuka, T., and Haxby, J. V. (2004). Combinatorial codes in ventral temporal lobe for object recognition: Haxby (2001) revisited: is there a "face" area? *Neuroimage* 23, 156–166. doi: 10.1016/j.neuroimage.2004.05.020

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., and Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* 293, 2425–2430. doi: 10.1126/science.1063736

Hebiri, M., and Van de Geer, S. (2011). The smooth-lasso and other $\ell1+\ell2$-penalized methods. *Electron. J. Stat.* 5, 1184–1226. doi: 10.1214/11-EJS638

Hoyos-Idrobo, A., Varoquaux, G., Schwartz, Y., and Thirion, B. (2018). FReM-scalable and stable decoding with fast regularized ensemble of models. *Neuroimage* 180, 160–172. doi: 10.1016/j.neuroimage.2017.10.005

Jenatton, R., Audibert, J.-Y., and Bach, F. (2011). Structured variable selection with sparsity-inducing norms. *J. Mach. Learn. Res.* 12, 2777–2824. doi: 10.48550/arXiv.0904.3523

Ji, H., Chen, B., Petro, N. M., Yuan, Z., Zheng, N., and Keil, A. (2019). Functional source separation for EEG-fMRI fusion: application to steady-state visual evoked potentials. *Front. Neurorobot.* 13, 24. doi: 10.3389/fnbot.2019.00024

Ji, H., Petro, N. M., Chen, B., Yuan, Z., Wang, J., Zheng, N., et al. (2018). Cross multivariate correlation coefficients as screening tool for analysis of concurrent EEG-fMRI recordings. *J. Neurosci. Res.* 96, 1159–1175. doi: 10.1002/jnr.24217

Kia, S. M., Vega Pons, S., Weisz, N., and Passerini, A. (2017). Interpretability of multivariate brain maps in linear brain decoding: definition, and heuristic quantification in multivariate analysis of MEG time-locked effects. *Front. Neurosci.* 10, 619. doi: 10.3389/fnins.2016.00619

Lemm, S., Blankertz, B., Dickhaus, T., and Ḿuller, K.-R. (2011). Introduction to machine learning for brain imaging. *Neuroimage* 56, 387–399. doi: 10.1016/j.neuroimage.2010.11.004

Li, Z., Liu, J., Yang, Y., Zhou, X., and Lu, H. (2014). Clustering-guided sparse structural learning for unsupervised feature selection. *IEEE Trans. Knowl. Data Eng.* 26, 2138–2150. doi: 10.1109/TKDE.2013.65

Meinshausen, N., and Bühlmann, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B* 72, 417–473. doi: 10.1111/j.1467-9868.2010.00740.x

Micchelli, C. A., Morales, J. M., and Pontil, M. (2013). Regularizers for structured sparsity. *Adv. Comput. Math.* 38, 455–489. doi: 10.1007/s10444-011-9245-9

Michel, V., Gramfort, A., Varoquaux, G., Eger, E., and Thirion, B. (2011). Total variation regularization for fMRI-based prediction of behavior. *IEEE Trans. Med. Imaging* 30, 1328–1340. doi: 10.1109/TMI.2011.21 13378

Mourao-Miranda, J., Bokde, A. L., Born, C., Hampel, H., and Stetter, M. (2005). Classifying brain states and determining the discriminating activation patterns: support vector machine on functional MRI data. *Neuroimage* 28, 980–995. doi: 10.1016/j.neuroimage.2005.06.070

Nemirovski, A. (2000). Topics in non-parametric statistics. *Ecole d'Eté de Probabilités de Saint-Flour* 28, 85.

Pan, S. J., and Yang, Q. (2010). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* 22, 1345–1359. doi: 10.1109/TKDE.2009.191

Park, M. Y., Hastie, T., and Tibshirani, R. (2006). Averaged gene expressions for regression. *Biostatistics* 8, 212–227. doi: 10.1093/biostatistics/kxl002

Pereira, F., Mitchell, T., and Botvinick, M. (2009). Machine learning classifiers and fMRI: a tutorial overview. *Neuroimage* 45, S199-S209. doi: 10.1016/j.neuroimage.2008. 11.007

Petro, N. M., Gruss, L. F., Yin, S., Huang, H., Miskovic, V., Ding, M., et al. (2017). Multimodal imaging evidence for a frontoparietal modulation of visual cortex during the selective processing of conditioned threat. *J. Cogn. Neurosci.* 29, 953–967. doi: 10.1162/jocn_a_01114

Raina, R., Ng, A. Y., and Koller, D. (2006). "Constructing informative priors using transfer learning," in *Proceedings of the 23rd International Conference on Machine Learning* (Pittsburgh, PA: ACM), 713–720.

Rao, R. P., and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* 2, 79. doi: 10.1038/4580

Rasmussen, P. M., Hansen, L. K., Madsen, K. H., Churchill, N. W., and Strother, S. C. (2012). Model sparsity and brain pattern interpretation of classification models in neuroimaging. *Pattern Recognit.* 45, 2085–2100. doi: 10.1016/j.patcog.2011.09.011

Rondina, J. M., Hahn, T., de Oliveira, L., Marquand, A. F., Dresler, T., Leitner, T., et al. (2014). Scors—a method based on stability for feature selection and mapping in neuroimaging. *IEEE Trans. Med. Imaging* 33, 85–98. doi: 10.1109/TMI.2013.2281398

Saha, S., Mamun, K. A., Ahmed, K., Mostafa, R., Naik, G. R., Darvishi, S., et al. (2021). Progress in brain computer interface: challenges and opportunities. *Front. Syst. Neurosci.* 15, 578875. doi: 10.3389/fnsys.2021.578875

Shah, R. D., and Samworth, R. J. (2013). Variable selection with error control: another look at stability selection. *J. R. Stat. Soc. Ser. B* 75, 55–80. doi: 10.1111/j.1467-9868.2011.01034.x

Shi, J., and Malik, J. (2000). *Normalized Cuts and Image Segmentation.* (IEEE). 107.

Shimizu, Y., Yoshimoto, J., Toki, S., Takamura, M., Yoshimura, S., Okamoto, Y., et al. (2015). Toward probabilistic diagnosis and understanding of depression based on functional MRI data analysis with logistic group LASSO. *PLoS ONE* 10, e0123524. doi: 10.1371/journal.pone.0123524

Spratling, M. W. (2008). Predictive coding as a model of biased competition in visual attention. *Vision Res.* 48, 1391–1408. doi: 10.1016/j.visres.2008.03.009

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., et al. (2002). Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* 15, 273–289. doi: 10.1006/nimg.2001.0978

Wan, J., Zhang, Z., Rao, B. D., Fang, S., Yan, J., Saykin, A. J., et al. (2014). Identifying the neuroanatomical basis of cognitive impairment in Alzheimer's disease by correlation-and nonlinearity-aware sparse bayesian learning. *IEEE Trans. Med. Imaging* 33, 1475–1487. doi: 10.1109/TMI.2014.2314712

Wang, J., and Zheng, N. (2014). Measures of linear correlation for multiple variables. *arXiv.*

Wang, Y., Zheng, J., Zhang, S., Duan, X., and Chen, H. (2015). Randomized structural sparsity via constrained block subsampling for improved sensitivity of discriminative voxel identification. *Neuroimage* 117, 170–183. doi: 10.1016/j.neuroimage.2015.05.057

Wen, Z., Yu, T., Yu, Z., and Li, Y. (2019). Grouped sparse bayesian learning for voxel selection in multivoxel pattern analysis of fMRI data. *Neuroimage* 184, 417–430. doi: 10.1016/j.neuroimage.2018.09.031

Witten, D. M., Shojaie, A., and Zhang, F. (2014). The cluster elastic net for high-dimensional regression with unknown variable grouping. *Technometrics* 56, 112–122. doi: 10.1080/00401706.2013.810174

Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G., and Vaughan, T. M. (2002). Brain-computer interfaces for communication and control. *Clin. Neurophysiol.* 113, 767–791. doi: 10.1016/S1388-2457(02)00057-3

Xiang, S., Shen, X., and Ye, J. (2015). Efficient nonconvex sparse group feature selection via continuous and discrete optimization. *Artif. Intell.* 224, 28–50. doi: 10.1016/j.artint.2015.02.008

Yan, J., Li, T., Wang, H., Huang, H., Wan, J., Nho, K., et al. (2015). Cortical surface biomarkers for predicting cognitive outcomes using group l2, 1 norm. *Neurobiol. Aging* 36, S185-S193. doi: 10.1016/j.neurobiolaging.2014.07.045

Ye, J., Farnum, M., Yang, E., Verbeeck, R., Lobanov, V., Raghavan, N., et al. (2012). Sparse learning and stability selection for predicting MCI to AD conversion using baseline ADNI data. *BMC Neurol.* 12, 46. doi: 10.1186/1471-2377-12-46

Yuan, M., and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B* 68, 49–67. doi: 10.1111/j.1467-9868.2005.00532.x

Yuan, L., Liu, J., and Ye, J. (2011). "Efficient methods for overlapping group LASSO," in *Advances in Neural Information Processing Systems* (IEEE), 352–360.

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B* 67, 301–320. doi: 10.1111/j.1467-9868.2005.00503.x

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 27 September 2023
DOI 10.3389/fnins.2023.1246778

Learning-based sliding mode synchronization for fractional-order Hindmarsh-Rose neuronal models with deterministic learning

Danfeng Chen[1]*, Junsheng Li[1], Chengzhi Yuan[2], Jun He[1] and Wenbo Zhu[1]

[1]School of Mechatronic Engineering and Automation, Foshan University, Foshan, China, [2]Department of Mechanical, Industrial and Systems Engineering, University of Rhode Island, Kingston, RI, United States

OPEN ACCESS

EDITED BY
Shihui Ying,
Shanghai University, China

REVIEWED BY
Njitacke Tabekoueng Zeric,
University of Buea, Cameroon
Pei Cheng,
Anhui University, China

*CORRESPONDENCE
Danfeng Chen
✉ cdf2017@fosu.edu.cn

RECEIVED 24 June 2023
ACCEPTED 07 September 2023
PUBLISHED 27 September 2023

CITATION
Chen D, Li J, Yuan C, He J and Zhu W (2023) Learning-based sliding mode synchronization for fractional-order Hindmarsh-Rose neuronal models with deterministic learning. *Front. Neurosci.* 17:1246778. doi: 10.3389/fnins.2023.1246778

Introduction: In recent years, extensive research has been conducted on the synchronous behavior of neural networks. It is found that the synchronization ability of neurons is related to the performance of signal reception and transmission between neurons, which in turn affects the function of the organism. However, most of the existing synchronization methods are faced with two difficulties, one is the structural parameter dependency, which limits the promotion and application of synchronous methods in practical problems. The other is the limited adaptability, that is, even when faced with the same control tasks, for most of the existing control methods, the control parameters still need to be retrained. To this end, the present study investigates the synchronization problem of the fractional-order HindmarshRose (FOHR) neuronal models in unknown dynamic environment.

Methods: Inspired by the human experience of knowledge acquiring, memorizing, and application, a learning-based sliding mode control algorithm is proposed by using the deterministic learning (DL) mechanism. Firstly, the unknown dynamics of the FOHR system under unknown dynamic environment is locally accurately identified and stored in the form of constant weight neural networks through deterministic learning without dependency of the system parameters. Then, based on the identified and stored system dynamics, the model-based and relearning-based sliding mode controller are designed for similar as well as new synchronization tasks, respectively.

Results: The synchronization process can be started quickly by recalling the empirical dynamics of neurons. Therefore, fast synchronization effect is achieved by reducing the online computing time. In addition, because of the convergence of the identification and synchronization process, the control experience can be constantly replenished and stored for reutilization, so as to improve the synchronization speed and accuracy continuously.

Discussion: The thought of this article will also bring inspiration to the related research in other fields.

KEYWORDS

fractional-order Hindmarsh-Rose system, synchronization control, sliding mode control, deterministic learning, system identification

1. Introduction

In recent years, the dynamic behavior of complex networks, especially neural networks, has attracted extensive attention. It is found that the performance of signal reception and transmission between neurons affecting the functions of the organism depends on the synchronization ability of neurons. The most commonly mentioned neurological diseases such as Epileptic, Parkinson's, Alzheimer's, autism, and schizophrenia, are closely related to the synchronization ability of brain neurons (Yang et al., 2021; Zeng et al., 2023). Precisely, it has been proved that decreased synchronization can lead to neural disorders such as schizophrenia, while increased synchronization abnormalities may induce neurological diseases such Parkinson's disease and epilepsy (Uhlhaas et al., 2006). In addition, as presented in Brown et al. (2004), the firing rate of neurons in the subthalamic basement nucleus (STN) and the medial Globus Pallidus (GPI) in Parkinson's patients not only increased but also tended to synchronize abnormally.

For these neuropathies mentioned above, electrical stimulation method (Liu et al., 2019) was the most commonly used clinical treatment method. The abnormal synchronization of neurons is calibrated by adjusting the direction, frequency, and amplitude of the stimulation current. However, for different neurological diseases, how to choose or set optimal parameters of the stimulation current is a difficult problem.

Considering that the process of information generation, transmission, and decoding between neurons are closely related to their complex discharge activities, it is of paramount significance to simulate the electrical activity of neurons through mathematical models. Thus, many research studies are conducted on differential equation models to further analyze the influence of the parameter variations on the neuronal electrophysiological processes and firing activities. Among the various differential neuron models, such as the Hodgin-Huxley (HH) (Hodgkin and Huxley, 1952), FitzHugh-Nagumo (FHN) (Fitzhugh, 1961), Hindmarsh-Rose (HR) (Hindmarsh et al., 1984) and Ermentrout (Ermentrout, 2014) neuronal models, the HR neuronal model is the most commonly used one for non-linear dynamic and synchronization analysis (Parastesh et al., 2019; Liu et al., 2021; Remi et al., 2022).

The HR model possesses simple polynomial expression and can accurately describe the process of signal transmission across neurons. In Boaretto et al. (2018), the HR model was introduced to study the dynamic mechanism of abnormal phase synchronization. As discussed in Simo et al. (2021), the effect of the electromagnetic on the HR model under weak electric environment was considered to simulate the electrical activities and the synchronization process of neurons. In Ding et al. (2022), the dynamics of fractional-order memristor-coupled Hindmarsh-Rose neuron model considering synaptic crosstalk was investigated. It revealed that there were differences between the number and stability of equilibrium points for different crosstalk strength parameters. As discussed in Li et al. (2023), Fourier coefficients are introduced to investigate the effect of electric field on vibrational resonance for signal detection in a single neuron model and a bidirectionally coupled neuron model, respectively. It was found that the periodic external electric field of an appropriate frequency significantly enhances the vibrational resonance, which indicate that the external electric field may play a constructive role in the detection of weak signals in the brain and neuronal systems. In addition, the Hopf bifurcation, one of the typical non-linear dynamic behaviors was investigated in a memristor-coupled Hindmarsh-Rose and FitzHugh-Nagumo neurons with two time delays in Guo et al. (2023). It revealed that the time delay in HR neurons has a greater effect on blocking the synchronization than the time delay in FHN neuron.

Apart from the dynamic analysis using differential models, a large number of electrophysiological experiments have been conducted for the electrophysiological mechanism of abnormal synchronization of neurons (Jia et al., 2012). Through biological experiments given by Gu et al. (2014), it was found that the discharge frequency of the neuronal system became faster with the increase of potassium (K^{+}) concentration. Furthermore, with the variation of the concentration of potassium, the neuronal system showed different firing models, which was consistent with the dynamic performance of HR model under external stimulus current. In Jia et al. (2017), the authors conducted corresponding biological experiments on the calcium ion (Ca^{2+}) of neuron cells. Similar results have been achieved, that is, with the change of calcium concentration in a certain range, the neurons exhibit single-cycle, double-cycle, and chaotic electrical activity. In , it indicates that by adjusting the concentration of calcium ions (Ca^{2+}) of neurons, the corresponding inter spike interval (ISI) demonstrates similar features with the dynamic behaviors of the HR model with the variation of system parameters. These results further demonstrate that it is feasible and effective to analyze the electrical activity of neurons according to the non-linear dynamics of the HR model with different system parameters, such as the external stimulating current and other related parameters.

Recently, increasing attention has been focused on fractional-order (FO) calculus (Rihan et al., 2019; Wang et al., 2020; Jin et al., 2021), which is also very popular in the field of neuroscience (Dong et al., 2014). It was found that compared with the integer order model, the fractional neuron models reveal more advantages, such as the FO neuron models can describe the physical memory and genetics more accurately and can illustrate the biological characteristics more correctly in the presence of noise (Dong et al., 2014). Moreover, the stimulating dynamical features show that many neural computing features can be implemented in FO systems, which enriches the functional neuronal mechanisms. Therefore, the neural dynamic analysis method based on fractional HR model makes the model-based modeling of abnormal synchronization of neurons step up to a new stage.

In addition to the mechanism analysis based on the model and biological experiments, the synchronization control between neurons is also one of the core problems attracting people's attention. Over the past few decades, various control techniques, including neural network control (Motallebzadeh et al., 2012), feedback control (Semenov and Fradkov, 2021), adaptive control (Deng et al., 2006), fuzzy control (Nirvin et al., 2021), and sliding mode control (Chen et al., 2012; Vafaei et al., 2019), have been proposed and applied to the control and synchronization of the HR model as well as the FOHR models. As presented in Rajagopal et al. (2019), a feedback synchronization controller was designed for the fractional-order HR neuronal model, whose gain was limited to some parameter conditions. The authors in Giresse et al. (2019) designed controllers for the synchronized behaviors of coupled FOEHR neurons. Among these methods, sliding mode control and adaptive control techniques have attracted much attention due to their positive features such as guaranteed stability, strong robustness against parameter variations, and simplicity in implementation (Meng et al., 2020). As presented in Che et al. (2010), for unidirectional complete synchronization of HR neurons, a sliding mode control scheme with additional conditions was considered. However, the chattering phenomena is the main problem faced by the sliding mode control methods. Thus, many research studies are conducted to reduce the chattering problem by using different sliding mode surfaces. However, most of the results show that there is a tradeoff between control error and the control smoothness. In addition to the problem mentioned above, most of these control methods depend heavily on the system models. For most of the actual dynamic systems, the system models have some uncertainty because of the influence of dynamic environment (Rabah et al., 2017; Xu et al., 2020). How to avoid the influence of system uncertainty and disturbance on the control system performance is of great significance for the synchronization control of chaotic system. As discussed in Liu et al. (2021), the adaptive radial basis function (RBF) neural network was introduced for the identification of the unknown system dynamics of the HR model. However, the training time and computation cost of neural network observer inevitably increase greatly.

In Wang and Hill (2018), Wang proposed a deterministic learning (DL) theory mainly discussing the problem of knowledge learning and reutilization of non-linear dynamic systems under unknown dynamic environment by using the RBF neural networks. It has proved that for any period or period-like system input, the persistence of excitation (PE) condition can be satisfied and the precise convergence of the neural network weights can be achieved. With the development of deterministic learning theory, it was further applied for the problem of dynamic pattern recognition (Lin et al., 2019), period-doubling bifurcation detection (Chen and Wang, 2016), and intelligent control (Zhang et al., 2023). The DL algorithm emphasizes the preservation and reutilization of system dynamic knowledge. When faced with similar recognition or control tasks, it can quickly recall the identified and stored system knowledge so as to reduce the online computation time.

Inspired by the above discussion, the dynamic characteristics of the HR model, especially the FOHR neuron model under unknown dynamic environment is considered in this study and the model-based and learning-based sliding mode control algorithm are proposed by using the deterministic learning (DL) mechanism. Since the system dynamics of the slave system is unknown as considered in this study, in order to achieve ideal robustness effect of the control system, the traditional sliding mode control method usually sets too large gain parameters to overcome the system uncertainties, which in turn leads to serious chattering problem. In our study, the sliding mode gain parameter is effectively reduced by compensating the system dynamics with locally accurate system identification. First, the unknown dynamics of the FOHR system under unknown dynamic environment is locally accurately identified and stored in the form of constant weight neural networks through deterministic learning without dependency of the system parameters. Then, the model-based and learning-based sliding mode controllers based on the identified and stored system dynamics are designed for the similar and new synchronization tasks, respectively. Therefore, the fast synchronization effect is achieved through recalling the empirical dynamics of neurons. Moreover, the control experience can be constantly replenished and stored for reutilization due to the convergence of the identification and synchronization process, which help improves the synchronization speed and accuracy continuously.

2. Methods and innovations

In this section, the method proposed in this study and the main innovations are briefly introduced. Aiming at the problem of abnormal synchronization of neurons under unknown dynamic environment, the sliding-mode control method is introduced. Different from traditional sliding-mode control policy, the human experience of knowledge acquisition, storage, and re-application is introduced to the control process of our study. Precisely, the unknown dynamic information of the neuron system is identified and stored according to the deterministic learning mechanism by using the RBF neural network. The dynamic information is further applied to the controller to achieve more accurate synchronization effect, which is called the model-based sliding-mode control. Considering the case that the stored dynamic information is limited and the unknown slave system can not be well matched, the relearning-based sliding-mode control is proposed. During this propose, the identified and control experience can be updated and supplemented to the dynamic patter database, which can provide experience for new and unfamiliar synchronization tasks. Thus, the online computing time is shortened, and a better synchronization effect can be achieved. In addition, the problem of excessive chattering faced by traditional sliding mode control method can be solved skillfully by selection small sliding-mode gain by using experiential information. The flowchart of the method is available in Figure 1. The emphasis of our study is not only on the effect of synchronization but also on the efficient storage and reutilization of the experience knowledge in the process of neural system synchronization, which is the essence of intelligent learning and intelligent control and not covered by most existing research studies.

3. Preliminary knowledge

3.1. Fractional-order definition and basic properties

The fractional order (FO) calculus has a very long mathematical history and has gained extensive attention in the areas of science and engineering with the advent of high computational devices recent years. The FO calculus can be seen as the comprehensive and generalized version of the conventional integer-order calculus, which encompassed both fractional and integer-order differential and integral equations (Dar et al., 2022). Correspondingly, the FO derivative possesses complex or real arbitrary order, for which various mathematical operators have been proposed. Among those

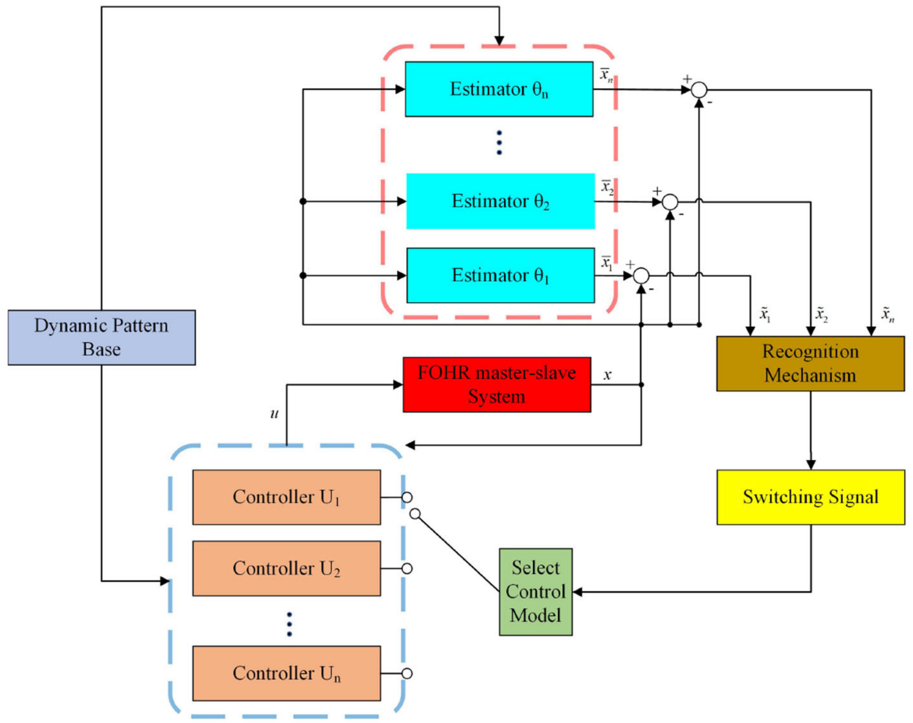


FIGURE 1
Flowchart of the method.

operators, the Grunwald-Letnikov (GL) (Huang, 2016), Liouvill (L) (Huang, 2016), Riemamn-Liouville (RL) (Efe, 2009), and Caputo (C) (Gorenflo and Mainardi, 1997) are most commonly used. Compared with the L, RL, and C operators, the GL operator pays more attention to the numerical calculation of fractional-order differentiation. Since the fractional derivative describes memory and hereditary properties in such an appropriate manner that it demonstrates much advantages in system representation compared to the integer-order models, the GL-based fractional order definition (Huang, 2016) is introduced in this study:

$$ {}_a^G D_t^q f(t) = \frac{d^q f(t)}{d(t-a)^q} = \lim_{h\to 0} h^{-q} \sum_{r=0}^{n} (-1)^r \binom{q}{r} f(t-rh), \quad (1) $$

where, G means the GL based fractional calculus, $nh = t - a$, if $q < 0$, the Equation (1) is the G-L based fractional integral definition; on the contrary, if $q > 0$, the Equation (1) is the G-L based differential definition.

3.2. Properties of fractional calculus

The main and commonly used properties of fractional derivatives are given as follows:

1. For $q = n$, where n is an integer, the operation ${}_0D_t^q f(t)$ gives the same result as classical differentiation of integer order n. Meanwhile, the fractional derivative degenerate to integer derivative.
2. For $q = 0$, the operation becomes the identity operator given as

$$ {}_0D_t^q f(t) = f(t). \quad (2) $$

3. The additive index law (semigroup property)

$$ {}_0D_t^\alpha\, {}_0D_t^\beta f(t) = {}_0D_t^\beta\, {}_0D_t^\alpha f(t) = {}_0D_t^{\alpha+\beta} f(t) \quad (3) $$

holds under some reasonable constraints on the function $f(t)$. In particular, there is

$$ D_t^q (D_t^{1-q} f(t)) = D_t^1 f(t) = \frac{d}{dt} f(t), 0 < q < 1. \quad (4) $$

3.3. The deterministic learning theory

In 2009, the deterministic learning (DL) theory was proposed for the problem of learning in uncertain dynamic environments (Wang et al., 2009). It mainly focuses on the dynamic process of knowledge learning, representation, and utilization in unknown dynamic environment. With deterministic learning, fundamental knowledge on system dynamics can be accumulated, stored, and represented by constant RBF networks in a deterministic manner. Moreover, in a scenario whereby an adaptive neural network (NN) controller achieves tracking of a periodic or periodic-like reference orbit, the deterministic learning mechanism is shown capable of achieving closed-loop identification of partial system dynamics during tracking control.

In detail, for any unknown continuous non-linear function $f(X) : \Omega_X \to R$ with recurrent system trajectories $\psi(x_0)$, in which

$\Omega_X \subset R^q$ is a compact set, an ideal constant weight vector W^* of the RBF networks exists, that is, $f(X) = W^{*T}\phi(X) + \varepsilon^*, \forall X \in \Omega_X$, where $\varepsilon^* > 0$ is the approximation error and $X \in \Omega_X \subset R^q$ denotes the input vector of the radial basic function networks (RBFNs), $W^* = [w_1^*, \cdots, w_n^*]^T \in R^N$ is the ideal RBFNs weight with N being the number of neurons. $\phi(X) = [\varphi_1(\|X - c_1\|), \cdots, \varphi_n(\|X - c_n\|)]^T$ is the regression vector of RBFs with $\varphi_i(\cdot)$ being one of the radial basic function, and c_i is the center of neurons distributed in the input space. For the radial basic function, the Gaussian function is one of the most commonly used kernel RBFs given as $\varphi_i(\|X - c_i\|) = exp[\frac{-(X-c_i)^T(X-c_i)}{\eta_i^2}]$, in which η_i is the adjacent width of the radial base kernel. It satisfies the Schoenberg theorem (Schoenberg, 1938) and is localized basis function in the sense that $\varphi_i(\|X - c_i\|) \to 0$ as $\|X\| \to \infty$. All these properties of the Gaussian function provide a rich source of RBFs that are suitable for interpolation of data in Euclidean spaces. The conditional non-singularity property is essential in proving the partial persistent excitation (PE) condition of RBF networks, which is the key to the accurate identification ability for the deterministic learning theory.

4. Dynamic identification of the fractional-order HR model via deterministic learning

4.1. The fractional-order HR model

With the development of neuroscience, various differential models have been proposed for describing the neuron system, including the Hodgin-Huxley (HH) model, the FitzHugh-Nagumo (FHN) model, the Hindmarsh-Rose (HR) model, and the Ermentrout neuronal model. Among those models, the HR model, possessing the simplest system form, can accurately describe the signal transmission process across the nerve fiber membrane. Thus, the HR model is commonly used for neuron dynamic describing and analysis. The classical three-variable HR neuronal model can be described by the following equations:

$$\begin{aligned} \dot{x} &= y - ax^3 + bx^2 - z + I \\ \dot{y} &= c - dx^2 - y, \\ \dot{z} &= r\left(s_0(x - q_0) - z\right) \end{aligned} \tag{5}$$

where x is the membrane potential, y is the recovery variable standing for the gating dynamics of the potassium (K^+) channel, and z represents the adaptation current corresponding to the dynamics of calcium (Ca^{2+}) channel. Moreover, the model parameters $a, b, c, r,$ and s_0 are positive constants, while the parameter q_0 stands for the resting potential, and I represents the external stimulation input.

The FO differential model has more advantages in neuronal dynamic description compared to that of the integer-order model. In addition, the FO system has a wider stability region. Thus, in this study, the following fractional order HR (FOHR) neuronal model is introduced, that is,

$$\begin{aligned} D_t^q x &= y - ax^3 + bx^2 - z + I \\ D_t^q y &= c - dx^2 - y, \\ D_t^q z &= r(s_0(x - q_0) - z) \end{aligned} \tag{6}$$

in which, the operator D_t^q represents the GL fractional derivative as shown in Equation (1).

The state variables and model parameters of the FOHR model possess the same physical meaning with the integer-order HR model. Through bifurcation analysis under different values of the external stimulation input I and fractional order q, the FOHR model demonstrates a wealth of dynamic behaviors, such as the subthreshold oscillations, spiking, bursting as well as chaotic behaviors.

In detail, when taking the fractional order $q = 1$, the FOHR model degenerates to an integer-order HR model. By setting $q = 1$ and the corresponding system parameters $a = 1.0, b = 3.0, c = 1.0, d = 6.0, r = 0.013, s = 4.0$, and $q_0 = -1.56$, diverse non-linear dynamics under different external stimulus I of the HR model are generated. By changing the control parameter I, the membrane potential x presents different state characteristics, which can be seen from Figure 2, in which the initial system state (x_0, y_0, z_0) is set as $(0.1, 1.0, 0.2)$.

Precisely, when setting $I = 1.5$, the neuron produces slow-peak regular spiking (single-cycle spiking) state as given in Figure 2A. Gradually increasing I to 1.8, 2.3, 2.8, the HR system exhibits regular bursting state, shown as the period-2, period-3, and period-4 bursting behaviors, respectively, which can be seen from Figures 2B–D. When I increased to 3.2, the state x of the HR system becomes chaotic as shown in Figure 2E. Further increasing I to 3.58, the system regresses to a fast single-cycle spiking state as demonstrated in Figure 2F, in which the period interval is significantly shorter and the rate of the dynamic activity is much faster than that of the interval demonstrate in Figure 2A.

4.2. The dynamic behavior of the FOHR model under fractional order *q*

Except for the non-linear behavior of the time response of the membrane potential x, the inter-spike interval (*ISI*) (Rabinovich and Abarbanel, 1998) is one of most commonly used physiological indicators, which carries important information of neuronal firing. In the following discussion, the bifurcation diagram of the peak membrane potential x_{max} and the *ISI* sequence of the FOHR neural system with different bifurcation parameters are considered.

First, the dynamic non-linearity of the FOHR model under different fractional orders q with the external excitation $I = 3$ is considered. As shown in Figure 3A, the bifurcation diagram of the *ISI* sequence exhibits a comb-shaped region with the increase of fractional order q. Correspondingly, the bifurcation diagram of the peak of the membrane potential x (denoted as x_{max}) demonstrates that the discharge characteristics of the system varies more obviously according to the fractional order. That is, with the increase of the fractional order q within a certain range, the system as a whole shows the tendency of periodic decline, and the periodic bursting phenomenon occurs as demonstrated in Figure 3B. In other words, the firing behavior of neurons becomes more complex and unstable with the increase of the fractional order of the neuronal system, exhibiting richer dynamic activity characteristics.

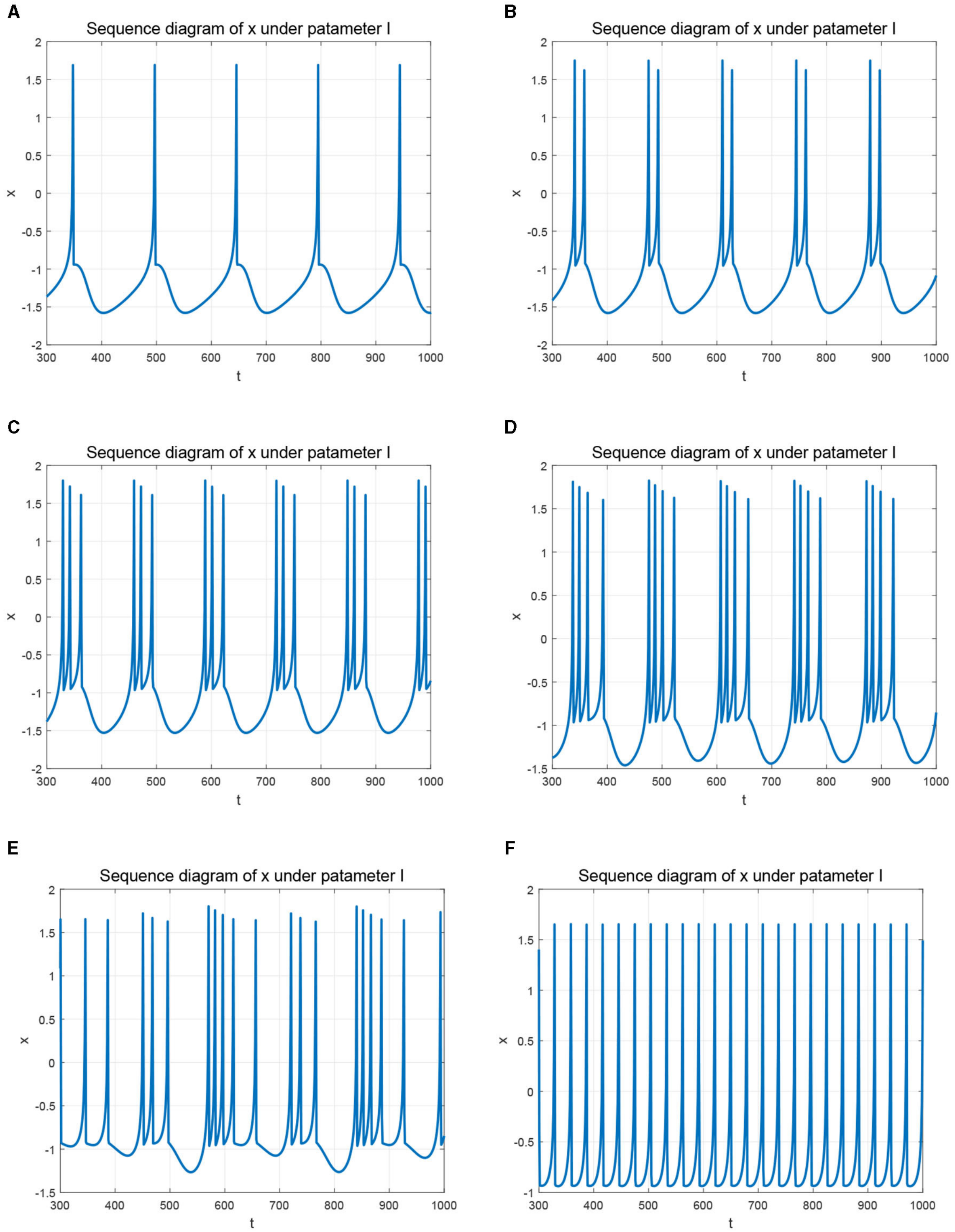


FIGURE 2
Time response of membrane potential x with different parameters I. **(A)** $I = 1.5$. **(B)** $I = 1.8$. **(C)** $I = 2.3$. **(D)** $I = 2.8$. **(E)** $I = 3.2$. **(F)** $I = 3.58$.

4.3. The dynamic behavior of the FOHR model under external excitation *I*

Second, take the external excitation I as the control parameter for analyzing the dynamic behaviors of the FOHR model with a certain fractional order. The parameter I is taken within the interval $[1.2, 4.3]$ and $r = 0.013$.

The simulation results of the bifurcation diagram of the *ISI* sequence (shown in Figure 4) exhibit that the dynamic characteristics of the FOHR system become more complex with

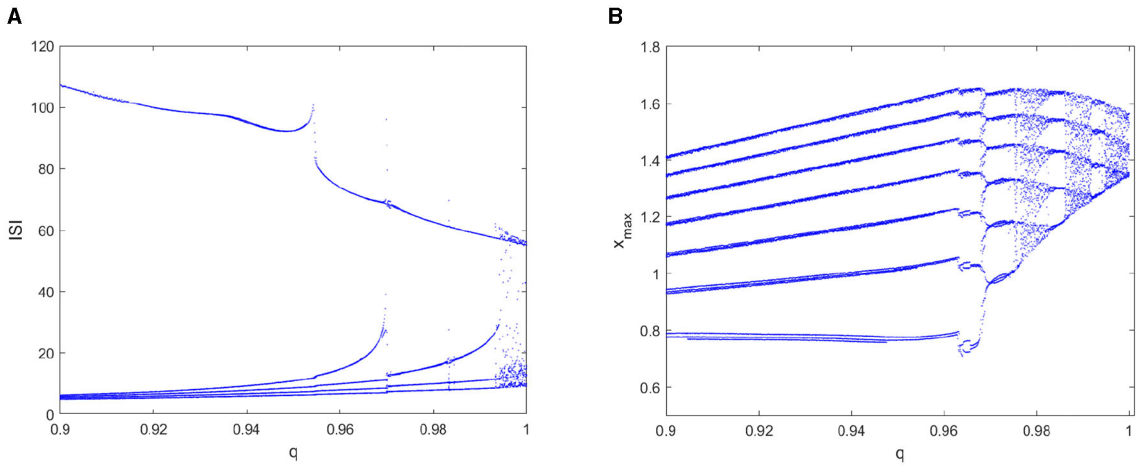


FIGURE 3
Bifurcation diagram of the *ISI* and the x_{max} sequence with parameter *q*. **(A)** bifurcation diagram of *ISI*. **(B)** bifurcation diagram of x_{max}.

the increase of I. Taking the integer order as an example (given in Figure 4A), the *ISI* sequence experiences the process of period-2 bifurcation to chaos and then back to single period by the period-doubling bifurcation process. Correspondingly, the *ISI* sequence of the 0.98-order FOHR model indicates similar discharge behaviors with that of the integer-order model, but the chaos duration is reduced. In addition, from the amplitude of the *ISI* sequence, it can be seen that the effect of external stimulus current on the system dynamics was much obvious.

According to the bifurcation diagram of the x_{max} sequence shown in Figure 5, the dynamical behaviors correspond to the same fractional order has similar and abundant dynamic characteristics with that of the *ISI* sequence. In addition, some hidden information contained in the integer order can be clearly demonstrated in the 0.98-order HR model as shown in Figure 5B, such as the period-4 cluster bursting under $I = 3.0$, the period-5 cluster discharge when $I = 3.3$, the comb-shaped region and the chaotic region. If a further decrease in the fractional order *q* to 0.96 and 0.95 as can be seen from Figures 5C, D, the dynamic structure of the FOHR system changes qualitatively. Precisely, with the increase of parameter *I*, the dynamic behavior of the FOHR model becomes more complex. The structure and stability of the system is influenced correspondingly.

4.4. The dynamic behavior of the FOHR model under parameter *r*

To further analyze the dynamic characteristics of the FOHR model, another important parameter *r* which relates to the calcium (Ca^{2+}) concentration and significant to many neurological disorders, is considered as the control parameter in this part. All the other parameters are kept as the same as mentioned above, while parameter *I* is fixed to 3.5. When ranging the parameter *r* from 0.0015 to 0.06, a variety of dynamic behaviors of the FOHR neuron system are presented. As shown in Figures 6, 7, the bifurcation diagram of the *ISI* sequence and the x_{max} sequence demonstrate similar non-linear characteristics. Moreover, compared to the integer-order HR model, the 0.98-order HR model presents a more detailed and clear dynamic process.

In conclusion, the dynamic simulations given above suggest that compared to the integer-order HR model, the fractional-order HR model can describe the numerous computational features and the non-linear dynamics of the neuron model more accurately, which help enrich the functional neuron mechanisms and further ensures more accurate dynamic analysis. Thus, it is necessary and important to introduce the FOHR model, and the FOHR model with fractional order $q = 0.98$ is taken into consideration in the following study.

5. Identification of the FOHR model via deterministic learning

The above numerical simulations are obtained based on the assumption that the parameters of neurons are known, which is also commonly used in most related research studies. However, it is too ideal for most practical neuron systems. More precisely, the neuron parameters are actually unknown and vary dynamically with the dynamic environment. Therefore, how to identify the non-linear dynamics of the neuronal model under unknown dynamic environment is essential for comprehensive understanding of the non-linear characteristics of the actual HR model. This will be the focus of the discussion below.

To identify the unknown system dynamics of the fractional order HR model, the RBF neural network is considered:

$$f_i(x;\mu) = \hat{W}_{\xi_i}^T \phi_{\xi_i}(x), \tag{7}$$

where $x \in \Omega_f \subset R^n$ is the neural networks (NNs) input, $\hat{W}_\xi = [\hat{W}_{\xi_1}, \cdots, \hat{W}_{\xi_n}]^T$ is the estimate of the ideal weight matrix, and $\phi_\xi(x) = [\varphi_{\xi_1}(x), \cdots, \varphi_{\xi_n}(x)]^T$ is chosen as a vector of Gaussian functions, that is given as

$$\varphi_{\xi_i}(x) = exp(\frac{-\|X - c_i\|}{\eta_i^2}), i = 1, \cdots, N_i, \tag{8}$$

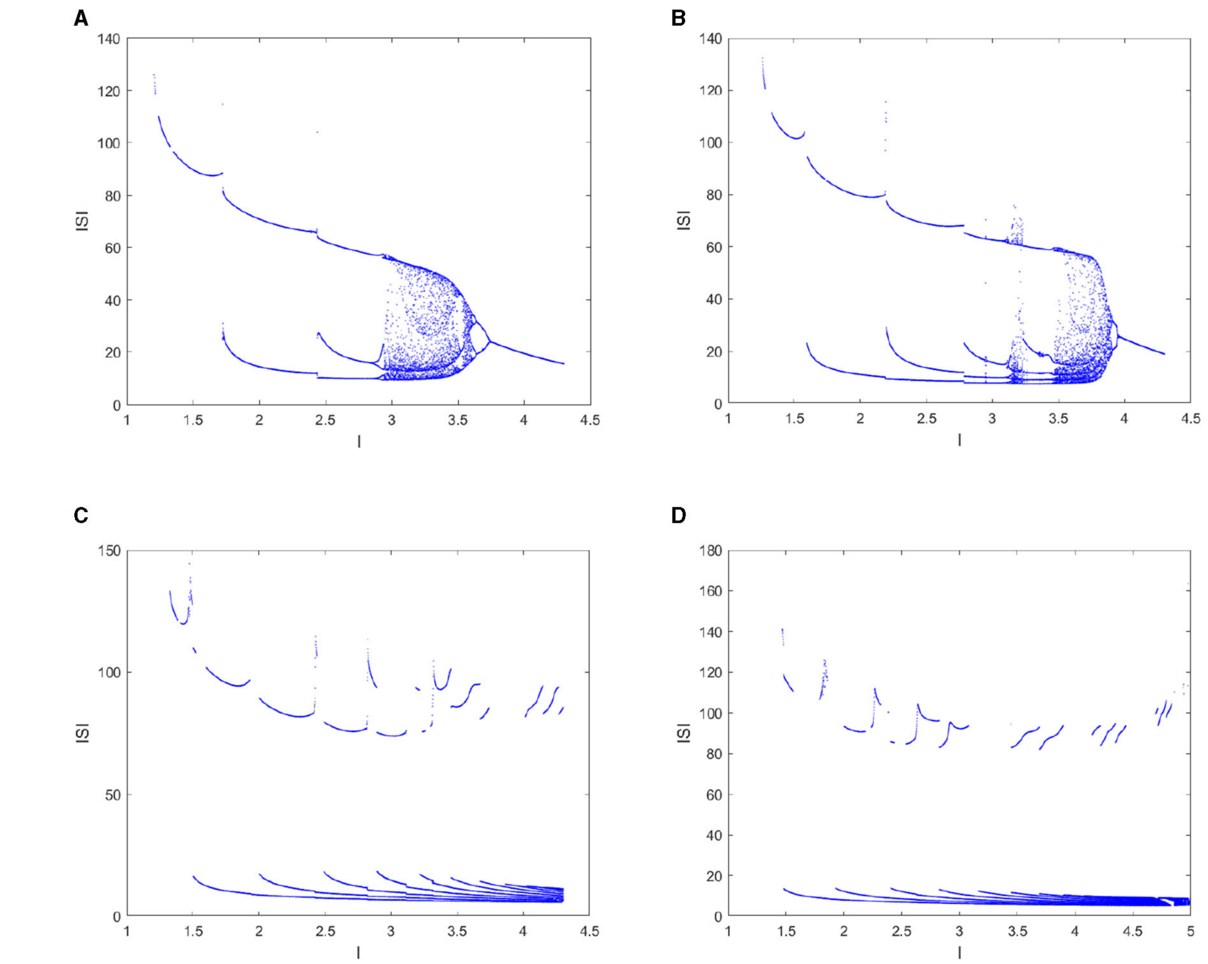


FIGURE 4
Bifurcation diagram of the *ISI* sequence with parameter *I*. **(A)** $q = 1$. **(B)** $q = 0.98$. **(C)** $q = 0.96$. **(D)** $q = 0.95$.

where c_i denotes the center vector of the *ith* basis function and η_i is the adjacent width of the basis function, $(\cdot)_\xi$ represents the neurons that was distributed close to the system trajectory, which plays the main role during the process of the neuronal dynamic identification.

Since the input of the RBFNN possesses regression property, the RBF NNs can locally accurately approximate the non-linear function along the system trajectory, that is,

$$f_i(x;\mu) = W_{\xi_i}^{*T}\varphi_{\xi_i}(x) + \varepsilon_{i_1}, \tag{9}$$

where $W_{\xi_i}^*$ is the optimal weights vector and ε_{i_1} is the bounded identification error close to zero.

The dynamic investigations and simulations of the HR model discussed above have revealed the regression characteristics of the neuron system. It is the regression property of the HR model that reminds us of the deterministic learning theory, which emphasizes that almost any period or period-like (recurrent) NN input can lead to the satisfaction of the partial persistent excitation condition (PE) along the system trajectory by using the localized RBFNs. Furthermore, the identified system dynamics can be stored due to the convergence of the NN weights; that is,

$$\bar{W}_i = mean_{t\in[t_a,t_b]}\hat{W}_i(t), \tag{10}$$

where $t_a > t_b > 0$ is the time segment referring to a piece of time segment within the convergence process of the NN weights and "mean" is the arithmetic mean. Then, the unknown system dynamics can be accurately identified and stored by the constant vector of neural networks, giving as

$$f_i(x;\mu) = \bar{W}_{\xi_i}^T\varphi_{\xi_i}(x) + \varepsilon_{i_2}, \tag{11}$$

where $\varepsilon_{i_2} = \varepsilon_{i_1} - \tilde{W}_{\xi_i}^T\varphi_{\xi_i}(x)$ is the practical approximation error of the system dynamics by using the constant NN vector $\bar{W}_{\xi_i}^T\varphi_{\xi_i}$ with $\tilde{W}_i = \hat{W}_i - W_i^*$ being a small positive number approaching zero.

For different dynamic external excitation I of the FOHR system under the given fractional order, different state trajectories are generated. Based on the approximate process by using the DL method, accurate identification of unknown system dynamics $f_i(x;\mu)$ are obtained and stored as constant RBF neural networks $\bar{W}_{\xi i}^T\varphi_{\xi i}^T$. Then, a certain number of constant RBF neural networks compose a pattern base which denoted as $\chi = \{\chi^k = \bar{W}_i^{kT} \mid k = 1,\cdots,K\}$.

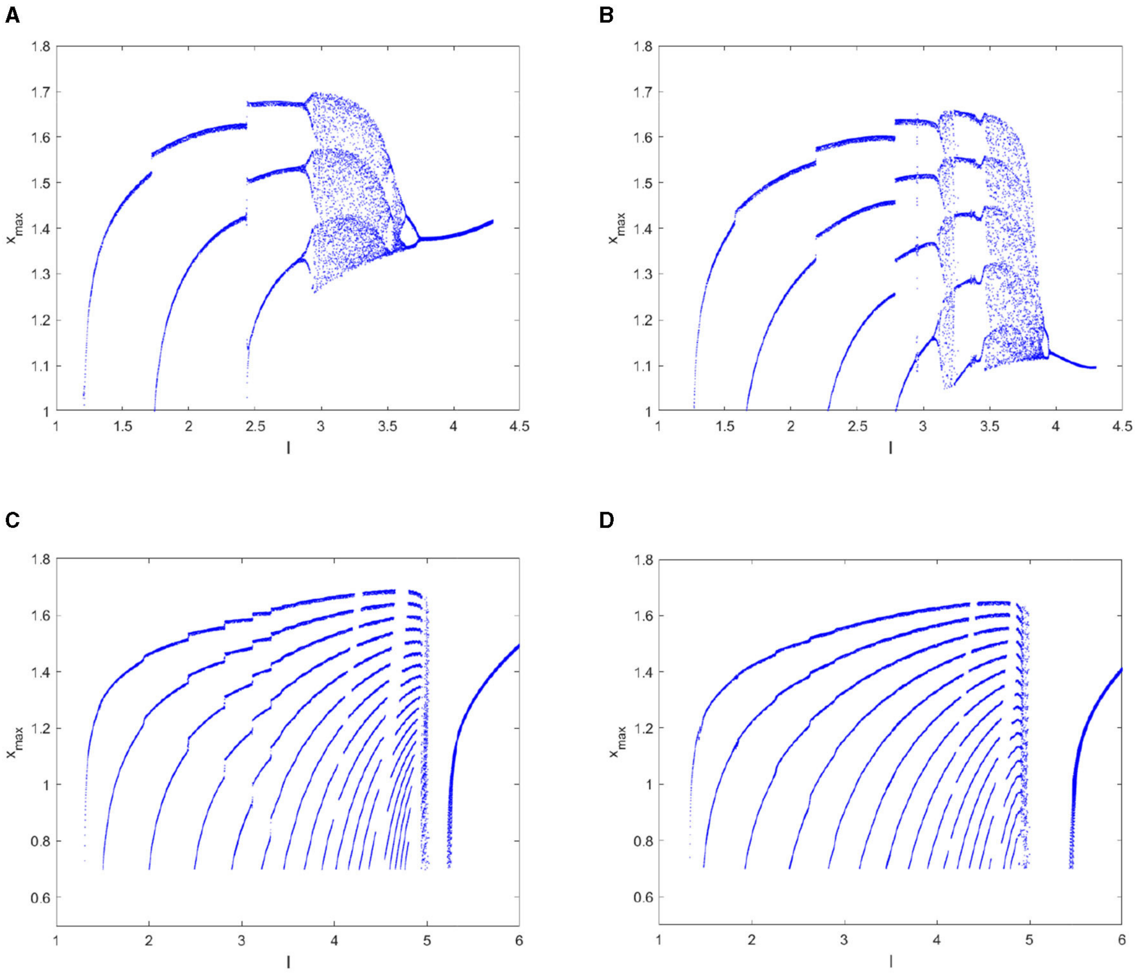


FIGURE 5
Bifurcation diagram of the x_{max} sequence with parameter I. **(A)** $q = 1$. **(B)** $q = 0.98$. **(C)** $q = 0.96$. **(D)** $q = 0.95$.

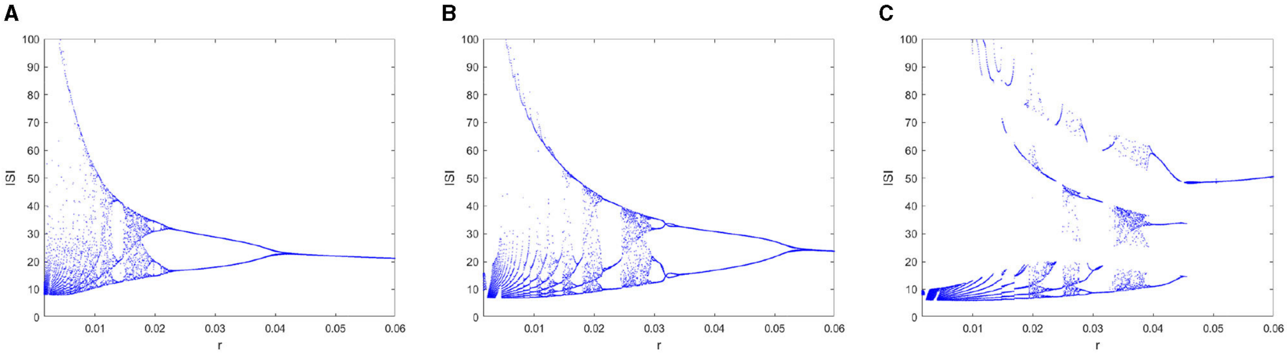


FIGURE 6
Bifurcation diagram of the ISI sequence with parameter r. **(A)** $q = 1$. **(B)** $q = 0.98$. **(C)** $q = 0.96$.

Remark 1: The process of system dynamics identification and storage of the FOHR system in unknown dynamic environment will simulate the way of human learning and memorizing new knowledge. The created pattern base, that is, the memory in the mind of knowledge, can be directly invoked in the control process.

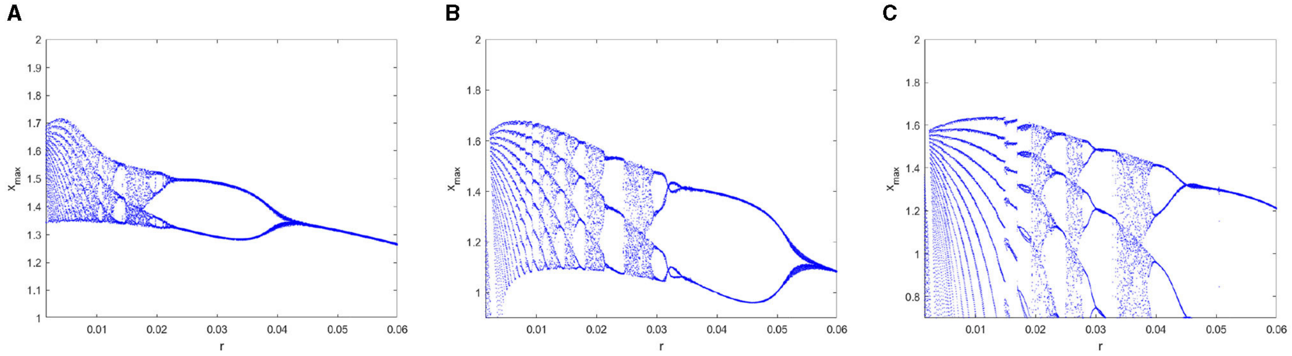


FIGURE 7
Bifurcation diagram of the x_{max} sequence with parameter r. **(A)** $q = 1$. **(B)** $q = 0.98$. **(C)** $q = 0.96$.

6. Sliding mode control of the FOHR system by using deterministic learning

6.1. Problem description

In this sub-section, the model-based sliding mode control problem of two FOHR neuronal models is considered. The two neurons interconnect in a master-slave configuration. The master FOHR neuronal model is given as follows:

$$\begin{aligned} D_t^q x_{m,1} &= x_{m,2} - a x_{m,1}^3 + b x_{m,1}^2 - x_{m,3} + I, \\ D_t^q x_{m,2} &= c_m - d x_{m,1}^2 - x_{m,2}, \\ D_t^q x_{m,3} &= r(s_0(x_{m,1} - q_0) - x_{m,3}), \end{aligned} \tag{12}$$

and the slave FOHR neuronal model under control is denoted by

$$\begin{aligned} D_t^q x_{s,1} &= x_{s,2} - a x_{s,1}^3 + b x_{s,1}^2 - x_{s,3} + I + d_1 + u_1, \\ D_t^q x_{s,2} &= c_s - d x_{s,1}^2 - x_{s,2} + d_2 + u_2, \\ D_t^q x_{s,3} &= r(s_0(x_{s,1} - q_0) - x_{s,3}) + d_3 + u_3, \end{aligned} \tag{13}$$

where $d_i, i = 1,2,3$ represents the bounded unknown external disturbance; that is, $|d_i| \leq \bar{d}_i, i = 1,2,3$ and the terms u_i and $i = 1,2,3$ denote the control inputs of the state variables.

For the convenience of discussion, the simplified master-slave neuron system models are presented as follows:

$$\begin{aligned} D_t^q x_m &= f_m(x_m), \\ D_t^q x_s &= f_s(x_s) + d_i + u_i, \end{aligned} \tag{14}$$

where $x_m = [x_{m,1}, x_{m,2}, x_{m,3}]^T$ and $x_s = [x_{s,1}, x_{s,2}, x_{s,3}]^T$ are the state vectors of the master and slave neuronal system, respectively. f_m represents the known system dynamics vectors of the master FOHR model. Correspondingly, f_s represents the unknown system dynamics vectors of the slave FOHR model. Precisely, f_s is smooth, but unknown non-linear dynamics of the slave system. d_i and u_i have the same meaning as the formula given in Equation (13). The main task in this part is to realize the synchronization of the master-slave system with proper amount of calculation and correct the synchronization error by adjusting the parameters.

6.2. Model-based sliding mode control of the FOHR system

The synchronization of the master-slave neuronal system is to drive the slave neuron system to track the state as well as the trajectory of the master system under certain external disturbance in unknown dynamic environment by properly designed controller. In order to achieve ideal stability effect of the control system, the gain parameters of the traditional sliding mode control algorithm are usually set too large, which leads to serious chattering problem. In this part, the obtained system dynamics $\bar{W}_i$ stored in the pattern base χ is applied for the sliding mode control to achieve fast synchronization performance for the master-slave neuron system. In addition, the accurate modeling of the system dynamics help reduce the synchronization error of the master-slave system without large gain, thus reducing the chattering caused by sliding mode gain.

The synchronization error of the master-slave FOHR system is defined as follows;

$$\begin{aligned} e_i &= x_{s,i} - x_{m,i}, \\ D_t^q e_i &= f_{s,i}(x) + d_i + \mu_i - f_{m,i}(x), \end{aligned} \tag{15}$$

where $i = 1,2,3$. To achieve fast synchronization of the master-slave FOHR system, the identified and stored model-based sliding mode control method is proposed. First, the fractional order proportional integral sliding surface is designed as follows:

$$\begin{aligned} s_i &= c_i D_t^{1-q} e_i + e_i, \\ D_t^q s_i &= c_i e_i + D_t^q e_i. \end{aligned} \tag{16}$$

where $s_i, (i = 1,2,3)$ is the fractional order proportional integral sliding surface. The derivative of the sliding mode surface can be achieved according to the properties of fractional order models discussed in Section 3.2, that is,

$$\dot{s}_i = c_i e_i + f_{s,i}(x) + d_i + u_i - f_{m,i}(x), \tag{17}$$

where the corresponding constant rate of convergence is designed as

$$\dot{s}_i = -\eta_i sgn(s_i). \tag{18}$$

The following sliding mode control rate is designed according to the Equations (17) and (18)

$$\mu_i = -\eta_i sgn(s_i) - c_i e_i + f_{m,i}(x) - f_{s,i}(x). \tag{19}$$

For the unknown system dynamics $f_{m,i}(x)$ of the slave system, the rapid recognition process is introduced, that is,

$$\dot{\bar{x}}_i^k = -b_i(\bar{x}_i^k - x_i) + \bar{W}_i^{kT}\varphi_i(x), k = 1, \cdots, K, \tag{20}$$

in which $\bar{x}_i^k$ represents the state of the dynamic model and the corresponding dynamic information of the system has been identified and stored in the pattern base χ as mentioned above, x_i is the *ith* state of the unknown slave system, and $b_i > 0$ is a design parameter.

For the unknown slave FOHR system, the recognition error system is given as

$$\dot{\tilde{x}}_i^k = -b_i\tilde{x}_i^k + (\bar{W}_i^{kT}\varphi_i(x) - f_{s,i}(x)), i = 1, \cdots, n, \tag{21}$$

where $\tilde{x}_i^k = \bar{x}_i^k - x_i$ is the state tracking error between the empirical pattern stored in the base and the unknown slave system.

Commonly, without identifying the unknown dynamics of the unknown slave FOHR system, the differences between the dynamic systems stored in the pattern base and the slave pattern denoted as $|\bar{W}_i^{kT}\varphi_i(x) - f_{s,i}(x)|$ shown in Equation (21) is unavailable for direct computation. However, as presented in Wang et al. (2009), the state tracking error $|\tilde{x}_i^k|$ can be explicitly measured.

For any unknown slave FOHR system with regression system trajectory $\varphi(x_{d0})$, the tracking error $|\tilde{x}_i^k|$ can be achieved within finite time by properly selecting the design parameters; that is, by introducing the average L_1-norm based dynamic similarity measure, that is given as

$$\|\tilde{x}_i^k(t)\|_1 = \frac{1}{T}\int_t^{t+T} |\tilde{x}_i^k(\tau)|d\tau, \tag{22}$$

where $T > 0$ is a design parameter, and the difference between system dynamics can be explicitly measured. Based on the similarity measure, the smallest tracking error between certain unknown slave system and the system identified as well as stored in the pattern base χ can be obtained, that is

$$x_i^0 = \min(\|\tilde{x}_i^k\|, k = 1, \cdots, m), \tag{23}$$

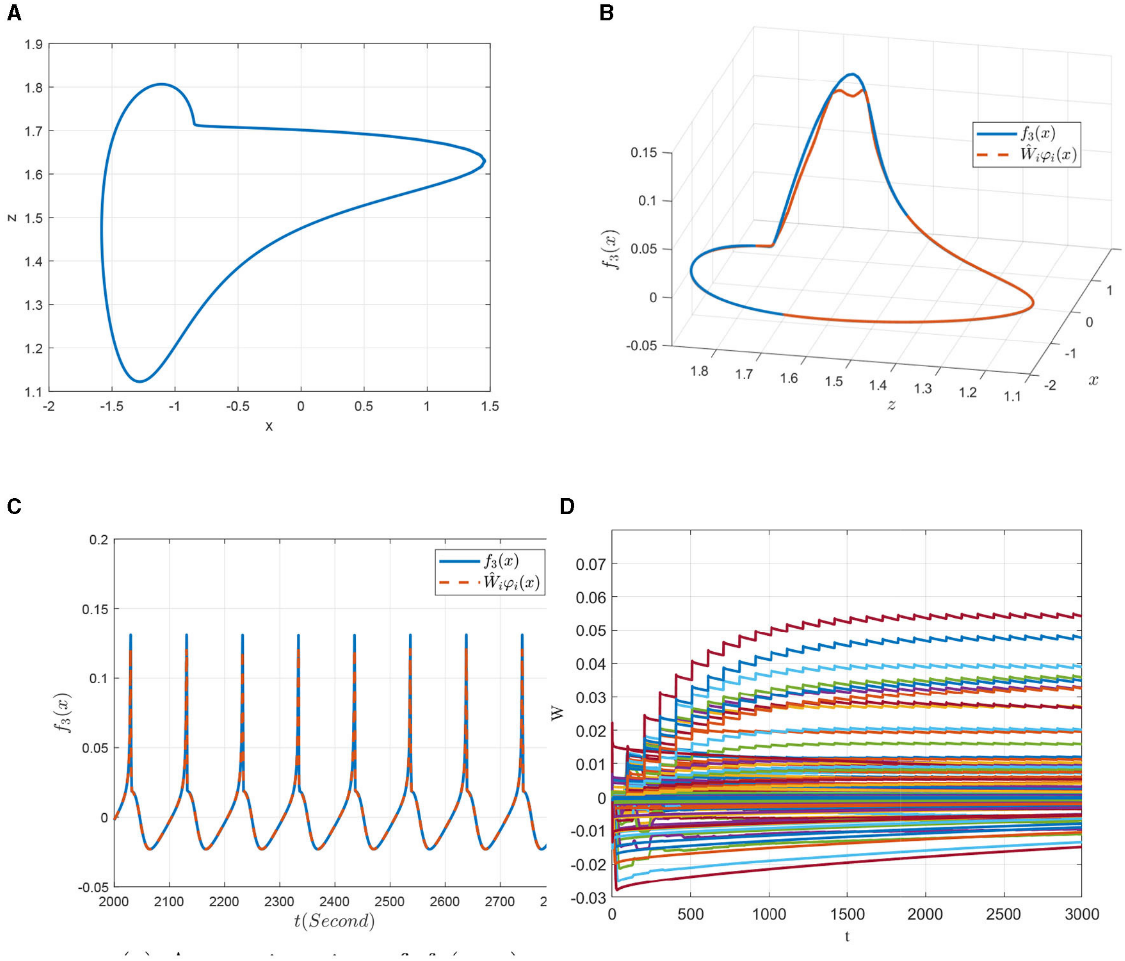


FIGURE 8
Non-linear dynamic identification of the 0.98-order HR system with $I = 1.5$ (χ^1). **(A)** State trajectory on the $x - z$ plane. **(B)** Approximation of the state trajectory. **(C)** Approximation of $f_3(x; \mu)$. **(D)** Weight convergence.

in which m denotes the number of models stored in the pattern base χ.

Remark 2: According to the recognition process discussed above, the dynamic differences between the slave system and those systems stored in the pattern base can be accurately measured without identifying of the dynamic information of the slave system. This process is therefore referred to as rapid recognition. In particular, the most similar dynamic pattern χ^{k_0} can be selected from the pattern base according to the minimum recognition error, and the dynamic information of the selected model denoted as $\bar{W}_i^{k_0 T}$ can be used to replace the unknown dynamics $f_{s,i}(x)$ of the slave FOHR model in the following control process.

Based on the recognition process, the unknown system dynamics $f_{s,i}(x)$ of the slave system can be locally accurately identified as well as stored by the constant weight NNs along the system trajectory, that is,

$$f_{s,i}(x) = \bar{W}_i^{k_0 T}\varphi_i(x) + \varepsilon_{i_2}. \tag{24}$$

Substituting Equation (24) into Equation (19), the following control rate is obtained:

$$u_i = -\eta_i sgn(s_i) - c_i e_i + f_{m,i}(x) - \bar{W}_i^{k_0 T}\varphi_i(x), \tag{25}$$

where $\bar{W}_i^{k_0 T}\varphi_i(x)$ denotes the most similar dynamic model recognized from the pattern base to the unknown slave system by using the localized RBFNNs located close to the system trajectory.

Remark 3: The mode-based sliding mode control is designed to fit the unknown dynamics of the slave system quickly by calling the acquired dynamic information of the neurons, and the experience is applied to the control process. During this process, the generalization ability of the rapid recognition mechanism based on deterministic learning provides the right decisions for invoking right dynamic patterns for better control performance. Put it another way, the empirical dynamic information learned and stored in the pattern base is so sufficiently utilized that the on-line control time is reduced and the fast synchronization is achieved. Compared with the traditional sliding mode control method, the model based sliding mode control algorithm can effectively reduce the sliding mode gain so as to reduce the chattering problem of the system.

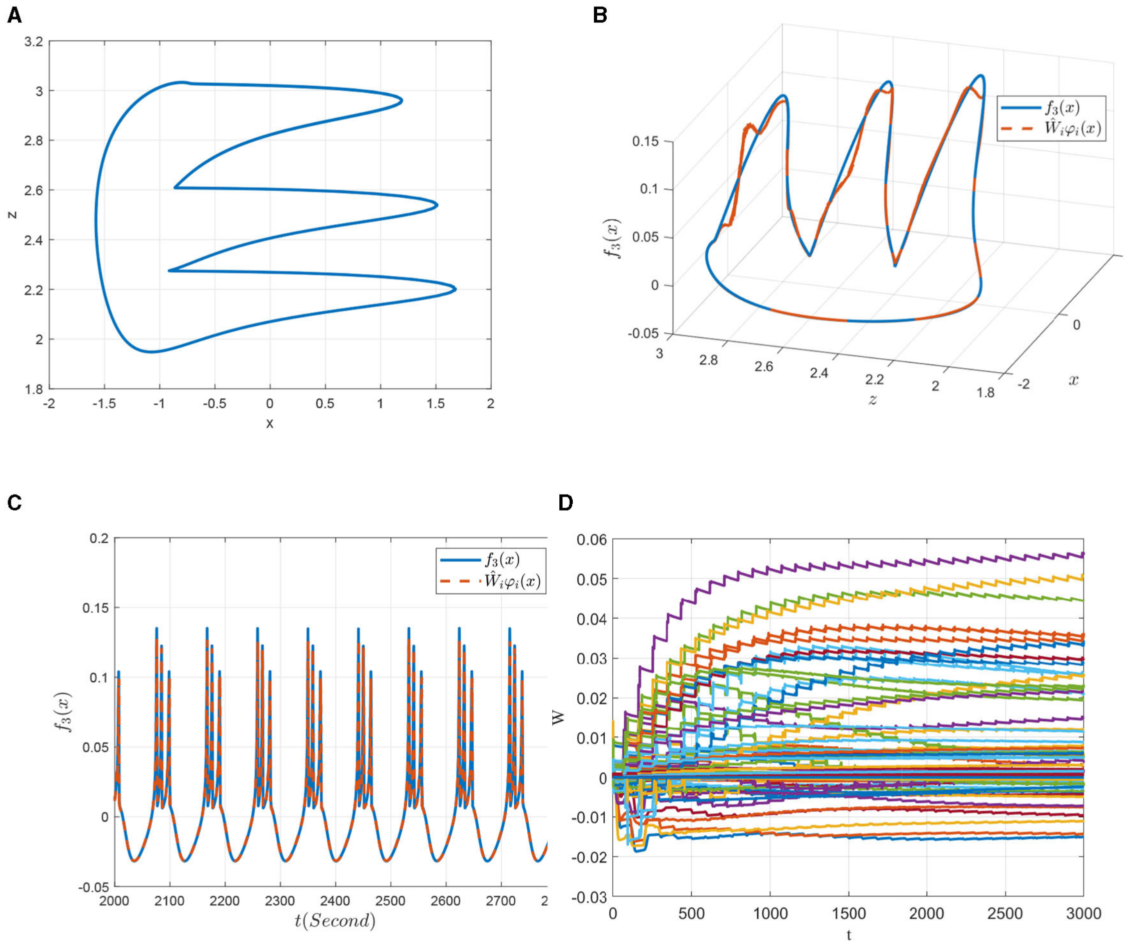


FIGURE 9
Non-linear dynamic identification of the 0.98-order HR system with $I = 2.5$ (χ^3). **(A)** State trajectory on the $x - z$ plane. **(B)** Approximation of the state trajectory. **(C)** Approximation of $f_3(x;\mu)$. **(D)** Weight convergence.

6.3. Stability analysis

To verify the stability of the master-slave synchronization control system, consider the following Lyapunov function candidate:

$$V_i = \frac{1}{2}s_i^2. \tag{26}$$

The derivative of V is

$$\dot{V}_i = s_i\dot{s}_i. \tag{27}$$

By taking the differential equation of the sliding surface given in Equation (17) and the sliding mode rate given in Equation (25) to Equation (27), we have

$$\begin{aligned}\dot{V}_i &= s_i(c_ie_i + f_{s,i}(x) + d_i + u_i - f_{m,i}(x)),\\ &= s_i(f_{s,i}(x) + d_i - \eta_i sgn(s_i) - \bar{W}_i^{k_0T}\varphi_i(x)).\\ &= s_i(\varepsilon_{i2} + d_i - \eta_i sgn(s_i)),\end{aligned} \tag{28}$$

as shown in Equation (28), the external disturbance d_i and the identification error ε_{i2} have an upper bound. Therefore, to ensure that the function V_i is negative definite, just need to set appropriate sliding mode gain η_i to make the equation $\eta_i > \varepsilon_{i2}+d_i$ work, which will further ensure the convergence of synchronization error.

6.4. Relearning-based sliding mode control of the master-slave FOHR system

As discussed above, the robustness and generalization ability of the recognition system are greatly related to the richness of the patterns in the dynamic pattern database. When considering the condition that there is no ideal similar dynamic pattern in the pattern base for the unknown slave system, that is,even if the smallest tracking error exists, the corresponding constant system dynamics $\bar{W}_i^{k_0T}\varphi_i(x)$ utilized in the control rate may result in large synchronization error and affects the stability and convergence of the control process. This analysis suggests that it is necessary to further explore how to improve the synchronization effect under limited off-line pattern base.

In order to solve the above problems to ensure a stable and rapid control effect, further identification of the unknown slave system is considered. Based on the selected dynamics $\bar{W}_i^{k_0T}\varphi_i(x)$ according to the smallest recognition error, the improved control

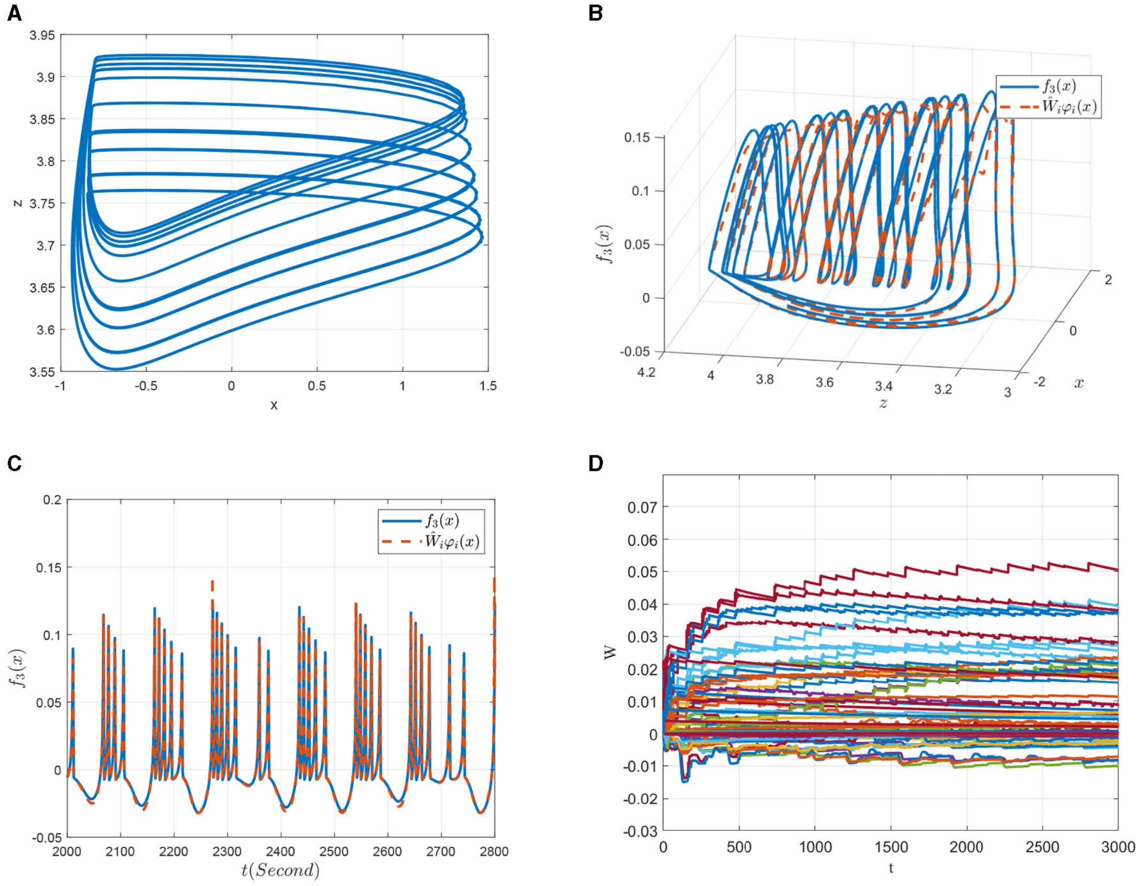


FIGURE 10
Non-linear dynamic identification of the 0.98-order HR system with $I = 3.6$ (χ^3). **(A)** State trajectory on the $x - z$ plane. **(B)** Approximation of the state trajectory. **(C)** Approximation of $f_3(x;\mu)$. **(D)** Weight convergence.

rate is proposed below:

$$\begin{aligned} u_i &= -\eta_i sgn(s_i) - c_i e_i + f_{m,i}(x) - \hat{W}_i^T\varphi_i(x), \\ u_i(0) &= -\eta_i sgn(s_i) - c_i e_i + f_{m,i}(x) - \bar{W}_i^{k_0 T}\varphi_i(x). \end{aligned} \tag{29}$$

Based on the Equation (29) and Equation (17), the time derivative of the sliding mode variable is given as

$$\dot{s}_i = -\tilde{W}_i^T\varphi_i(x) + \varepsilon_i - \eta_i sgn(s_i) + d_i, \tag{30}$$

where $-\tilde{W}_i^T\varphi_i(x) = f_{s,i}(x) - \hat{W}_i^T\varphi_i(x)$, $|d_i(t)| \leq D_i$, and $|\varepsilon_i| \leq \bar{\varepsilon}$ are external excitation and identification error with upper bound, respectively. For convenience of presentation, define $D_i + \bar{\varepsilon}_i = \kappa_i$, the derivation of the sliding mode surface is given as follows:

$$\dot{s}_i = -\tilde{W_i}^T\varphi_i(x) - \eta_i sgn(s_i) + \kappa_i. \tag{31}$$

In addition, the NN adaptive update law of the sliding mode control is designed as

$$\dot{\hat{W}}_i = \dot{\tilde{W}}_i = \Gamma_i\varphi_i(x)s_i - \sigma_i\Gamma_i|s_i|\hat{W}_i, \tag{32}$$

where Γ_i and σ_i are positive adjustable parameters. Since $\kappa_i = D_i + \bar{\varepsilon}$, the synchronization error is precisely related to the identification accuracy; that is, the higher the identification accuracy of the unknown slave system, the better the synchronization effect of the master-slave neuronal system.

Theorem 1 Consider the master-slaver neuron FOHR system as shown in Equation (14), the learning-based controller Equation (29), and the NN weight updating law Equation (32). For initial condition $x_d(0)$ which generates the recurrent orbit $\varphi_d(x_0)$, and with corresponding initial condition $x(0)$ selected in a close vicinity of the recurrent orbit, the control error of the master-slave system described by Equation (15) converges exponentially to a small neighborhood around zero.

Proof: For the sliding mode-based control system, consider the following Lyapunov function:

$$V = \frac{1}{2}s_i^2 + \frac{1}{2}\tilde{W}_i^T\Gamma^{-1}\tilde{W}_i. \tag{33}$$

The derivative of V is

$$\dot{V}_i = s_i\dot{s}_i + \tilde{W}_i^T\Gamma_i^{-1}\dot{\tilde{W}}_i. \tag{34}$$

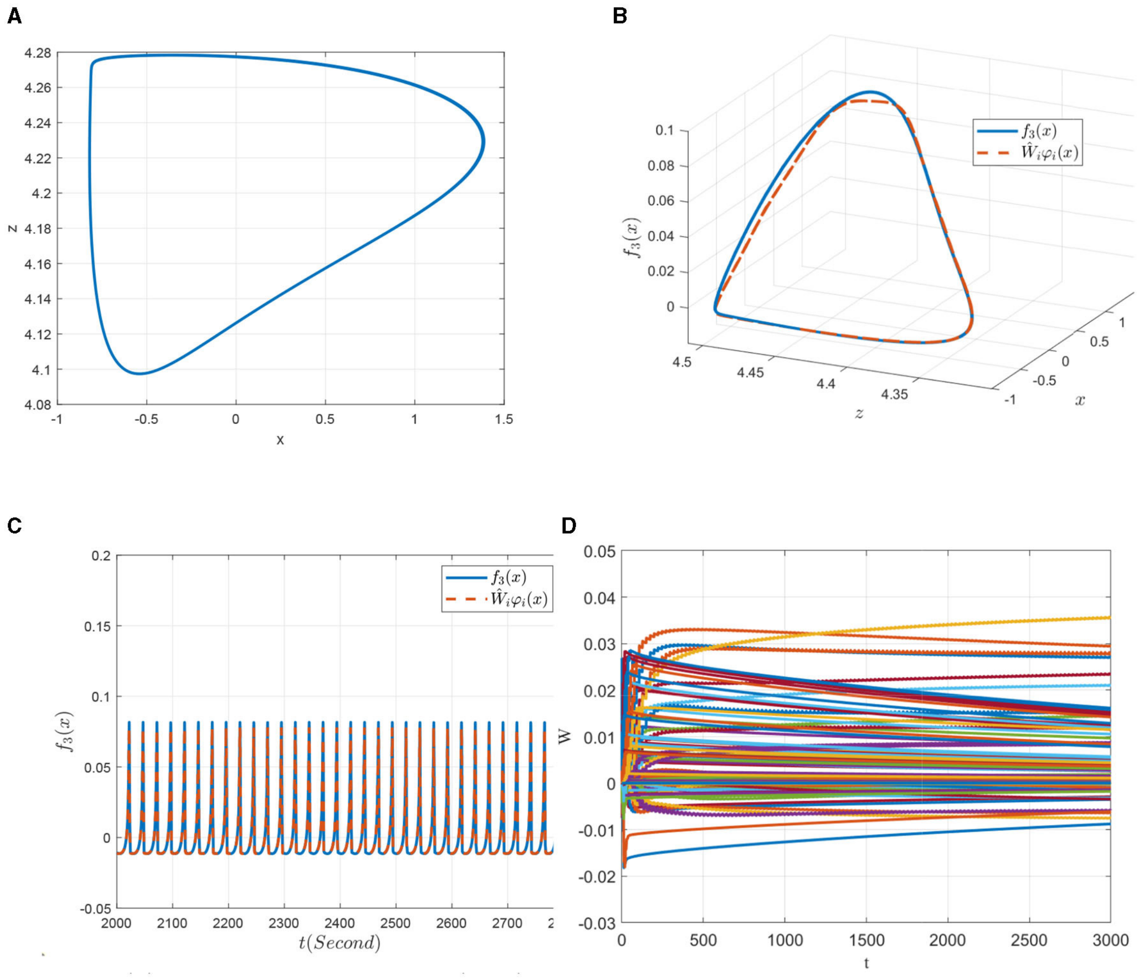


FIGURE 11

Non-linear dynamic identification of the 0.98-order HR system with $I = 4$ (χ^4). **(A)** State trajectory on the $x - z$ plane. **(B)** Approximation of the state trajectory. **(C)** Approximation of $f_3(x;\mu)$. **(D)** Weight convergence.

By introducing the designed sliding mode surface and the adaptive update rate equation, there is,

$$\begin{aligned}\dot{V}_i &= s_i(-\tilde{W}_i^T\varphi_i(x) - \eta_i sgn(s_i) + \kappa_i) + \tilde{W}_i^T\Gamma_i^{-1}(\Gamma_i\varphi_i(x)s_i \\ &\quad - \sigma_i\Gamma_i|s_i|\hat{W}_i), \\ &= s_i\kappa_i - s_i\eta_i sgn(s_i) - \sigma_i\tilde{W}_i^T|s_i|\hat{W}_i, \\ &\le |s_i|(\kappa_i - \eta_i - \sigma_i\tilde{W}_i^T\hat{W}_i),\end{aligned} \tag{35}$$

in which,

$$\begin{aligned}-\sigma_i\tilde{W}_i^T\hat{W}_i &\le -\sigma_i\|\tilde{W}_i\|^2 + \sigma_i\|\tilde{W}_i\|\|W_i^*\|, \\ &\le -\frac{\sigma_i}{2}\|\tilde{W}_i\|^2 + \frac{\sigma_i}{2}\|\bar{\bar{W}}_i\|^2,\end{aligned} \tag{36}$$

with $\bar{\bar{W}}_i$ being the upper bound of the ideal identification NN weight W_i^*. Thus, it follows that

$$\dot{V}_i \le |s_i|(\kappa_i - \eta_i - \frac{\sigma_i}{2}\|\tilde{W}_i\|^2 + \frac{\sigma_i}{2}\|\bar{\bar{W}}_i\|^2). \tag{37}$$

It is clear that $\dot{V}$ is negative definite when the following conditions are met:

$$|\eta_i| > \frac{\sigma_i}{2}\|\bar{\bar{W}}_i\|^2 + \kappa_i \; or \; \|\tilde{W}_i\| > \frac{\sigma_i}{2}\|\bar{\bar{W}}_i\| + \sqrt{\frac{2\kappa_i}{\sigma_i}}. \tag{38}$$

Since the ideal identification NN weight W_i^*, the external excitation d_i and the estimate error ε_i are all upper bounded; therefore, all signals in a closed-loop control system remain bounded, including the estimate NN weight $\hat{W}_i$ and the sliding mode variable s_i.

In addition, to the convergence of the sliding mode variable, the following Lyapunov function is given as

$$V = \frac{1}{2}s_i^2. \tag{39}$$

The corresponding derivative is given as

$$\begin{aligned}\dot{V}_i &= s_i\dot{s}_i, \\ &= s_i(-\tilde{W}_i^T\varphi_i(x) - \eta_i sgn(s_i) + d_i + \varepsilon_i), \\ &\le -|s_i|(\eta_i + \tilde{W}_i^T\varphi_i(x) - \kappa_i).\end{aligned} \tag{40}$$

Considering that the Gauss function $\varphi_i(x)$ and $-\tilde{W}_i^T\varphi_i(x) + \kappa_i$ are both bounded, when the gain η_i satisfies the condition that $\eta_i > -\tilde{W}_i^T\varphi_i(x) + \kappa_i$, there is

$$\dot{V}_i \le -\gamma_i|s_i| = -\gamma_i\sqrt{V_i}, \tag{41}$$

where $\gamma_i = \eta_i + \tilde{W}_i^T\varphi_i(x) - \kappa_i$. As long as the parameter η_i is reasonably designed, the convergence of the tracking error

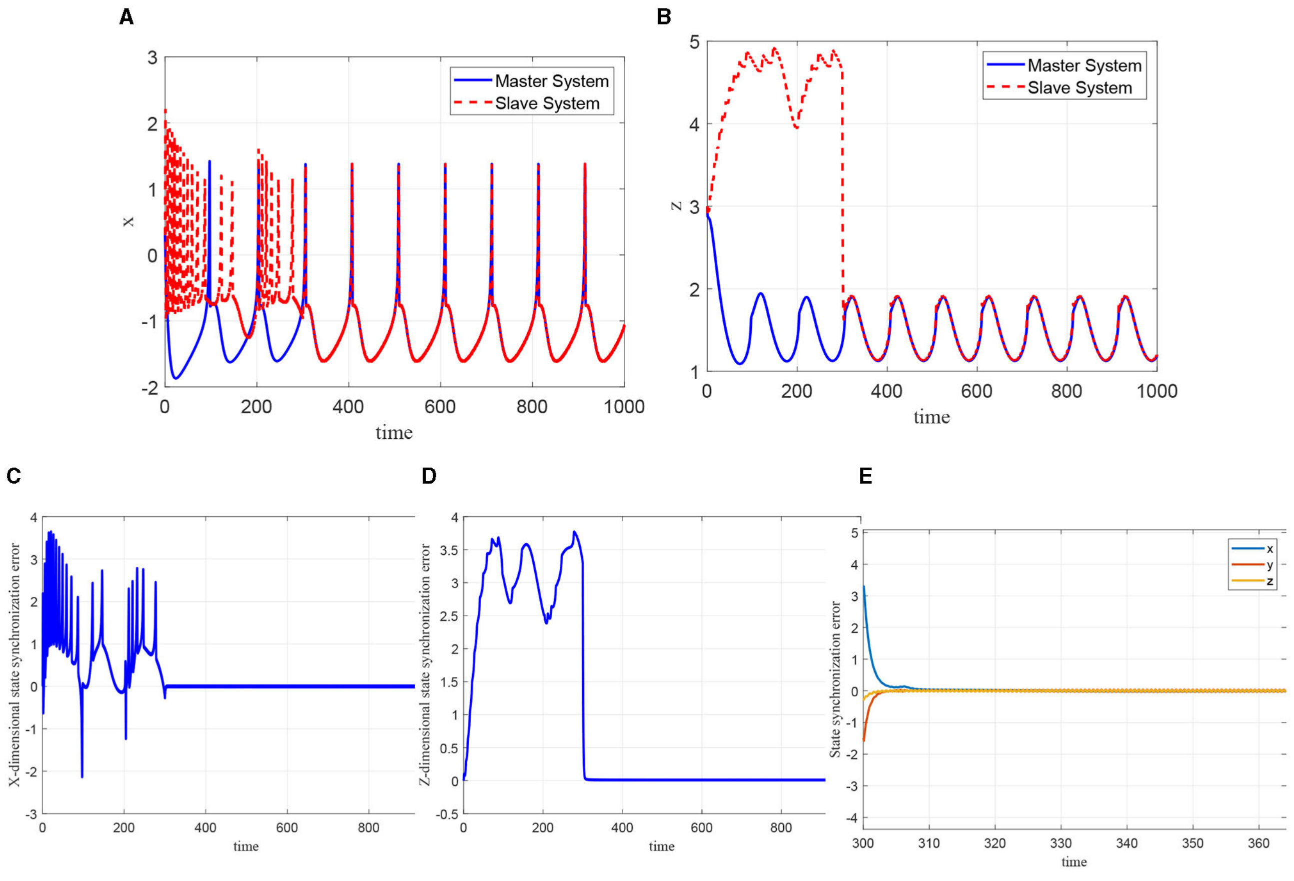


FIGURE 12
Synchronization of the master-slave neuron system. **(A)** Synchronization of *x*. **(B)** Synchronization of *z*. **(C)** Synchronization error of *x*. **(D)** Synchronization error of *z*. **(E)** Synchronization error of *x*, *y*, *z*.

is ensured, and the sliding mode variable do converge to some neighborhood of zero. In addition, the size of the convergence neighborhood depends on the control parameter; that is, by properly design the control parameters, ideal synchronization control performance can be achieved.

Remark 4: According to the relearning-based sliding mode control algorithm given above, if there is no dynamic pattern that is sufficiently similar to the unknown slave system in the pattern base χ, the on-line identification process for the unknown slave system is started. Different from the initial identification process, the initial weights of the neural network during the identification process for unfamiliar synchronization objects are taken from the constant weight of the dynamic system corresponding to the minimum recognition error rather than iterating from zero. Thus, the learned and stored dynamic information help reduce the on-line identification time. Additionally, the identified dynamic information of the slave system will be restored in the form of constant weights and can further utilized to new synchronization problems. This process will help enrich the empirical dynamics information of the pattern base to improve the accuracy and efficiency of the new synchronization tasks.

7. Simulation research

To verify the effectiveness of the control strategy proposed in this study, simulations of the master-slave FOHR system under unknown dynamic environment are conducted.

7.1. Identification of the unknown dynamics of the FOHR system

In this part, the identification of the FOHR system shown in Equation (6) under unknown dynamic environment is considered. For the convenience of presentation, the system state x, y, and z are denoted as x_1, x_2, and x_3, respectively. The corresponding state vector $x = [x_1, x_2, x_3]^T \in R^3$ of the FOHR model is available from measurement and the parameter $\mu = [a, b, c, d, r, s_0, q_0]^T$ is taken as a constant vector and chosen as $a = 1, b = 3, c = 1, d = 6, r = 0.013, s_0 = 4, q_0 = -1.56$. As demonstrated in Section 3, by varying the fractional order parameter q and fixing all the other parameters unchanged, the FOHR system presents

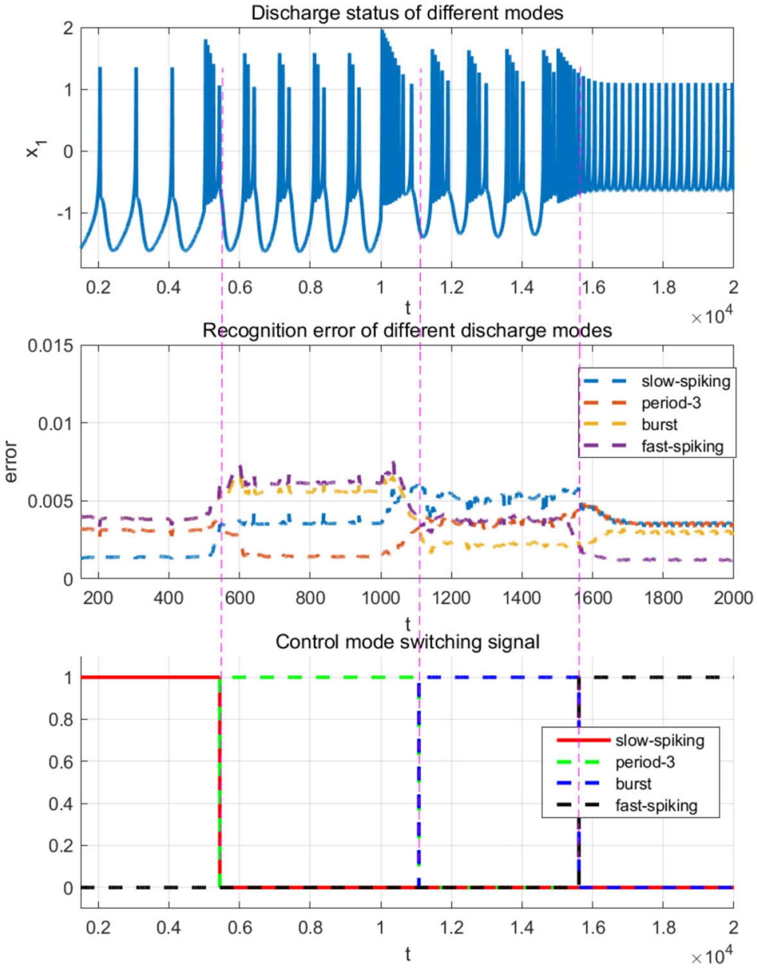


FIGURE 13
Switching control based on recognition error.

diverse non-linear behaviors. Moreover, the 0.98-order HR model can best describe the abundant non-linear dynamic characteristic of neurons. Thus, the 0.98-order HR system is considered for dynamic identification with the external excitation I being taken as the control parameter.

To verify the identification effects, four kinds of representative discharge models of the 0.98-order HR system with the parameters given above are chosen, that is, the slow-spiking model χ^1 with $I = 1.5$, the period-3 bursting model χ^2 with $I = 2.5$, the chaotic bursting model χ^3 with $I = 3.6$ and the fast-spiking model χ^4 with $I = 4$. The dynamic analysis about FOHR system have demonstrated that the corresponding state trajectories of the four dynamic models mentioned above possess regression properties. Thus, the DL algorithm is introduced for the unknown dynamic identification process.

According to the DL algorithm, the dynamical RBF network $\dot{\hat{x}} = -A(\hat{x} - x) + \hat{W}\phi(x)$ is employed to identify the unknown system dynamics $f_i(x;\mu)(i = 1, 2, 3)$ as shown in Equation (6). For the space limitation, the unknown dynamic $f_3(x;\mu) = r(s_0(x - q_0) - z)$ is taken as an example to show the identification effects. The center of the neural network is evenly placed on $[-2.1, 2.1] \times [0.9, 5.1]$ and the widths are set as $\eta_i = 0.3$. The weights of the RBF networks are updated online according the equation $\dot{\hat{W}}_i = \dot{\tilde{W}}_i = -\Gamma_i\varphi_i(x)\tilde{x}_i - \sigma_i\Gamma_i\hat{W}_i$, within which the parameters are chosen as $\Gamma_i = diag\{2, 2, 2\}$, $\sigma_i = 0.0001, i = 1, 2, 3$ and a_3 from $A = [a_1, a_2, a_3]^T$ is set as $a_3 = 10$. The initial condition of the dynamical system is set as $[x_1(0), x_2(0), x_3(0)]^T = [0.3, 1, 3]^T$, $[\hat{x}_1(0), \hat{x}_2(0), \hat{x}_3(0)]^T = [0.2, 0.3, 0.0]^T$, and the initial weights are $\hat{W}_i(0) = 0.0$.

First, the 0.98-order HR system with external excitation $I = 1.5$ denoted in a slow-spiking model as χ^1 is to be identified. Figure 8A is the projection of the state trajectory of the slow-spiking model on the $x - z$ plane. In Figures 8B, C, it is seen that the state trajectory can be accurately identified by using the DL algorithm. More importantly, in addition to the state tracking, the NN approximation of the system dynamics $f_3(x;\mu)$ along the system trajectory is shown in Figure 8C. The convergence of the weights of the RBF neural network is further obtained from the Figure 8D. That is, by introducing the DL algorithm, the unknown dynamic information $f_3(x;\mu)$ of the FOHR model is locally accurately approximated by $\hat{W}_i\varphi_i(x)$, and the identified non-linear dynamic information can be further stored in the constant weights of networks given as $\bar{W}_i\varphi_i(x)$.

Second, similar results are obtained for the identification of the non-linear dynamics of the 0.98-order HR system with $I = 2.5$ that exhibiting a period-3 bursting model denoted as χ^2. It can be seen from the Figure 9A that the non-linear dynamics of the period-3 bursting model are richer than that of the dynamics presented in Figure 8A. Even though, ideal approximation effects of both the system state and the unknown system function are obtained as demonstrated in Figures 9B, C. The parameters of the corresponding RBF networks also converge to an ideal value, which can be seen from Figure 9D.

Third, consider the identification of the dynamics of model χ^3 with $I = 3.6$, as shown in Figure 10. The system state given in Figure 10A presents a complex state of chaos, which contains more dynamic information of the FOHR system. By properly designing the identification parameters, locally accurate NN approximations of the system state as well as the unknown system dynamics are achieved along the system trajectory, which can be seen from Figures 10B, C. In addition, it is noticed from the Figure 10D that more neurons are involved and activated in the identification of the chaotic bursting model χ^3. Moreover, the oscillation of the NN weights during the convergence process is so obvious that more time is needed for it converge to the ideal values.

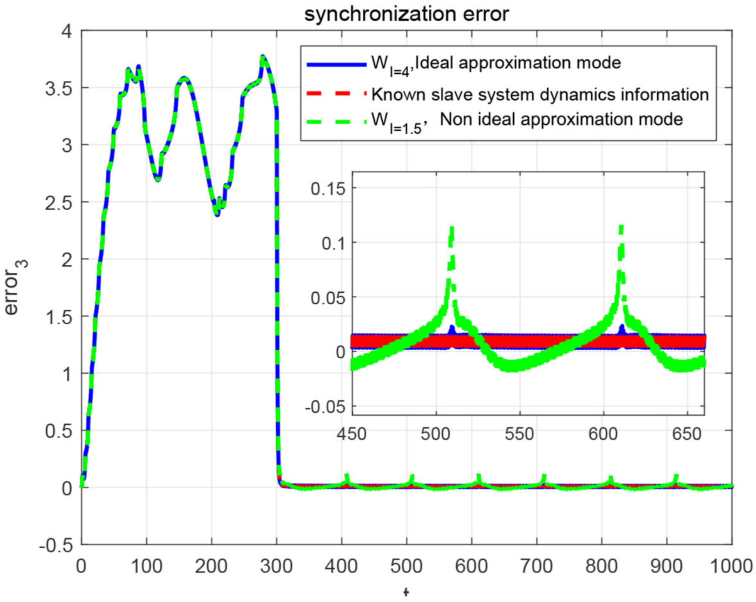


FIGURE 14
Synchronization error of master-slave neuron system with different modes.

Finally, a further increase in the external excitation I to 4 (denoted as model χ^4), the system returns back to a simple discharge state. As can be seen from Figure 11A, the state trajectory of model χ^4 is a typical period-1 behavior, but the discharge rhythm is faster compared to that of the model χ^1 shown in Figure 8A. As for its identification simulations demonstrated in Figures 11B–D, it is shown that it achieves better state and dynamic tracking effects, and the parameter convergence process is much smooth and faster.

7.2. Model-based sliding-mode control of the FOHR system

Based on the acquisition and storage of the unknown dynamic information of the FOHR system, the rapid recognition of the FOHR model is demonstrated in this part. The dynamic models $\chi^{1,2,3,4}$ mentioned above are taken as the training patterns. The testing patterns are generated from the FOHR system presented in Equation (7), with $I = 1.43$ denoted as χ^5, $I = 2.3$ denoted as χ^6, $I = 3.4$ denoted as χ^7, and $I = 4.2$ denoted as χ^8. The other parameters are set as the same to the training patterns, that is, $q = 0.98, a = 1, b = 3, c = 1, d = 6, r = 0.013, s = 4$, and $q_0 = -1.56$. For the recognition process, the dynamic NN network system is introduced, that is,

$$\dot{\tilde{x}}_i = -\bar{b}_i\tilde{x}_i + (\bar{W}_i^{kT}\varphi_i(x) - f_i'(x; \mu')), i = 1, \cdots, n \qquad (42)$$

for which, the initial states is given as $[x_0, y_0, z_0]^T = [0.3, 1, 3]^T$ and $[\tilde{x}_0, \tilde{y}_0, \tilde{z}_0]^T = [0, 0, 0]^T$.

Based on the obtained dynamic pattern database χ, which contains the learned system dynamics as experience of the slaw peak regular spiking model, period-doubling, period-3, period-4 bursting model, and chaotic bursting model, the simulation of the learning-based sliding-mode control of the master-slave neural system is discussed in this part. The corresponding parameters are given as $\eta_i = 1, c_i = 1, \Gamma_i = 2$, and $\sigma_i = 0.01, (i = 1, \cdots, n)$. The external disturbances are set as $d_1(t) = 0.6 + 0.2cos(t), d_2(t) = 0.0, d_3(t) = 0.01 + 0.05sin(t)$, and the other parameters of the master-slave system are given as the same as shown in the previous section. The external stimulus current of the master system is set as $I = 1.5$, while for the slave system, the external stimulus current is set as $I = 3.8$. The other parameters are designed as $q = 0.98, a = 1, b = 3, c = 1, d = 6, r = 0.013, s = 4$, and $q_0 = -1.56$, the initial state of the master-slave system is given as $[x_0, y_0, z_0]^T = [0.313]^T$, and the control will be added at $t = 300ms$.

As can be seen from the Figures 12A, B, when the control quantity is added to the slave system at $t = 300ms$, the state of the master-slave neurons can quickly reach consistency, and the selected NN controller achieves good synchronization to the master neuron system. Moreover, the synchronization error demonstrated in Figures 12B–D shows that the FOHR master-slave neuronal system achieves fast synchronization performance.

Since the external excitation of the master and slave system are set as $I = 1.5$ and $I = 3.8$, respectively, it means that the master system is in slow-spiking state and the slave system is in a state of rapid-peak spiking, as described in the identification phase. For accurate synchronization effect, the rapid-peak spiking model shown be recalled from the pattern base χ^4, which can be validated from the Figure 13.

In addition, through the simulation comparison by recalling the rapid-peak spiking model, the ideal known dynamic model corresponding to the slave system and the slow-spiking model, respectively, the synchronization errors are shown in Figure 14. It demonstrate that in terms of convergence speed, accuracy, and buffeting size, the more accurate the dynamic model is selected, the better the synchronization effect will be. It further indicates that the performance of the sliding-mode control algorithm is highly related to the dynamic information accuracy of the invoked dynamic models.

7.3. Relearning-based sliding-mode control of the FOHR system

To verify the effectiveness of the relearning-based sliding-mode control performance to the master-slave neuronal system, the third dimension dynamics of the neuron system is taken as an example,

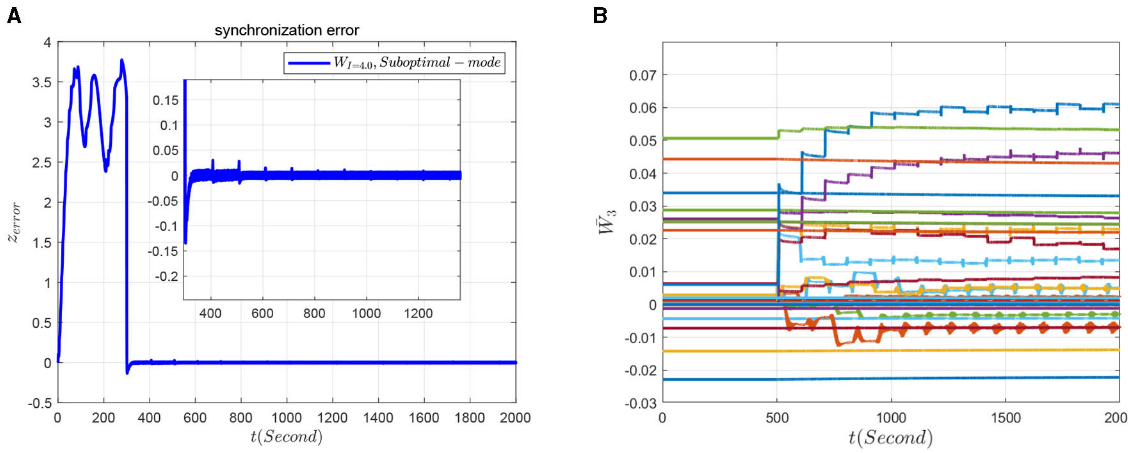


FIGURE 15
Synchronization effect of the master-slave neuron system based on relearning control. **(A)** Synchronization error. **(B)** Convergence of the NN weight.

and the sliding-mode controller is set as

$$\begin{aligned} &\mu_3 = -\eta_3 sgn(s_3) - c_3 e_3 + f_{m,3}(x) - \hat{W}_3^T \varphi_3(x), \\ &\mu_3(0) = -\eta_i sgn(s_3) - c_i e_3 + f_{m,3}(x) - \bar{W}_3^{K_0 T} \varphi_3(x), \end{aligned} \tag{43}$$

in which the initial NN weight is set as $\hat{W}_3(0) = \bar{W}_3$, with $\bar{W}_3$ being the constant NN weight. During the control process, the model-based sliding mode controller is added to the system at 300 ms and at 500 ms switch to the relearning-based sliding mode controller. The synchronous response of the system can be seen from Figure 15A. Furthermore, it can be seen from Figure 15A, when the system switch to the relearning-based sliding mode control policy, the synchronization error is getting smaller because of more accurate identification of the dynamics of slave system, and the cusp error is obviously improved. In addition, the NN weight of the relearning process can convergence to ideal values as shown in Figure 15B.

8. Conclusion

Aiming at the problem of abnormal synchronization of fractional-order Hindmarsh-Rose (FOHR) neuronal system in unknown dynamic environment, the identification, rapid recognition, and synchronization control of the unknown dynamic FOHR system is discussed in this study. For accurate synchronization of the FOHR neuronal system, the unknown dynamic information has been identified by using the deterministic leaning theory. Based on the achieved system dynamics, the unknown different dynamic patterns generated from the FOHR system can be rapidly recognized without relearning process. In addition, the achieved dynamic information has been applied to the sliding mode controller, resulting in more accurate and efficient synchronization performance of the master-slaver neuronal system. From system identification to pattern construction, then to model-based and relearning-based sliding mode control, this study emphasizes the whole linkage process, which kindly displays the human experience of learning and application of unknown knowledge, which is the essence of intelligent learning and intelligent control.

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

DC, JL, and CY contributed to the methodology, concept, and design of the study. DC, JL, and JH performed the experiments and simulation analysis. JL, JH, and WZ prepared the draft manuscript. All authors participated manuscript organization and approved the submitted version.

Funding

This study was supported by the National Natural Science Foundation of China (Grant No. 62106048) by Guangdong Natural Science Foundation (Grant No. 2022B1515120003) and by Guangdong Provincial Laboratory of Marine Science and Engineering (Grant No. SML2022SP101).

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

Boaretto, B. R. R., Budzinski, R. C., Prado, T. L., Kurths, J., and Lopes, S. R. (2018). Neuron dynamics variability and anomalous phase synchronization of neural networks. *Chaos* 28, 106304. doi: 10.1063/1.5023878

Brown, P., Mazzone, P., Oliviero, A., Maria, G. A., Fabio, P., Pietro, A. T., et al. (2004). Effects of stimulation of the subthalamic area on oscillatory pallidal activity in Parkinson's disease. *Exp. Neurol.* 188, 480–490. doi: 10.1016/j.expneurol.2004.05.009

Che, Y. Q., Wang, J., Tsang, K. M., and Chan, W. L. (2010). Unidirectional synchronization for Hindmarsh-Rose neurons via robust adaptive sliding mode control. *Nonlinear Anal. Real World Appl.* 11, 1096–1104. doi: 10.1016/j.nonrwa.2009.02.004

Chen, D., and Wang, C. (2016). Prediction of period-doubling bifurcation based on dynamic recognition and its application to power systems. *Int. J. Bifurc. Chaos* 26, 1650157. doi: 10.1142/S0218127416501571

Chen, D., Zhang, R., Ma, X., and Liu, S. (2012). Chaotic synchronization and anti-synchronization for a novel class of multiple chaotic systems via a sliding mode control scheme. *Nonlinear Dyn.* 69, 35–55. doi: 10.1007/s11071-011-0244-7

Dar, M. R., Kant, N. A., and Khanday, F. A. (2022). Dynamics and implementation techniques of fractional-order neuron models: a survey. *Fract. Order Syst.* 1, 483–511. doi: 10.1016/B978-0-12-824293-3.00017-X

Deng, B., Wang, J., and Fei, X. (2006). Synchronizing two coupled chaotic neurons in external electrical stimulation using backstepping control. *Chaos Solitons Fractals* 29, 182–189. doi: 10.1016/j.chaos.2005.08.027

Ding, D., Chen, X., Yang, Z., Hu, Y., Wang, M., Zhang, H., et al. (2022). Coexisting multiple firing behaviors of fractional-order memristor-coupled HR neuron considering synaptic crosstalk and its ARM-based implementation. *Chaos Solitons Fractals* 158, 112024. doi: 10.1016/j.chaos.2022.112014

Dong, J., Zhang, G., Xie, Y., Yao, H., and Wang, J. (2014). Dynamic behavior analysis of fractional-order Hindmarsh–Rose neuronal model. *Cogn. Neurodyn.* 8, 167–175. doi: 10.1007/s11571-013-9273-x

Duan, Y. (2002). *Nonlinear Dynamics in Interspike Intervals of an Experimental Neural Pacemaker.* Fourth Military Medical University.

Efe, M. O. (2009). Adaline based robust control in robotics: a Riemann-Liouville fractional differintegration based learning scheme. *Soft Comput.* 13, 23–29. doi: 10.1007/s00500-008-0289-9

Ermentrout, B. (2014). Linearization of $f - I$ curves by adaptation. *Neural Comput.* 10, 1721–1729. doi: 10.1162/089976698300017106

Fitzhugh, R. (1961). Impulses and physiological states in models of nerve membrane. *Biophys. J.* 1, 455.

Giresse, T. A., Crepin, K. T., and Martin, T. (2019). Generalized synchronization of the extended Hindmarsh-Crose neuronal model with fractional order derivative. *Chaos Solitons Fractals* 118, 311–319. doi: 10.1016/j.chaos.2018.11.028

Gorenflo, R., and Mainardi, F. (1997). *Fractional Calculus*. Springer. Available online at: https://www.researchgate.net/publication/216225140

Gu, H., Pan, B., Chen, G., and Duan, L. (2014). Biological experimental demonstration of bifurcations from bursting to spiking predicted by theoretical models. *Nonlinear Dyn.* 78, 391–407. doi: 10.1007/s11071-014-1447-5

Guo, Z. H., Li, Z. J., Wang, M. J., and Ma, M. L. (2023). Hopf bifurcation and phase synchronization in memristor-coupled Hindmarsh–Rose and FitzHugh–Nagumo neurons with two time delays. *Chin. Phys. B* 32,038701. doi: 10.1088/1674-1056/aca601

Hindmarsh, J. L., and Rose, R. M. (1984). A model of neuronal bursting using three coupled first order differential equations. *Proc. R. Soc. Lond. Ser. B.* 221, 87–102.

Hodgkin, A. L., and Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol.* 117, 500–544.

Huang, C. (2016). *Dynamical Analysis and Control for Several Classes of Fractional Systems*. Southeast University. doi: 10.7666/d.Y3186252

Jia, B., Gu, H., Li, L., and Zhao, X. (2012). Dynamics of period-doubling bifurcation to chaos in the spontaneous neural firing patterns. *Cogn Neurodyn.* 6, 89–106. doi: 10.1007/s11571-011-9184-7

Jia, B., Gu, H., and Xue, L. (2017). A basic bifurcation structure from bursting to spiking of injured nerve fibers in a two-dimensional parameter space. *Cogn. Neurodyn.* 11, 189–200. doi: 10.1007/s11571-017-9422-8

Jin, T., Gao, S., Xia, H., and Ding, H. (2021). Reliability analysis for the fractional-order circuit system subject to the uncertain random fractional-order model with Caputo type. *J. Adv. Res.* 4, 15–26. doi: 10.1016/j.jare.2021.04.008

Li, X. X., Xue, X. P., Liu, D. J., Yu, T. Y., He, Q. Q., et al. (2023). Effects of electric field on vibrational resonances in Hindmarsh-Rose neuronal systems for signal detection. *Chin. Phys. B* 32, 048701. doi: 10.1088/1674-1056/ac9cc0

Lin, P., Wang, M., and Wang, C. (2019). Abrupt stall detection for axial compressors with non-uniform inflow via deterministic learning. *Neurocomputing* 338, 163–171. doi: 10.1016/j.neucom.2019.02.007

Liu, D., Zhao, S., Luo, X., and Yuan, Y. (2019). Unidirectional synchronization of Hodgkin-Huxley neurons with prescribed performance under transcranial magneto-acoustical simulation. *Front. Neurosci.*, 13, 1061. doi: 10.3389/fnins.2019.01061

Liu, Y., Sun, Z., Yang, X., and Xu, W. (2021). Rhythmicity and firing modes in modular neuronal network under electromagnetic field. *Nonlinear Dyn.* 104, 4391–4400. doi: 10.1007/s11071-021-06470-8

Meng, F., Zeng, X., Wang, Z., and Wang, X. (2020). Adaptive synchronization of fractional-order coupled neurons under electromagnetic radiation. *Int. J. Bifurc. Chaos.* 30, 2050044. doi: 10.1142/S0218127420500443

Motallebzadeh, F., Motlagh, M. R. J., and Cherati, Z. R. (2012). Synchronization of different-order chaotic systems: adaptive active vs. optimal control. *Commun. Nonlinear Sci. Num. Simul.* 17, 3643–3657. doi: 10.1016/j.cnsns.2012.01.012

Nirvin, P., Rihan, F. A., Rakkiyappan, R., and Pradeep, C. (2021). Impulsive sampled-data controller design for synchronization of delayed T–S fuzzy Hindmarsh–Rose neuron model. *Math. Comput. Simul.* 201, 588–602. doi: 10.1016/j.matcom.2021.03.022

Parastesh, F., Azarnoush, H., Jafari, S., Hatef, B., and Perc, M. (2019). Repnik R Synchronizability of two neurons with switching in the coupling. *Appl. Math. Comput.* 350, 217–223. doi: 10.1016/j.amc.2019.01.011

Rabah, K., Ladaci, S., and Lashab, M. (2017). A novel fractional sliding mode control configuration for synchronizing disturbed fractional-order chaotic systems. *Pramana.* 89, 1–13. doi: 10.1007/s12043-017-1443-7

Rabinovich, M. I., and Abarbanel, H. D. (1998). The role of chaos in neural systems. *Neuroscience* 87, 5–14. doi: 10.1016/S0306-4522(98)00091-8

Rajagopal, K., Khalaf, A. J. M., Parastesh, F., Moroz, I., and Jafari, S. (2019). Dynamical behavior and network analysis of an extended Hindmarsh-Crose neuron model. *Nonlinear Dyn.* 98, 477–487. doi: 10.1007/s11071-019-05205-0

Remi, T., Subha, P. A., and Usha, K. (2022). Collective dynamics of neural network with distance dependent field coupling. *Commun. Nonlinear Sci. Num. Simulat.* 110, 106390. doi: 10.1016/j.cnsns.2022.106390

Rihan, F. A., Al-Mdallal, Q. M., AlSakaji, H. J., and Hashish, A. (2019). A fractional-order epidemic model with time-delay and nonlinear incidence rate. *Chaos Solitons Fractals* 126, 97–105. doi: 10.1016/j.chaos.2019.05.039

Schoenberg, I. J. (1938). Metric spaces and completely monotone functions. *Ann. Math.* 39, 811–841.

Semenov, D. M., and Fradkov, A. L. (2021). Adaptive synchronization in the complex heterogeneous networks of Hindmarsh–Rose neurons. *Chaos Solitons Fractals* 150, 111170. doi: 10.1016/j.chaos.2021.111170

Simo, G.R., Njougouo, T., Aristides, R.P., Louodop, P., Tchitnga, R., et al. (2021). Chimera states in a neuronal network under the action of an electric field. *Phys. Rev. E.* 103, 062304. doi: 10.1103/PhysRevE.103.062304

Uhlhaas, P.J., Linden, D. E. J., Singer, W., Haenschel, C., Lindner, M., Maurer, K., et al. (2006). Dysfunctional long-range coordination of neural activity during gestalt perception in schizophrenia. *J. Neurosci.* 26, 8168–8175. doi: 10.1523/JNEUROSCI.2002-06.2006

Vafaei, V., Kheiri, H., and Akbarfam, A. J. (2019). Synchronization of fractional-order chaotic systems with disturbances via novel fractional-integer integral sliding mode control and application to neuron models. *Math. Methods Appl. Sci.* 42, 2761–2773. doi: 10.1002/mma.5548

Wang, C., Chen, T., Chen, G., and Hill, D.J. (2009). Deterministic learning of nonlinear dynamical systems. *Int. J. Bifurc. Chaos* 1, 1307–1328. doi: 10.1142/S0218127409023640

Wang, C., and Hill, D. J. (2018). *Deterministic Learning Theory: For Identiflcation, Recognition, and Conirol*. CRC Press. doi: 10.1201/9781420007763

Wang, S., He, S., Yousefpour, A., Jahanshahi, H., Repnik, R., and Perc, M. (2020). Chaos and complexity in a fractional-order financial system with time delays. *Chaos Solitons Fractals* 131, 109521. doi: 10.1016/j.chaos.2019.109521

Xu, J., Li, N., Zhang, X., and Qin, X. (2020). Fuzzy synchronization control for fractional-order chaotic systems with different structures. *Front. Phys.* 8, 155. doi: 10.3389/fphy.2020.00155

Yang, H., Rong, G., Huang, S., Cui, S., Wang, M., et al. (2021). Research on the effects of neural network damage on neuronal firing patterns and synchronous behavior. *J. Anhui Norm. Univ.* 44, 233–237. doi: 10.14182/J.cnki.1001-2443.2012.03.005

Zeng, W., Shan, L., Su, B., and Du, S. (2023). Epileptic seizure detection with deep EEG features by convolutional neural network and shallow classifiers. *Front. Neurosci.* 17, 1145526. doi: 10.3389/fnins.2023.1145526

Zhang, F., Wu, W., and Wang, C. (2023). Pattern-based learning and control of nonlinear pure-feedback systems with prescribed performance. *Sci. China Inform. Sci.* 66, 1–22. doi: 10.1007/s11432-021-3434-9

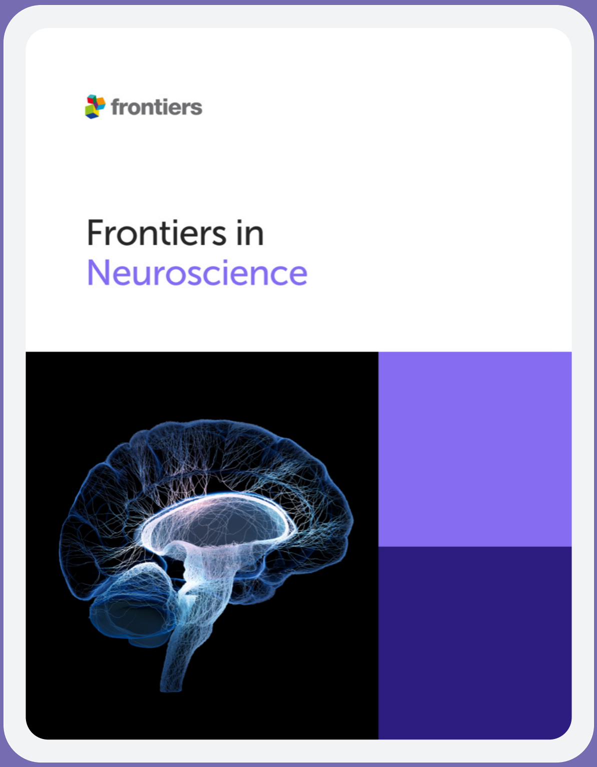
Frontiers in Neuroscience

Provides a holistic understanding of brain function from genes to behavior

Part of the most cited neuroscience journal series which explores the brain - from the new eras of causation and anatomical neurosciences to neuroeconomics and neuroenergetics.

Discover the latest Research Topics

See more →



frontiers

Frontiers in
Neuroscience

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact