Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Transformer-based hybrid systems to combat BCI illiteracy

Maximilian Achim Pfeffer [a], Johnny Kwok Wai Wong [b], Sai Ho Ling [a],*

[a] Faculty of Engineering and Information Technology, University of Technology Sydney, New South Wales, Australia
[b] Faculty of Design, Architecture and Building, University of Technology Sydney, New South Wales, Australia

ARTICLE INFO

Dataset link: https://physionet.org/content/eegmmidb/1.0.0/, https://www.bbci.de/competition/iv/

Keywords:
Transformers
Neural networks
Brain–computer-interface
Electroencephalography
Artificial intelligence
Convolutional neural networks
Hybrid-models
Biomedical engineering
BCI illiteracy
Signal processing

ABSTRACT

This study addresses the challenge of enhancing Brain–Computer Interfaces (BCIs), focusing on low Signal-to-Noise Ratios and "BCI illiteracy" often affecting up to 20% of users. Transformer-based models show promise but remain underexplored. Three experiments were conducted. Experiment A assessed the performance of architectures combining Convolutional and Transformer Blocks for binary Motor Imagery (MI) classification. Experiment B introduced a hybrid system, refining both block types and adding a Noise Focus Block to infuse Stochastic Noise, enhancing multi-class classification robustness. Experiment C evaluated the emerging architectures on 106 subjects, focusing on robustness across weak and strong learners. In Experiment A, the best networks achieved a validation accuracy of 0.914 and a loss of 0.146 (p=0.000967, F=12.675). In Experiment B, the proposed architecture improved multi-class MI classification to 84.5% on Dataset II, significantly improving performance for BCI-illiterate users. Experiment C showed a Kappa >83%, reduced standard deviation, and a highest validation accuracy of 88.69% across all individuals. The hybrid integration of Transformers, CNNs, and Noise-Resonance-based layers significantly enhances classification performance, particularly for weak BCI learners. Further research is recommended to optimize hybrid system architectures and hyperparameter settings to overcome current limitations in BCI performance.

List of abbreviations

- AI: Artificial Intelligence
- BCI: Brain-Computer Interface
- BN: Batch Normalization
- CEL: Cross-Entropy Loss
- CNN: Convolutional Neural Network
- EEG: Electroencephalography
- ELU: Exponential Linear Unit
- EMG: Electromyogram
- EOG: Electrooculogram
- ERD: Event-Related Desynchronization
- ERP: Event-Related Potentials
- ERS: Event-Related Synchronization
- FBCSP: filter-bank common spatial patterns
- FD: Frequency-Domain
- GFP: Global Field Power
- TC: Temporal Convolution
- HSD: Tukey's Honestly Significant Difference
- ICA: Independent Component Analysis
- MI: Motor Imagery
- NLP: Natural Language Processing
- OFAT: One-Factor-At-a-Time
- PCA: Principal Component Analysis
- SNR: Signal-to-Noise Ratio
- STF: Stochastic Transformer Focus
- t-SNE: t-distributed Stochastic Neighbor Embedding
- XAI: Explainable Artificial Intelligence

1. Introduction

Electroencephalography (EEG) captures brain activity by recording postsynaptic potentials generated by neurons in the cerebral cortex [1]. These electrical potentials, which occur perpendicular to the cortical surface, can be detected non-invasively through electrodes placed on the scalp [2]. The EEG signals represent the summation of all local field potentials, offering a cost-effective method to monitor and analyze brain activity with high temporal resolution. This capability has paved the way for the development, and deployment of Brain–Computer Interfaces (BCIs), which translate neural signals into commands that can control external devices or software applications. BCIs have become a critical area of research due to their potential to provide communication and control pathways for individuals with motor disabilities and to enhance human–computer interaction in general [3].

* Corresponding author.
E-mail addresses: maximilianAchim.pfeffer@student.uts.edu.au (M.A. Pfeffer), Johnny.Wong@uts.edu.au (J.K.W. Wong), steve.ling@uts.edu.au (S.H. Ling).

https://doi.org/10.1016/j.compbiomed.2025.111378
Received 9 January 2025; Received in revised form 27 October 2025; Accepted 5 December 2025
Available online 12 December 2025

However, despite the promising applications of BCIs, there remain significant challenges that hinder their widespread adoption and effectiveness [4]. One of the primary challenges is the low Signal-to-Noise Ratio (SNR) inherent in EEG data [5,6]. The scalp-recorded signals are often contaminated with noise from various sources, including muscle activity, eye movements, and external electrical interference [4,6]. This noise makes it difficult to isolate the brain's electrical activity related to specific tasks or commands, reducing the accuracy and reliability of BCIs. Researchers have employed various signal processing techniques to mitigate noise and improve the SNR, but achieving consistently high performance remains elusive [3–5,7,8].

Another critical challenge in BCI research is the phenomenon known as “BCI illiteracy” or “weak learners”, which refers to the inability of some individuals to use BCI systems effectively. Studies estimate that approximately 15%–20% of the population struggles to achieve proficiency with BCIs, regardless of the specific approach or technology used [9,10]. The underlying reasons for BCI illiteracy are not fully understood, but it is believed that individual variations in brain structure and function play a significant role [11–13]. Some users may not produce detectable patterns of brain activity necessary for the BCI to interpret their intentions accurately [14]. Additionally, other factors such as excessive muscle artifacts, misunderstanding of instructions, or environmental noise can contribute to poor BCI performance. While these latter issues are often surmountable, the individual variations in brain structure present a more intractable problem, which was observed in this study as well.

In response to these challenges, the exploration of novel BCI approaches has gained momentum. One such promising development is the application of Transformer-based models to BCI tasks. Transformers, originally developed for natural language processing (NLP), have demonstrated remarkable success in various tasks by capturing long-range dependencies in data through self-attention mechanisms, and are currently being investigated in many AI research fields such as image analysis, medical image segmentation, and time-series forecasting [15–19]. This capability is particularly relevant to EEG data, where temporal dependencies across multiple time points can provide crucial information for interpreting neural signals [20]. By applying Transformers to BCIs, researchers aim to enhance the robustness and accuracy of these systems, potentially overcoming some of the limitations associated with traditional methods. Several hybrid approaches have already explored this potential. EEG-TCNet integrates temporal convolutional modules with self-attention to capture sequential dependencies efficiently, while EEG-ITNet combines inception-style convolutional blocks with Transformer layers to strengthen multi-scale feature extraction [21,22]. Conformer-based variants have also been adapted for EEG decoding, blending convolutional front-ends with Transformer encoders to leverage both local spectral patterns and global temporal context [18,23]. Similarly, DRDA introduces dual residual attention modules to refine spatiotemporal representations [24], and time-series Transformer frameworks have been adapted to EEG for cross-subject generalization [17,25]. While these studies report improved classification accuracy and robustness compared to traditional CNN-only models, they generally focus on optimizing average-case learners and do not explicitly address variance reduction or weak-learner performance. This gap motivates our design of STFNet, which embeds stochastic resonance directly within a CNN-Transformer backbone to enhance both mean accuracy and stability across heterogeneous subject populations.

Recent studies have shown that Transformer-based models can achieve improved performance in BCI tasks, particularly in terms of classification accuracy and robustness to noise. These models have outperformed conventional approaches in several benchmark datasets, demonstrating their potential to enhance BCI performance for a broader range of users. However, despite these promising results, the application of Transformers in BCI research is still in its infancy, with only a handful of studies exploring this area [18,26]. This knowledge gap was recently highlighted as a call to action to further assess the capability of Transformers in improving multi-class classification performance and robustness in EEG-based BCI, particularly in overcoming the existing limitations related to signal-to-noise ratio (SNR) and overall accuracy [26]. The limited research thus far suggests that Transformers could be particularly beneficial for users who struggle with traditional BCI approaches, including those affected by BCI illiteracy.

To further investigate the potential of Transformers in BCIs, this study conducted two experiments focused on Motor Imagery (MI) classification, a common BCI task in which users imagine specific movements to control external devices. In the first experiment, various network architectures combining Convolutional and Transformer Blocks were assessed under different settings to determine their effectiveness in MI classification. The results indicate that networks incorporating both spatial convolution and transformer attention blocks achieved high validation accuracies, with the best-performing model reaching a validation accuracy of up to 91.4% with a loss of 0.146. The findings were statistically significant, highlighting the potential of combining these architectural elements to improve BCI performance.

In the second experiment, a novel combinatory approach is proposed, integrating Transformer Blocks, Self-Attention, Convolutional Blocks, and a Noise Focus Block designed to introduce stochastic noise within the network during both training and classification. This architecture aimed to enhance the model's robustness to noise, a critical factor in real-world BCI applications by leveraging properties of all aforementioned building blocks. The results from this experiment were particularly noteworthy, as the proposed model not only improved the overall accuracy of MI classification but also significantly enhanced performance for a subject previously deemed BCI illiterate. Lastly, a third experiment was conducted to investigate the performance of the developed hybrid models on Dataset I as well, confirming the superior robustness of the feature extraction and noise resistance of all model architectures put forward.

For multi-class classification, the proposed models achieved average accuracies of up to 83.3% and 90.6% on Dataset I and Dataset II, respectively. Hence, this set a new benchmark for multi-class MI tasks and demonstrates the superior robustness of the proposed approach across different subjects by utilizing transformer-based and noise-inducing layers in conjunction with traditional deep-learning methodologies.

These findings suggest that the integration of Transformers with Convolutional Neural Networks (CNNs) and noise-resonance mechanisms offers a promising pathway to address some of the most persistent challenges in BCI research. By leveraging the strengths of these different architectural components, it was hypothesized to enable the development of more universally effective BCIs that can accommodate a wider range of users, including those who have previously struggled with traditional systems. As the field continues to evolve, further exploration of Transformer-based models and their application to BCIs could lead to significant advancements in both the robustness and accessibility of these technologies.

In this study, the potential of combining Transformers, CNNs, and noise-resonance-based layers to improve BCI classification performance is investigated, particularly for weak BCI learners. This approach represents a significant step toward overcoming the current limitations of BCI technology and achieving more reliable and universally applicable systems. As research in this area progresses, the integration of these advanced computational techniques may ultimately lead to the development of BCIs that can truly work for all users, regardless of individual variability in brain function or external factors [26–28].

2. Background

2.1. Event-related potentials and event-related desynchronization in brain–computer interfaces

Event-Related Potentials (ERPs) and Event-Related Desynchronization (ERD) as exemplary displayed in Fig. 1 (subplots a, c) are foundational concepts in the realm of BCI applications [29]. ERPs are

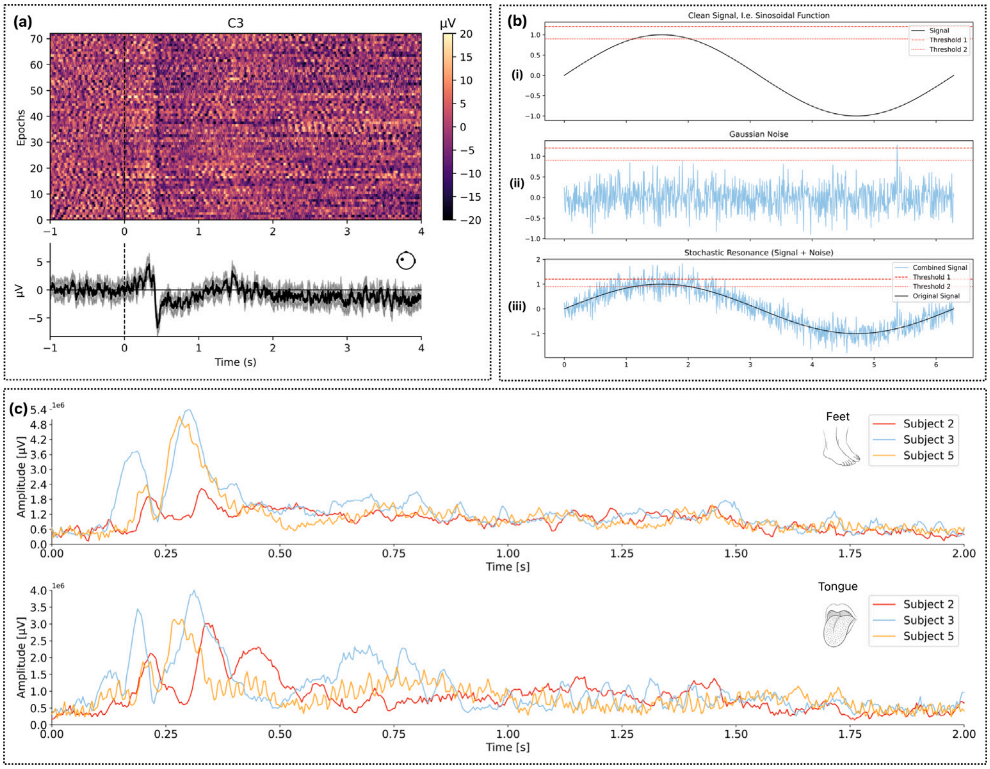

Fig. 1. Overview of event-related Desynchronization and stochastic resonance using sample data from Dataset II. (a) Epoched from MI-related channel (C3) of Subject 3, showing archetypal ERD around left-hand movement cue onset, averaged across all generated epochs. (b) Visualization of Stochastic Resonance. (i): A clean sinusoidal signal that does not independently surpass the preset detection thresholds (Threshold 1 and Threshold 2). (ii): Gaussian noise, which oscillates around zero and fails to cross the thresholds on its own. (iii): The combined signal (original signal + noise), demonstrating the phenomenon of stochastic resonance. The added noise allows the signal to exceed both thresholds intermittently. (c) GFP for all four tasks of Dataset II for the best-performing subject (Subject 3), the worst-performing subject (Subject 2), and the subject with the biggest improvement using STFNet (Subject 5).

time-locked EEG responses elicited by specific sensory, cognitive, or motor events, reflecting the brain's processing of these stimuli. Notably, the P300 component, a positive deflection occurring approximately 300 ms after stimulus presentation, is widely used in BCI applications for its robustness and reliability in signal detection [5,30,31]. ERDs, conversely, represent a decrease in power within specific frequency bands, typically the Mu (8–13 Hz) and Beta (13–30 Hz) bands, associated with motor imagery or execution tasks [31].

In BCI systems, ERPs are often utilized in paradigms like the P300 speller, where the user's focus on a target stimulus generates detectable ERP components [30,32]. ERD-based BCIs capitalize on the modulation of sensorimotor rhythms during imagined movements, enabling users to control external devices through motor imagery [30,31,33,34]. The differentiation between imagined movements such as left-hand, right-hand, or foot movement generates distinct ERD patterns that the BCI can classify [2,33,35].

However, the efficacy of ERD-BCIs relies heavily on the user's ability to produce consistent and distinguishable EEG patterns. Some users struggle to generate sufficient ERD signals, necessitating alternative approaches that encourage users to explore different mental strategies to enhance signal generation [36].

2.2. Limitations of EEG-based BCIs

Despite significant advancements, a considerable subset of individuals cannot achieve effective control over BCI systems, a challenge often referred to as "BCI illiteracy" [37–39]. This phenomenon, generally assumed to affect up to 20%–30% of potential users [9,17, 38,40] (and which was further substantiated by our investigation as delineated in Fig. 2, arises from a combination of neurophysiological differences [41,42], cognitive and attentional factors [43], and psychological influences such as motivation, fatigue, and stress [44]. These user-specific attributes shape the generation and stability of EEG features such as ERD/ERS patterns, leading to pronounced variability in BCI performance. Throughout this paper, we use the term "weak learners" to denote subjects who consistently achieve low motor imagery classification accuracy across runs and models, reflecting unstable or weak ERD/ERS patterns. Conversely, "strong learners" are subjects with consistently high classification accuracy and clear ERD/ERS expression. These terms are descriptive rather than formal categories, and are used to differentiate subject-level performance variability in line with prior reports of BCI illiteracy.

Fig. 2 presents a comparative analysis of motor imagery decoding performance across nine participants and illustrates the manifestation of BCI illiteracy within the analyzed datasets. In panel (a), two individuals (S2 and S5) fall below the predefined weak-learner threshold of 35% classification accuracy, thereby confirming their limited capacity for generating reliably classifiable neural representations. However, when examining the broader cohort, it becomes evident that reduced classification performance cannot be attributed to a single neurophysiological marker. Panels (b)–(d) reveal pronounced variability in the spectral characteristics of the mu (8–13 Hz), beta (13–30 Hz), and SMR (12–14 Hz) bands, indicating that the mechanisms underlying low BCI performance are not uniform across individuals. For instance, S2 exhibits moderate beta activity yet markedly lower mu and SMR power, while S5 demonstrates comparatively stronger mu modulation but remains within the weak-learner range. Conversely, several participants with weaker oscillatory amplitudes in specific bands still achieve substantially higher decoding accuracies. These observations emphasize that motor imagery classifiability arises from an interplay of multiple spectral and spatiotemporal factors rather than from any single dominant frequency component.

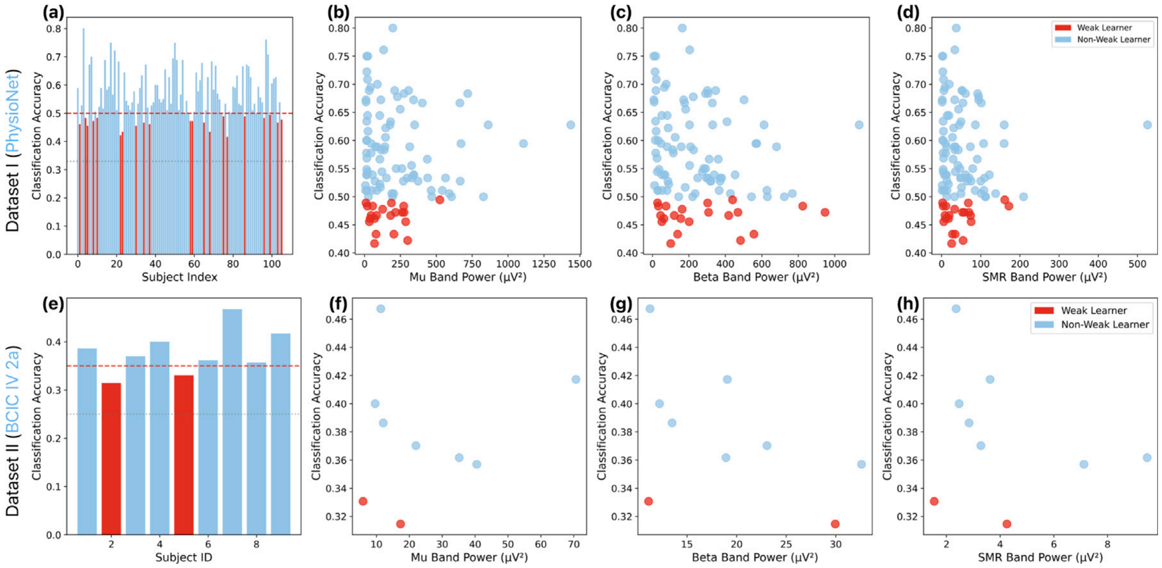

Fig. 2. Dataset characteristics using metrics designed to highlight weak BCI learners. Top row: (a) PhysioNet Dataset subjects for multi-class classification, with by-chance threshold (gray dotted line) and weak-learner threshold (red dotted line). (b, c, d) Mu, Beta, and SMR Band Power, respectively, highlighting weak learners based on the criteria as per subplot (a) within the bands, showing lower overall Band Power strengths for most weak learners. Bottom row: (e, f, g, d): Same as top row for Dataset II.

This variability supports the interpretation that “weak learner” should be understood as a general descriptor encompassing individuals whose EEG features deviate from population-level discriminability, rather than as a neurophysiologically distinct subgroup. The herein employed 35% threshold serves primarily as an operational reference point based on the per-chance threshold of 25%, to ultimately enable to exemplify below-average performers; it does not imply a discrete boundary or underlying homogeneity in neural mechanisms. Weak learners may therefore reflect diverse underlying causes, including diminished event-related desynchronization, inconsistent task engagement, atypical spatial patterns of cortical activation, or suboptimal signal-to-noise characteristics. The overlap observed between weak and non-weak learners across frequency bands reinforces the view that poor BCI performance cannot be reliably predicted by any single spectral measure. The variability is further exemplified in Fig. 1 subplot (c), which displays the global field power (GFP) across motor imagery tasks for three representative subjects [45]. GFP enables simultaneous visualization of all electrodes, offering a measure of overall neural activity. Subject 3 illustrates a strong learner with consistent ERD/ERS responses, Subject 2 a weak learner with variable and indistinct patterns, and Subject 5 an intermediate case. Such contrasts highlight how neurophysiological and cognitive differences directly contribute to BCI illiteracy and performance variability.

Beyond individual differences, EEG recordings are inherently low in signal-to-noise ratio (SNR) and sensitive to artifacts. Physiological contamination arises from EOG, EMG, and cardiac activity [46,47], while environmental interference such as electromagnetic sources further degrades data quality [48]. To mitigate these effects, numerous preprocessing approaches have been proposed, including band-pass and notch filtering, regression techniques, and Independent Component Analysis (ICA) and its variants [6,47,49]. However, while such methods can attenuate artifacts, they risk discarding subtle but task-relevant neural signals, and their performance is highly sensitive to electrode placement, scalp conductivity, and inter-individual brain anatomy [50,51]. Consequently, extensive calibration and individualized models remain necessary, and weak or subthreshold features often remain undetectable. The delicate balance between noise suppression and signal preservation continues to pose a central challenge in EEG signal processing.

Consequently, the weak-learner classification should be regarded as an integrative performance descriptor, summarizing individuals whose EEG-based representations yield below-average decoding performance despite adequate training and calibration conditions. This perspective underscores the necessity of adaptive, individualized modeling approaches that account for inter-subject variability and promote inclusive and reliable BCI operation across diverse user populations.

2.3. Stochastic resonance

Stochastic resonance, as shown in Fig. 1, is a counterintuitive phenomenon wherein the addition of a specific level of noise to a nonlinear system enhances the detection and transmission of weak signals [52]. Initially introduced to explain periodic climate changes, this concept has since found applications across various fields, including neuroscience, molecular systems, and mechanical oscillating systems. In molecular interactions, for instance, intrinsic noise can amplify subtle periodic signals, enhancing the precision of detection [53]. Similarly, in mechanical systems, controlled noise introduction can optimize the response to periodic forces, leading to improved accuracy in measurements [54]. These examples illustrate how noise, rather than being a detrimental factor, can be strategically leveraged to bring weak signals to the forefront.

In the realm of BCI development, stochastic resonance offers a novel strategy to address the perennial challenge of low SNR in EEG data. EEG signals, recorded from the scalp, are inherently weak and often buried in a sea of noise, making the extraction of meaningful neural patterns particularly challenging. Traditional noise reduction methods, such as filtering and ICA, focus on attenuating artifacts but may inadvertently eliminate subtle neural signals crucial for accurate interpretation. This delicate balance between noise suppression and signal preservation has limited the efficacy of BCI systems, especially in multiclass classification tasks where robustness and sensitivity to minute signal variations are critical.

Stochastic resonance presents a promising alternative by turning this challenge on its head: instead of attempting to eradicate noise, it introduces controlled noise to enhance the system’s sensitivity to subthreshold signals that would otherwise remain undetected [55,56]. This approach allows for a resonance effect, where weak EEG signals become more pronounced against the noisy background, improving their detectability. Unlike conventional methods that risk filtering out essential neural information, stochastic resonance modulates the noise level to achieve an optimal state, amplifying the weak neural signals of interest.

Studies have demonstrated the efficacy of this approach in enhancing neural signal detection. For example, [57] showed that the addition of noise could improve tactile sensation in humans, suggesting that stochastic resonance can be harnessed to augment sensory perception. In EEG analysis, McDonnell and Ward (2011) discussed how this phenomenon might be leveraged to improve neural signal processing [58]. By optimizing the noise level within the system, weak EEG signals that are typically obscured can emerge more distinctly, thereby enhancing the system's robustness and accuracy in translating neural activity into user intent.

In the context of BCI development, incorporating stochastic resonance into neural networks and deep learning models, such as CNNs and Transformers, holds the potential to circumvent existing limitations. By introducing noise not only to the input EEG data but also within the layers of these models, the phenomenon of stochastic resonance can be exploited to improve signal processing and classification performance. This paradigm shift offers a pathway to more resilient BCI systems, capable of detecting and interpreting weak neural signals with higher accuracy, even in the presence of significant noise. Therefore, applying stochastic resonance to BCI systems could mitigate the effects of BCI illiteracy by amplifying the neural signals of users who struggle to produce strong ERD patterns. This enhancement could lead to more reliable detection of user intent, thereby improving BCI performance and user experience.

Subplot (b) of Fig. 1 illustrates how noise can conceptually enhance the detectability of a weak signal, a principle that can be leveraged in neural signal processing, sensory systems, and other scientific fields where signal detection is challenged by low signal-to-noise ratios. This concept is particularly relevant in the context of this EEG signal-processing study: In panel A, a clean sinusoidal signal represents a weak neural component in EEG data. This signal, despite its importance, fails to independently surpass the preset detection thresholds (Threshold 1 and Threshold 2) due to its low amplitude. Such weak signals are often characteristic of subtle neural activity, like motor imagery or cognitive processing, which can be difficult to isolate in EEG recordings. Panel B introduces Gaussian noise, analogous to the inherent background noise present in EEG measurements. This noise fluctuates around zero and, on its own, does not cross the detection thresholds. In EEG recordings, this noise could stem from various sources, including muscle artifacts, electrical interference, or sensor noise. While this noise is typically considered detrimental, it can be harnessed constructively. Panel C shows the result of combining the original signal with the noise, demonstrating the phenomenon of stochastic resonance. With the added noise, the composite signal intermittently exceeds both thresholds. This enhanced detectability suggests that introducing controlled noise can amplify weak but meaningful EEG signals, making them more recognizable by neural network models. In the context of EEG processing with CNNs and transformers, this principle can be exploited by adding noise to both the input data and within neural network layers. This approach may improve the model's ability to detect and classify neural patterns, thereby enhancing the overall accuracy of EEG-based neural signal decoding. Note that EEG waveforms are not inherently sinusoidal; they are complex and vary with neural processes. The sinusoidal signal in Fig. 1 subplot (b) serves as a simplified model to illustrate how stochastic resonance can augment and hence, improve the detectability of weak neural signals.

2.4. Integrating transformers and self-attention mechanisms

Transformers and self-attention mechanisms have revolutionized deep learning, particularly in natural language processing, due to their ability to model long-range dependencies and focus on relevant input features [15]. In EEG signal processing, these architectures can capture complex temporal and spatial patterns, offering advantages over traditional convolutional and recurrent neural networks [26].

By integrating stochastic resonance with transformer-based models, it is possible to further enhance BCI performance. The self-attention mechanism allows the model to weigh the importance of different parts of the EEG signal, effectively filtering out irrelevant information while emphasizing critical neural patterns. When combined with stochastic resonance, which amplifies weak but relevant signals, the model becomes more adept at discerning user intent even in noisy conditions.

Recent studies have begun exploring transformer architectures in EEG-based BCIs. For instance, *Song et al. (2021)* proposed a transformer-based model for EEG classification, demonstrating improved performance over traditional methods [59]. However, as recently put forward in [26], transformer-based architectures for EEG-BCI signal processing in human–computer interactions have only begun to be explored, highlighting the need for further assessment and meticulous comparison of network architectures given the promising outcomes of the few existing studies.

3. Material and methods

3.1. Dataset

For the experiments, two datasets have been utilized: The first dataset utilized in this study was obtained from the Physionet Motor Imagery (MI) database [60], herein referred to as *Dataset I*. This dataset comprises over 1500 EEG recordings from 109 subjects, each recorded using a 64-channel EEG system with the BCI2000 platform. The experimental protocol included 14 runs for each subject, consisting of two baseline runs (one with eyes open and one with eyes closed) and three runs for each of the motor/imagery tasks. Each task run lasted two minutes, during which the subjects alternated between performing the designated motor or imagery task and relaxing. This comprehensive dataset offers a diverse range of motor and imagery tasks, providing a rich resource for the development and evaluation of EEG-based BCI systems.

Exclusion criteria were applied throughout all experiments and model training, resulting in a few subjects being excluded due to intrinsical errors or data structure abnormalities. In particular, subjects with IDs of 88, 92, and 100 were excluded for all runs due to formatting errors, resulting in a slightly reduced sample size of 106 subjects in total.

Additionally, four-class classification performance was assessed using Dataset II, which was taken from the Berlin Brain–Computer Interface Competition IV, Subset 2a [61]. Dataset II includes EEG recordings from 9 subjects engaged in a cue-based BCI experiment, featuring four different motor imagery tasks: imagining movements of the left hand (class 1), right hand (class 2), both feet (class 3), and tongue (class 4). Each participant underwent two recording sessions on separate days. Each session comprised 6 runs, with each run containing 48 trials, divided equally among the four tasks (12 trials per class). In total, each session provided 288 trials per subject. This structured protocol facilitated the assessment of multi-class motor imagery classification in a well-controlled experimental setting.

3.2. Experimental setup

Code was implemented using Python 3.11, PyTorch (Version 2.4.0), on MacOS Tahoe 26.0.1 using Apple's M1 Max Silicon Chip and Metal Performance Shaders for GPU training acceleration [62]. Training of the subject-dependent benchmarking of all models, UTS iHPC [63] (Saturn Shard) was used with Linux RedHatEnterprise (Ootpa 8.10) to enable 2 CUDA GPUs (Tesla V100-PCIE-32GB) and Python 3.11 inclusive PyTorch. Statistical analysis was implemented using both the statistical programming language R (Version 2023.06.2+561) and Python3.

For Experiment A, a total of 7 (seven) model architectures were implemented using Dataset I to assess cross-subject (or inter-subject) model performance and feature extraction capabilities of all denoted subjects, and a total of 756 (seven hundred and fifty six) models were developed using only individual subject's data only. Model architectures were built by combining different modifications and quantities of building blocks, mainly a) convolutional blocks and b) transformer blocks. In this paper, references to Model IDs such as C3 will denote three convolutional blocks, T2C2 denote a combination of two transformer blocks and two convolutional blocks, and so on.

For Experiment B, we repurposed the building blocks of Experiment A and added additional model complexity by introducing additional self-attention layers, noise layers, Frequency-Domain Augmentation, and more, to assess multi-class classification performance against the established architectures in Experiment A.

Lastly, in Experiment C, we utilize the proposed network architecture to test model performance on Dataset II, which is more extensively studied, hence providing a direct and evident comparison of model robustness, loss metrics and individual validation accuracy performance.

For all experiments and models, validation loss metrics were used as early stopping criteria as opposed to validation accuracy, to allow a more nuanced assessment of model convergence and prevent overfitting, particularly in cases where accuracy may not fully reflect subtle improvements in the model's learning process.

3.3. *Experimental workflow*

For reproducibility, we summarize the complete experimental pipeline applied across all datasets and experiments. Raw EEG data were first band-pass filtered between 0.5–40 Hz and resampled to 250 Hz. Trials were epoched relative to the cue onset (–0.5 s to 4.0 s) and baseline-corrected. Each epoch was standardized channel-wise, and only the 22 motor-related electrodes (10–20 montage) were retained for analysis.

The data splits were organized as follows. In Experiment A, within-subject training was performed using three runs for training and one run for testing, to evaluate the effect of transformer block depth. In Experiment B, subject-specific analyses emphasized weak learners, again using held-out runs for evaluation. In Experiment C, large-scale validation was performed on 106 subjects using a leave-one-subject-out strategy, where one subject served as the test set while all others were used for training. This design ensured that no data leakage occurred across train and test partitions.

Data augmentation was applied during training, including Gaussian noise infusion, temporal jittering, and frequency-domain dropout. The models were trained for 200 epochs using the Adam optimizer (learning rate 1×10^{-3}, weight decay 1×10^{-4}) with a batch size of 64. Evaluation metrics included mean accuracy and Cohen's Kappa, computed per subject and averaged across runs or folds as appropriate. Statistical significance of ablation results was assessed using repeated-measures ANOVA.

This workflow was applied consistently across all experimental conditions, with task-specific variations (e.g., number of transformer blocks, noise level Δ) described in the corresponding subsections of the Results.

3.4. *Data augmentation and class balancing*

During training we applied three complementary augmentations aimed at improving generalization under low SNR and class imbalance. (1) *Temporal jittering:* each trial was circularly shifted by a random offset drawn uniformly from ±50 ms (bounded so cues remained within the analysis window). (2) *Frequency-domain dropout:* after FFT, we zeroed randomly selected narrowbands (1–3 bins per trial; non-adjacent) before inverse FFT, simulating nonstationary spectral perturbations without destroying phase structure. (3) *Mixup:* we linearly combined pairs of trials and their one-hot labels with $\lambda \sim \text{Beta}(\alpha, \alpha)$, $\alpha = 0.2$ [64], which encourages smoother decision boundaries. In addition, we injected Gaussian noise ($\Delta \in \{0.1, 0.2\}$ as reported in Tables 5 and 7) to exploit stochastic resonance while regularizing the encoder.

To address imbalance in the four-class setting (Dataset II), we used per-class oversampling within each mini-batch to equalize class priors seen by the optimizer, with stratified train/validation splits to prevent leakage. All summary metrics (Accuracy, Kappa) are reported alongside macro-averaged F1 and macro-Recall to reflect minority-class behavior independent of class frequency.

3.5. *Classification of illiteracy*

Classification performance and statistical significance were assessed using a Random Forest classifier with 50 estimators and maximum depth of 10, coupled with standard scaling preprocessing in a scikit-learn pipeline. Each subject's motor imagery data was evaluated through 5-fold stratified cross-validation to ensure robust performance estimates while maintaining class balance across folds. Statistical significance of classification performance was determined through permutation testing with 100 label randomizations, where null distributions were generated by shuffling class labels and re-computing cross-validation accuracy to establish chance-level baselines. Subjects were classified as weak learners if their accuracy fell below 35% (10% above the 25% chance level for 4-class motor imagery) or below the 50% mark for 3-class motor imagery (where chance level would be around 33.33%), with additional significance testing at $\alpha = 0.05$ to control for multiple comparisons. Frequency band analysis focused on motor-relevant oscillations extracted from sensorimotor channels (C3, Cz, C4) using 4th-order Butterworth bandpass filters for mu (8–13 Hz), beta (13–30 Hz), and sensorimotor rhythm (12–14 Hz) bands, with spectral power quantified as the temporal variance of filtered signals averaged across trials and channels. Cohen's kappa was computed alongside accuracy to account for class imbalance, providing a chance-corrected measure of classification performance that ranges from 0 (chance level) to 1 (perfect agreement).

3.6. *Analysis*

For all models, cross-entropy loss was implemented using PyTorch using Theorem 1 for all binary classification models:

$$\mathcal{L}_{\text{binary}}(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right] \tag{1}$$

- $\mathcal{L}_{\text{binary}}$: Binary cross-entropy loss.
- N: Number of samples.
- y_i: True label for the ith sample (0 or 1).
- $\hat{y}_i$: Predicted probability for the ith sample.

Likewise, all multi-class model's loss was implemented as per Theorem 2:

$$\mathcal{L}_{\text{multi-class}}(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(\hat{y}_{i,c}) \tag{2}$$

- $\mathcal{L}_{\text{multi-class}}$: Multi-class cross-entropy loss.
- N: Number of samples.
- C: Number of classes.
- $y_{i,c}$: True label for the ith sample and cth class (one-hot encoded).
- $\hat{y}_{i,c}$: Predicted probability for the ith sample and cth class.

For all models, ANOVA was conducted to compare the validation accuracies across different numbers of transformer blocks using the following theorem:

$$F = \frac{\mathrm{MS}_{\mathrm{between}}}{\mathrm{MS}_{\mathrm{within}}} \tag{3}$$

- F: F-statistic for the ANOVA test.
- $\mathrm{MS}_{\mathrm{between}}$: Mean square between the groups.
- $\mathrm{MS}_{\mathrm{within}}$: Mean square within the groups.

Following the ANOVA, Tukey HSD post-hoc test was performed to identify which specific groups differ:

$$\mathrm{HSD} = q_\alpha \sqrt{\frac{\mathrm{MS}_{\mathrm{within}}}{n}} \tag{4}$$

- HSD: Honestly Significant Difference.
- q_α: Critical value from the Studentized range distribution.
- $\mathrm{MS}_{\mathrm{within}}$: Mean square within the groups.
- n: Number of observations per group.

3.7. Model architecture and building blocks

For this study's development of network architectures for EEG signal processing, a series of models were designed, varying in the number and arrangement of convolutional and transformer blocks. These networks, referred to as C2, C3, and C4, incorporate either two, three, or four convolutional layers, respectively. Additionally, networks labeled T1C2, T2C2, T3C2, and T4C2 combine up to four additional transformer blocks with convolutional layers. For simplicity, these architectures will be referred to as 'n' transformer blocks and 'm' convolutional modules, e.g., T1C2 as one transformer block with two convolutional modules.

The convolutional layers apply one-dimensional convolutions across the EEG channels to capture local spatial dependencies within the data. Each convolutional block consists of a convolution operation, followed by batch normalization, a rectified linear unit (ReLU) activation function, and dropout for regularization. This structure allows the network to learn hierarchical feature representations of the EEG data.

Transformer blocks, which are incorporated into the T1C2, T2C2, and T3C2 models, leverage multi-head self-attention mechanisms to capture long-range dependencies in the EEG signals. These blocks are particularly effective in modeling the temporal dynamics of EEG data, which is crucial for the accurate classification of neural signals. Each transformer block includes a multi-head attention layer, followed by a multi-layer perceptron (MLP), with layer normalization and dropout applied for regularization as shown in Fig. 3.

Gaussian noise layers are introduced during the training phase to simulate the effect of stochastic resonance and enhance the model's robustness. Noise is added to the input data and within intermediary layers, aiming to improve the model's ability to detect weak signals embedded in noise, a common characteristic of EEG data.

For classification, the output from the convolutional and transformer layers is flattened and passed through a fully connected MLP layer. This final layer produces predictions for three-class classification tasks, enabling the assessment of motor imagery performance.

The training process employs early stopping, guided by validation loss metrics, to prevent overfitting. The models are trained on augmented and balanced EEG data, utilizing techniques such as jittering, frequency-domain augmentation, and mixup. This ensures that the networks are exposed to a diverse set of training examples, promoting robustness and generalizability. By systematically varying the number and combination of convolutional and transformer blocks, this study aims to evaluate the impact of these architectural components on the classification of motor imagery tasks in EEG data.

The EEG data were preprocessed and structured as multi-dimensional arrays to be fed into the neural network models. The data preparation involved several key steps:

1. **Data Loading and Structuring:** EEG data for each subject were loaded using the MNE library, resulting in arrays with the shape $(\text{num\_trials}, \text{channels}, \text{timepoints})$. Each trial represents a segment of EEG recording, where *channels* refers to the number of EEG electrodes (22 in this case), and *timepoints* refers to the temporal resolution of the signal. This structure maintains the spatial and temporal characteristics of the EEG signal.
2. **Normalization and Conversion:** The raw EEG data were converted to microvolts to standardize the values across different subjects and sessions. The data were then transformed into NumPy arrays to facilitate efficient numerical operations and further processing.
3. **Data Augmentation:** To enhance the model's ability to generalize, various data augmentation techniques were applied, including:

    - *Time-Domain Jittering:* Random shifts were applied to the EEG data along the temporal axis to simulate variability in signal timing.
    - *Frequency-Domain Augmentation:* Noise was introduced in the frequency domain by adding random noise to the FFT-transformed signals and subsequently applying the inverse FFT to obtain time-domain signals.
    - *Mixup:* This technique involved linearly combining pairs of EEG trials and their corresponding labels, creating synthetic examples to augment the training data.

    This multi-modal augmentation strategy is designed to enhance performance by increasing training data diversity rather than addressing inherent class imbalances, as motor imagery datasets typically contain equal trial distributions across classes by design. Time-domain jittering specifically targets the temporal variability that naturally occurs in self-paced motor imagery tasks, where subjects exhibit trial-to-trial variations in imagery timing and duration. Frequency-domain augmentation enhances model robustness against the spectral noise commonly present in EEG recordings, particularly important for preserving the mu and beta rhythm modulations that are fundamental to motor imagery classification. The mixup technique's linear interpolation between trials creates decision boundaries that are more robust to the subtle inter-class differences often observed in motor imagery patterns. The subsequent class balancing step ensures that the augmentation process does not inadvertently create class distribution skews, maintaining equal representation across all motor imagery conditions and preventing classifier bias that could artificially inflate performance metrics. This systematic approach to data integrity ensures that performance improvements stem from enhanced model robustness rather than dataset artifacts.
4. **Balancing the Dataset:** Class balancing is achieved by identifying the maximum class count across all motor imagery categories and iteratively augmenting underrepresented classes through concatenated application of jittering, frequency-domain noise injection, and mixup operations until reaching the target sample size. For each class requiring augmentation, the three augmentation techniques are applied simultaneously to the existing class samples, quadrupling the available data per iteration (original + jittered + frequency-augmented + mixup), with this process repeated until the target count is reached and subsequently truncated to the exact maximum class size. This approach ensures deterministic class balance while maximizing data diversity through combined augmentation modalities applied in parallel rather than sequentially.
5. **Splitting Data into Training and Testing Sets:** The preprocessed data were split into training and testing sets, ensuring that the test set remained unseen during the model training process.

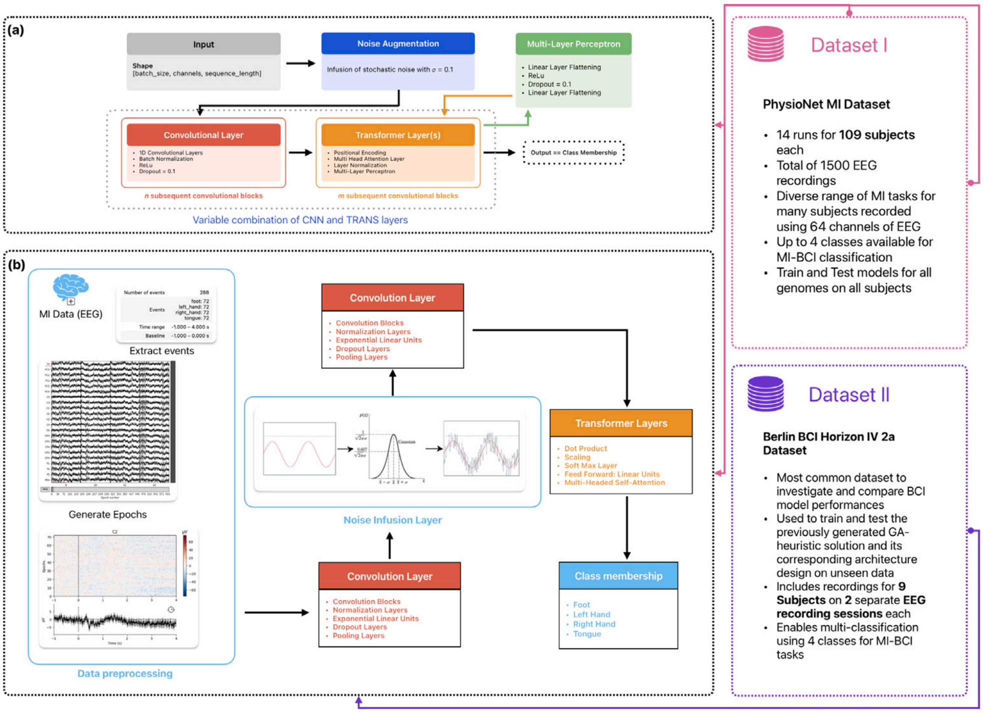
(a)
Input
Shape
[batch_size, channels, sequence_length]
Noise Augmentation
Infusion of stochastic noise with σ = 0.1
Multi-Layer Perceptron
• Linear Layer Flattening
• ReLu
• Dropout = 0.1
• Linear Layer Flattening
Convolutional Layer
• 1D Convolutional Layers
• Batch Normalization
• ReLu
• Dropout = 0.1
Transformer Layer(s)
• Positional Encoding
• Multi Head Attention Layer
• Layer Normalization
• Multi-Layer Perceptron
Output == Class Membership
n subsequent convolutional blocks
m subsequent convolutional blocks
Variable combination of CNN and TRANS layers
(b)
MI Data (EEG)
Number of events 288
Events foot: 72 left_hand: 72 right_hand: 72 tongue: 72
Time range -1.000 – 4.000 s
Baseline -1.000 – 0.000 s
Extract events
Generate Epochs
Data preprocessing
Convolution Layer
• Convolution Blocks
• Normalization Layers
• Exponential Linear Units
• Dropout Layers
• Pooling Layers
Gaussian
Noise Infusion Layer
Transformer Layers
• Dot Product
• Scaling
• Soft Max Layer
• Feed Forward: Linear Units
• Multi-Headed Self-Attention
Convolution Layer
• Convolution Blocks
• Normalization Layers
• Exponential Linear Units
• Dropout Layers
• Pooling Layers
Class membership
• Foot
• Left Hand
• Right Hand
• Tongue
Dataset I
PhysioNet MI Dataset
• 14 runs for 109 subjects each
• Total of 1500 EEG recordings
• Diverse range of MI tasks for many subjects recorded using 64 channels of EEG
• Up to 4 classes available for MI-BCI classification
• Train and Test models for all genomes on all subjects
Dataset II
Berlin BCI Horizon IV 2a Dataset
• Most common dataset to investigate and compare BCI model performances
• Used to train and test the previously generated GA-heuristic solution and its corresponding architecture design on unseen data
• Includes recordings for 9 Subjects on 2 separate EEG recording sessions each
• Enables multi-classification using 4 classes for MI-BCI tasks

Fig. 3. High-level overview of experimental setup and technical methodology. **(a) High-Level Visualization of Architectural Combination of Network Building Blocks. I.e., for Model T1C2 n=2 and m=1. (b)** High-level architecture of the proposed Stochastic Transformer Focus Network with Transformer-Based Self Attention Modules and Noise Resonance Imputation Layer. Right: Dataset use and validation logic.

Stratified splitting was used to maintain the class distribution in Dataset I. For Dataset II, the competition provides a pre-defined test split dataset, which was adopted accordingly.

6. **Data Format for Model Input:** The EEG data were then converted into PyTorch tensors and loaded into the model in the form of 3D arrays with the shape (batch_size, channels, sequence_length). Here, *batch_size* refers to the number of samples in each training iteration, *channels* corresponds to the EEG electrodes, and *sequence_length* denotes the temporal dimension. The data were passed through convolutional layers and transformer blocks to extract both spatial and temporal features of the EEG signal.
7. **Gaussian Noise Injection:** During the training phase, Gaussian noise was added to the input data as a form of regularization, aiming to improve the model's robustness to variations in EEG signals.

Noise augmentation layer

The NoiseLayer is a critical component designed to increase the robustness of the model by introducing Gaussian noise during training. This is mathematically represented as:

$$X' = X + \mathcal{N}(0, \sigma^2) \tag{5}$$

where X is the input data, $\mathcal{N}(0, \sigma^2)$ denotes the Gaussian noise with zero mean and variance σ^2. Here, σ is a hyperparameter representing the standard deviation of the noise, which can be adjusted based on validation performance. This technique simulates potential real-world environmental noise, thus preparing the model to handle unseen noisy data effectively.

Convolutional layers

The convolutional layers are designed to extract spatial and temporal features from the EEG signals. The architecture employs a sequence of convolutional operations, each followed by batch normalization and activation functions:

$$Z_1 = \mathrm{ELU}(\mathrm{BN}(W_1 * X' + b_1)) \tag{6}$$

where W_1 and b_1 are the weights and biases of the convolutional layer, respectively, and BN represents batch normalization. The choice of ELU (Exponential Linear Unit) as an activation function helps in capturing nonlinearities and maintaining robustness against vanishing gradients.

Integration of multi-head self-attention

The standard multi-head self-attention (MHA) as described by Vaswani et al.[15] is used to process temporally-structured convolutional feature embeddings $\mathbf{X} \in \mathbb{R}^{N \times d_{\text{model}}}$ through parallel attention heads. For each head $i \in \{1, \dots, h\}$, linear projections generate queries, keys, and values: $\mathbf{Q}_i = \mathbf{X}\mathbf{W}_i^Q$, $\mathbf{K}_i = \mathbf{X}\mathbf{W}_i^K$, $\mathbf{V}_i = \mathbf{X}\mathbf{W}_i^V$, where $\mathbf{W}_i^Q, \mathbf{W}_i^K, \mathbf{W}_i^V \in \mathbb{R}^{d_{\text{model}} \times d_k}$ with $d_k = d_{\text{model}}/h$. Scaled dot-product attention computes:

$$\mathrm{Attention}_i(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i) = \mathrm{softmax}\left(\frac{\mathbf{Q}_i\mathbf{K}_i^T}{\sqrt{d_k}}\right)\mathbf{V}_i \tag{7}$$

Head outputs are concatenated and projected: $\mathrm{MultiHead}(\mathbf{X}) = \mathrm{Concat}(\mathrm{head}_1, \dots, \mathrm{head}_h)\mathbf{W}^O$. This mechanism establishes global temporal dependencies complementing the CNN's local receptive fields, where convolutional layers extract hierarchical spatial–temporal features through localized kernels while self-attention captures long-range sequence relationships. Integration occurs via residual connections: $\mathbf{Y} = \mathrm{LayerNorm}(\mathbf{X} + \mathrm{MultiHead}(\mathbf{X}))$, enabling selective fusion of local convolutional representations with global attention-weighted features for enhanced motor imagery pattern discrimination across temporal sequences.

In contrast, configuration studies (non-STFNet models such as T1C2, T3C2, etc. as per Table 1 employ a fixed 4-head attention mechanism rather than the variable multi-head configurations used in the main STFNet ablation study, providing a controlled baseline for evaluating the impact of convolutional versus transformer layer depth while maintaining consistent attention head count.

3.8. Stochastic transformer focus network

During this study, the Stochastic Transformer Focus Network (STFNet) eventually emerged by combining various layer architectures to enhance the robustness and adaptability of EEG signal processing models for BCIs as shown in Fig. 3 by incorporating all features of the previous building blocks and concepts, such as:

- Stochastic Noise Layer
- Transformer Blocks
- Convolutional Blocks (spatial resolution)
- Convolutional Blocks (temporal resolution)

Following the initial convolutional layers, STFNet utilizes depthwise separable convolutions, which are efficient both computationally and in terms of model parameters. These layers perform a depthwise spatial convolution followed by a pointwise convolution:

$$Z_2 = \mathrm{ELU}(\mathrm{BN}(W_2 * (\mathrm{DWConv}(Z_1)))) \tag{8}$$

This configuration allows the model to learn spatial hierarchies more effectively and is particularly suited for handling high-dimensional EEG data.

To focus on the most relevant features across time series sequences, STFNet incorporates an attention mechanism:

$$A = \mathrm{softmax}(W_a Z_2) \tag{9}$$

where W_a are the trainable parameters of the attention module. This mechanism allows the network to dynamically weigh the importance of different features at each time step, which is crucial for tasks such as event-related potential detection.

The final layer of the STFNet is a fully connected layer followed by a softmax activation function, mapping the extracted features to the output classes:

$$P = \mathrm{softmax}(W_f A) \tag{10}$$

where W_f are the weights of the final dense layer. This setup ensures that the network outputs probabilities for each class, making it suitable for multi-class classification tasks.

Compared to state-of-the-art methodologies like the EEG Conformer as put forward by textitSong et al. [18], STFNet has an analytical advantage by incorporating stochastic resonance mechanisms and enhanced temporal processing to improve multi-class classification in EEG signals. While the EEG Conformer effectively merges convolutional and transformer layers to capture spatial and temporal features, STFNet introduces controlled noise to facilitate stochastic resonance. This added noise amplifies weak EEG signals, enhancing their detectability and interoperability. In parallel, STFNet incorporates an advanced temporal processing module, which further refines the network's ability to capture complex time-domain patterns, crucial for differentiating subtle neural activity.

To reiterate, the key features of the emerging model are:

- Controlled Gaussian noise injection during training to enhance weak EEG signals through stochastic resonance.
- Multi-stage feature extraction with convolutional layers and transformer encoders to capture spatial and temporal EEG features.
- Transformer encoders leverage multi-head self-attention to model long-range dependencies in EEG data.
- Time-domain jittering and mixup augmentation improve generalization and robustness by exposing the model to varied signal patterns.
- Dropout and Adam optimizer with weight decay prevent overfitting and ensure convergence.
- Optimization for Apple Silicon MPS hardware acceleration, enabling efficient training on large EEG datasets.

3.8.1. Ablation study and hyperparameter tuning

To investigate the implicit effects of hyperparameters and noise infusion modalities, an ablation study employed a controlled one-factor-at-a-time (OFAT) framework centered on a fixed baseline configuration (noise = 0.1, TCN = 1, embedding = 70, transformer depth = 2, heads = 10, forward expansion = 4), the results of which are presented in Table 6. Each architectural or stochastic variable was independently perturbed across multiple levels while all others remained constant, and key configurations were repeated to capture intra-condition variance. This 30-run design balances interpretability and reproducibility, allowing clear attribution of observed performance shifts to individual factors. Unlike brute-force factorial search, which scales exponentially and demands thousands of evaluations, the OFAT approach yields statistically analyzable results with a fraction of the computational load. The inclusion of repeated runs per factor further supports variance estimation and non-parametric testing, enabling reliable identification of influential parameters without prohibitive resource expenditure.

By employing a structured one-factor-at-a-time and sensitivity-oriented design rather than a full combinatorial sweep, the computational complexity was reduced by several orders of magnitude. A complete factorial exploration of six parameters would require over ten thousand distinct model trainings, equating to multiple years of computation given the 0.5–2-hour training time per subject across nine participants. In contrast, the selected 30 unique configurations strategically probe both individual parameter effects and key multi-factor sensitivities, reducing total runtime to approximately 15–60 GPU-hours while preserving interpretability, parameter coverage, and statistical tractability. To further deepen the analysis, we added manual adjustments strategically as to enable a more complex assessment of all hyperparameters, resulting in a total of 50 unique OFAT runs for the herein presented STFNet-centric ablation study.

4. Results

As shown in Table 1, the final validation accuracies for CNNs in Experiment A reached up to 77.0%±6.3%. In contrast, transformer-infused models achieved a higher maximum validation accuracy of 91.4%±2.5% and exhibited lower standard deviation across all validation metrics.

An initial one-way ANOVA was conducted to compare the mean validation accuracies across different groups defined by the number of T-blocks. The results indicated a significant effect of T-block number on validation accuracy, with an F-statistic of 12.675 and a p-value of 0.000967, suggesting that at least one group's mean validation accuracy is significantly different from the others.

The one-way ANOVA test was conducted to compare the effect of the number of transformer blocks on validation accuracy. The analysis revealed a statistically significant difference between the groups, $F(4, 9) = 12.675$, $p = 0.000967$.

To further identify which specific groups differed, a post-hoc analysis using Tukey's Honest Significant Difference (HSD) test was performed. Results are summarized Table 2.

The analysis reveals significant differences between models with zero transformer blocks and those with one or more transformer blocks. Specifically, the group comparisons indicate that models without transformer blocks (group1=0) differ significantly from models with transformer blocks (group2=1, 2, 3, 4). This finding confirms that incorporating transformer blocks into the network architecture leads to statistically significant improvements in validation accuracy. The positive mean differences (*meandiff*) across these comparisons further suggest that models utilizing transformer blocks consistently outperform their CNN-only counterparts.

In contrast, comparisons among models that already include transformer blocks (e.g., group1=1 vs. group2=2) do not show statistically significant differences. This suggests that while the introduction of transformer blocks enhances performance over CNN-only models, increasing the number of transformer blocks beyond one or two does

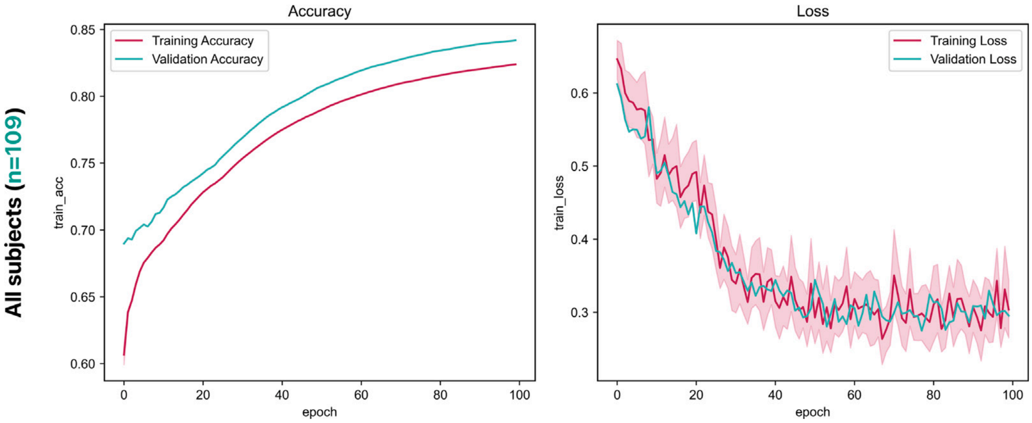

Fig. 4. Training and validation performance for cross-subject training of the T3C2-architecture using Dataset I.

Table 1
Training and validation performances of Experiment A for various combinations and epoch settings with bold highlights of the two best-performing validation loss and accuracies.

max_epochs	T-blocks	CNN-blocks	Parameters	Training		Validation	
				Accuracy	Loss	Accuracy	Loss
100	0	2	139,000	0.722	0.71	0.714	0.54
100	0	3	504,000	0.752	0.438	0.735	0.5
100	0	4	2,000,000	0.753	0.344	0.743	0.483
200	0	2	139,000	0.741	0.449	0.738	0.505
200	0	3	504,000	0.78	0.269	0.757	0.452
200	0	4	2,000,000	0.783	0.0752	0.77	0.422
100	1	2	208,000	0.827	0.279	0.84	0.306
200	1	2	208,000	0.875	0.103	**0.914**	0.146
100	2	2	277,000	0.823	0.0715	0.828	0.309
200	2	2	277,000	0.874	0.0136	0.896	0.144
100	3	2	345,000	0.841	0.638	0.857	0.285
200	3	2	345,000	0.888	0.296	**0.912**	**0.09**
100	4	2	414,000	0.824	0.206	0.834	0.279
200	4	2	414,000	0.878	0.0267	0.911	**0.103**

Table 2
Tukey HSD post-hoc test results for pairwise comparisons of validation accuracy across different numbers of transformer blocks in the cross-subject classification performance of Experiment A.

group1	group2	meandiff	p-adj	lower	upper	reject
0	1	0.1252	0.0092	0.0327	0.2176	True
0	2	0.1192	0.0124	0.0267	0.2116	True
0	3	0.1417	0.0041	0.0492	0.2341	True
0	4	0.1297	0.0073	0.0372	0.2221	True
1	2	−0.006	0.9997	−0.1193	0.1073	False
1	3	0.0165	0.9864	−0.0968	0.1298	False
1	4	0.0045	0.9999	−0.1088	0.1178	False
2	3	0.0225	0.9586	−0.0908	0.1358	False
2	4	0.0105	0.9975	−0.1028	0.1238	False
3	4	−0.012	0.9959	−0.1253	0.1013	False

Table 3
Individual performance of STFNet in Experiment B using Dataset II.

Subject ID	Test Accuracy	F1 Score	Recall	Kappa κ
A01	0.8648	0.8652	0.8644	0.8197
A02	0.6749	0.6705	0.6773	0.5671
A03	0.9391	0.9344	0.9339	0.9121
A04	0.8465	0.8461	0.8461	0.7952
A05	0.8261	0.8261	0.8262	0.7680
A06	0.6837	0.6819	0.6831	0.5782
A07	0.9314	0.9313	0.9329	0.9086
A08	0.8598	0.8608	0.8601	0.8130
A09	0.8712	0.8710	0.8706	0.8281
Average	0.8330 ± 0.095	0.8319 ± 0.095	0.8327 ± 0.094	0.7767 ± 0.125

not lead to further significant improvements. The lack of statistically significant differences between models with one or more transformer blocks implies the presence of a potential threshold effect. This observation is also reflected in the validation accuracies presented in Table 1 where models with two and three transformer blocks exhibit comparable performance levels.

Performance metrics during training of the network model T3C2 of Table 1 are displayed in Fig. 9. The findings indicate that not only does the low loss metric indicate good generalization of the model, but rather unexpectedly, validation performance is better than training performance (see Fig. 4).

In Experiment B, the 4-class classification performance of STFNet notably surpasses previously reported models on Dataset II. As shown in Table 3, validation accuracies exceed 83.3% for 7 out of 9 subjects, resulting in an average accuracy of 84.5% ± 7.3% across the entire dataset. Furthermore, Cohen's Kappa statistic, with an average of 0.75 ± 0.13, indicates a substantial level of agreement between the model's predictions and the actual class labels. The elevated kappa value underscores the robustness of STFNet in distinguishing between the four motor imagery tasks, reflecting not only accuracy but also the model's consistency in performance across different subjects, including weak learners.

As per Table 4, the performance of STFNet on Dataset II demonstrates a significant improvement over previous models in 4-class classification tasks. STFNet achieves an average accuracy of 84.5% with a standard deviation of 0.075 and a Cohen's Kappa of 0.75. Notably, this performance is superior to other models like FBCSP, ConvNet, and EEGNet, especially in the context of subjects traditionally considered as

Table 4

4-class performances across historical and cutting-edge models, sorted alphabetically. The best values for each column are highlighted in bold, STFNet stands out with high confidence Kappa scores and improved standard deviation, indicating the robustness of the model across weaker learners, whose validation accuracy is strongly increased using STFNet.

Model	A01	A02	A03	A04	A05	A06	A07	A08	A09	Average	SD σ	Kappa κ
C2CM [65]	0.875	0.652	0.903	0.667	0.625	0.455	0.896	0.833	0.795	0.744	0.153	0.6595
Conformer [18]	0.881	0.614	0.934	0.781	0.520	0.652	0.923	**0.881**	0.888	0.786	0.153	0.7155
ConvNet [66]	0.763	0.552	0.892	0.747	0.569	0.542	0.927	0.771	0.764	0.725	0.142	0.6337
DRDA [24]	0.838	0.551	0.874	0.753	0.623	0.572	0.862	0.753	0.820	0.747	0.126	0.6632
EEGNet [67]	0.857	0.615	0.885	0.670	0.559	0.521	0.896	0.833	0.868	0.745	0.152	0.6600
EEG-TCNet [68]	0.857	0.65	**0.945**	0.649	0.754	0.614	0.873	0.837	0.78	0.774	0.116	0.7
FBCSP [69]	0.760	0.565	0.813	0.610	0.550	0.453	0.828	0.813	0.708	0.678	0.137	0.570
FBCNet [70]	0.854	0.604	0.906	0.764	0.743	0.538	0.844	0.795	0.809	0.762	0.12	0.6827
Proposed [Table 6]	**0.883**	**0.724**	0.963	**0.847**	**0.837**	0.74	0.856	**0.882**	0.875	**0.845**	**0.073**	**0.75**

weak learners or BCI illiterate (e.g., A02 and A05). For instance, subject A02, which historically presented poor performance with models like ConvNet and FBCSP (accuracies of 55.2% and 56.5% respectively, and with low MI-related band-power strengths as per Fig. 2), achieved a significantly higher accuracy of 72.4% with STFNet. Similarly, subject A05, previously considered a weak learner with accuracies around 55%–62% in earlier models, reached an accuracy of 83.7% using the proposed model. This remarkable improvement in accuracy, alongside a higher Kappa score, indicates not only better classification performance but also suggests increased consistency and robustness across subjects. Moreover, the lower standard deviation observed with STFNet implies a reduction in performance variability, indicating that the model's architecture is more stable across diverse EEG signal patterns and varying subject-specific challenges.

The confusion matrices and task-specific metrics for false positives and false negatives in Fig. 5 provide insight into the model's performance and robustness across different subjects and tasks. For Subject 3, who is considered a high-performing learner, the confusion matrix shows high true positive rates across all tasks (left hand, right hand, feet, and tongue), with only minimal misclassifications. This subject serves as a gold standard for model performance, indicating the model's capacity to classify motor imagery tasks with high accuracy when the EEG signals are clear and distinct.

In contrast, Subjects 2 and 5, previously categorized as weak learners or BCI illiterate, exhibit more substantial improvements in classification performance using STFNet. For Subject 5, the confusion matrix shows significant improvements, especially in tasks like '`right hand`' and '`feet`', where previous models struggled. While there are still noticeable false negatives in the '`feet`' task, the overall distribution of true positives indicates that STFNet has successfully captured relevant features in these weaker learners' EEG signals. Subject 2's confusion matrix shows similar improvements, with fewer false positives and false negatives across tasks compared to earlier results. Notably, the false positive/false negative graph indicates a general reduction in false positives and negatives for both Subjects 2 and 5, suggesting that STFNet has increased the robustness of classification in weaker learners.

The decline in false positives and false negatives, particularly in challenging tasks like '`feet`' and '`tongue`' for Subjects 2 and 5, signifies a more balanced model performance. This reduction implies that STFNet can generalize better across subjects with varying levels of BCI proficiency. Additionally, the comparison with Subject 3 confirms that while weaker learners do not reach the same level of accuracy, the performance gap has been significantly reduced, indicating enhanced robustness and reliability in multi-class motor imagery EEG classification.

As for Experiment C, the results presented in Table 5 provide a comprehensive comparison of the performance metrics for different model architectures across 106 subjects. These models include various combinations of convolutional and transformer blocks (C3, C4, T1C2, T2C2, T3C2) as well as the more complex STFNet. The metrics presented — Accuracy, Loss, F1 Score, Recall, and Kappa — offer a detailed insight into each model's capability to handle EEG signal classification tasks under similar hyperparameter settings.

STFNet shows a marked improvement across nearly all performance metrics compared to the other models. With an accuracy of 0.8869 ± 0.0725, STFNet outperforms the simpler convolutional and transformer combinations such as C3 (0.6849 ± 0.1474) and C4 (0.5523 ± 0.1621). Even when compared to more sophisticated transformer-integrated architectures like T1C2 and T3C2, STFNet still maintains a higher accuracy, showcasing its enhanced capability in managing the complexities of EEG data. The F1 Score and Recall metrics further corroborate this, with STFNet achieving 0.8856 ± 0.0741 and 0.8869 ± 0.0725 respectively, indicating not only the model's precision in classification but also its effectiveness in correctly identifying relevant EEG signals across the dataset.

Another critical observation is the model's Kappa statistic, which measures the agreement between predicted and actual labels. STFNet achieves a Kappa value of 0.8304±0.1088, significantly higher than the values recorded for models like C3 (0.5274 ± 0.2211) and C4 (0.328 ± 0.2431). This elevated Kappa score suggests that STFNet's predictions align more consistently with the true classifications, implying better generalization across different subjects. Moreover, the standard deviation in STFNet's performance metrics is generally lower compared to other models, reflecting greater stability and robustness in its learning process across a diverse cohort of 106 subjects.

The hyperparameter settings, particularly the consistent application of noise ($\Delta = 0.1$), jitter augmentation (`J-Aug`), frequency-domain augmentation (`FD-Aug`), and mixup augmentation (`MU-Aug`), were uniform across all models, providing a controlled environment to assess the models' inherent capabilities. Despite the added complexity and parameter count in STFNet, it demonstrates superior performance with relatively balanced computational trade-offs, as evidenced by its maintained low loss value (0.0124 ± 0.009) similar to that of `T1C2` and `T3C2` models. Future implementation may explore NP-optimized implementation of such hyperparameter optimization as proposed by our foregoing work [26,71,72].

Table 7 presents the validation metrics for the same models under modified hyperparameter settings, specifically with increased noise infusion ($\Delta = 0.2$) and reduced data augmentations (J-Aug, `FD-Aug` set to 0.05, and `MU-Aug` to 0.1). STFNet continues to outperform the other models, achieving an accuracy of 0.906 ± 0.0596, which is an improvement from its performance in Table 5 (0.8869 ± 0.0725). This increase in accuracy, alongside an elevated F1 Score (0.9045 ± 0.0614) and Kappa (0.859 ± 0.0895), indicates that STFNet benefits from the adjusted noise and augmentation settings. The performance stability across subjects is also reflected in the reduced standard deviation, suggesting that STFNet remains robust even under different noise levels and augmentation strategies.

Comparing these results with those in Table 5, it is evident that the increased noise infusion and reduced data augmentation have a generally positive effect on STFNet, as well as on models like T1C2 and T3C2, which also show slight improvements in accuracy and F1 Score. For instance, T1C2's accuracy increased from 0.8443±0.1482 in Table 5

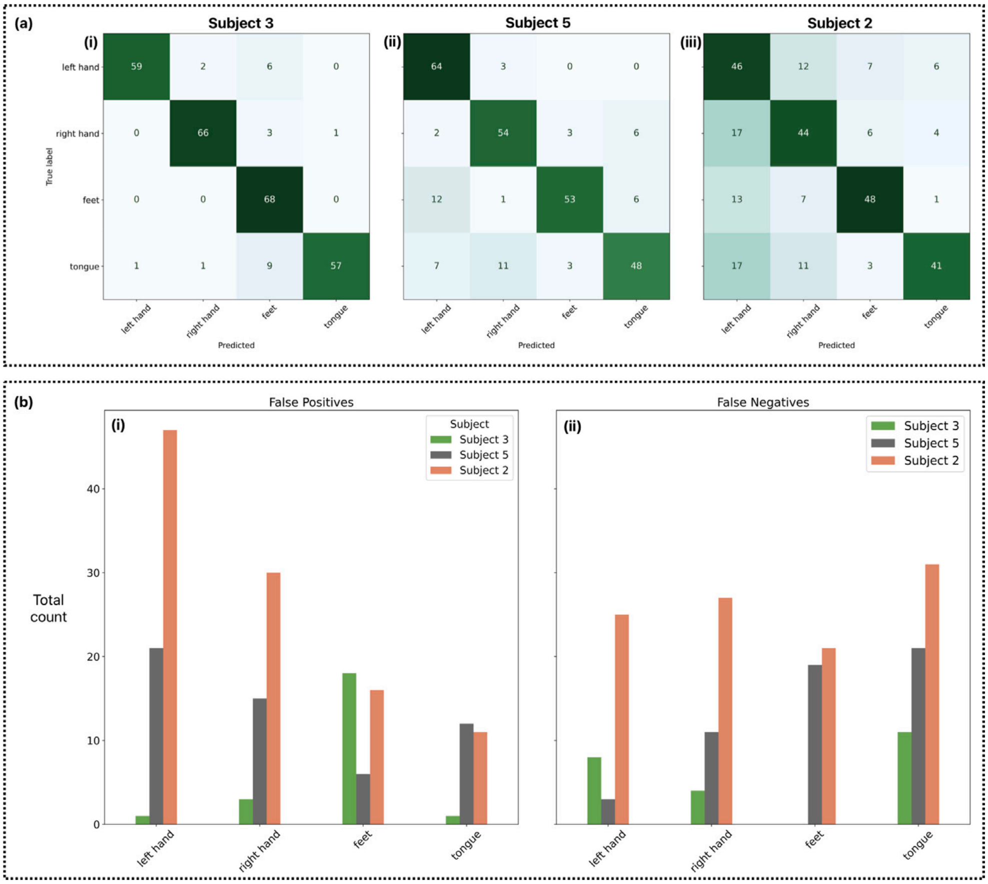

Fig. 5. Final model performance analysis for selected subjects 3, 5, and 2. **(a)** Confusion matrices for each subject's model performance. **(b)** Task-specific False Positive and False Negative Rates, respectively.

Table 5
Validation metrics for selected models of Experiment C (Configuration α).

Parameters	C3	C4	T1C2	T2C2	T3C2	STFNet
Noise [Δ]	0.1	0.1	0.1	0.1	0.1	0.1
J-Aug [Δ]	0.1	0.1	0.1	0.1	0.1	0.1
FD-Aug [Δ]	0.1	0.1	0.1	0.1	0.1	0
MU-Aug [α]	0.2	0.2	0.2	0.2	0.2	0.2
Max Epochs	200	200	200	200	200	200
Accuracy	0.6849 ± 0.1474	0.5523 ± 0.1621	0.8443 ± 0.1482	0.8071 ± 0.1270	0.8489 ± 0.1414	0.8869 ± 0.0725
Loss	0.0151 ± 0.0044	0.0175 ± 0.0035	0.0122 ± 0.0081	0.0147 ± 0.0081	0.0124 ± 0.0084	0.0124 ± 0.009
F1 Score	0.6717 ± 0.1627	0.4922 ± 0.2119	0.8356 ± 0.1738	0.8039 ± 0.1307	0.8408 ± 0.1618	0.8856 ± 0.0741
Recall	0.685 ± 0.1474	0.5523 ± 0.1620	0.8443 ± 0.1482	0.8071 ± 0.1270	0.8489 ± 0.1412	0.8869 ± 0.0725
Kappa	0.5274 ± 0.2211	0.328 ± 0.2431	0.7665 ± 0.2223	0.7107 ± 0.1904	0.7733 ± 0.2118	0.8304 ± 0.1088

to 0.8763±0.1219 in Table 7. This suggests that a higher noise level can aid in enhancing the generalizability of these models, likely due to the stochastic resonance effect. However, simpler models like C3 and C4 exhibit less improvement, indicating that while these modifications can benefit complex models, they may not suffice to significantly enhance the performance of less sophisticated architectures. Overall, the results imply that STFNet, with its inherent complexity and noise-handling capabilities, is more adept at leveraging increased noise levels and reduced augmentations to further improve classification performance.

As per Fig. 6 the inclusion of stochastic noise infusion ($\delta = 0.3$) overall yielded consistent gains in classification stability and performance across the evaluated motor-imagery classes. As illustrated in subplots (a) and (b), the mean class accuracies during validation testing were notably higher when the noise infusion layer was active compared to the $\delta = 0.0$ baseline, with the largest improvement observed for the left-hand class. This enhancement was accompanied by a narrower accuracy distribution for the left hand, right hand, and feet classes, indicating a reduction in variance and more stable class-specific predictions. Importantly, subplot (d) shows that for Subject 2 (the weakest learner) the validation loss remained comparable to that of the non-noise condition shown in (c), while validation accuracy improved, reflecting better generalization under noisy training. The higher training loss observed in the $\delta = 0.0$ configuration compared to the noise-infused case further supports this interpretation: the addition of stochastic perturbation during training likely mitigates overfitting by regularizing internal representations. Consequently, the model trained with $\delta = 0.3$ demonstrates improved robustness, more balanced class performance, and enhanced learning dynamics, particularly benefiting subjects and classes with weaker baseline performance. The corresponding training curves and confusion matrices are presented in Fig. 7.

Enhancing the previous analysis, the subplots (a) and (b) of Fig. 9 offer a more nuanced perspective on STFNet's robustness and generalizability through the distribution of Cohen's Kappa scores across 106

Table 6
Configuration-wise test scores for multi-classification performances during ablation study on STFNet hyperparameters: δ=noise infusion layer setting, nC=number of convolutional layers, emb=embedding size, nT=number of transformer layers, H=number of self-attention heads within the T layers, FW=forward expansion dimension for final embeddings. The best (ID=av) and worst (ID=b) performing architectures are highlighted in bold.

ID	δ	nC	emb	nT	H	FW	A01	A02	A03	A04	A05	A06	A07	A08	A09	Avg	SD
a	0.0	1	32	1	6	2	88.25	72.08	95.97	82.89	78.62	72.55	90.61	85.61	85.23	83.53	7.99
b	**0.0**	**1**	**96**	**2**	**6**	**4**	**87.18**	**61.83**	**92.67**	**86.84**	**80.79**	**69.30**	**92.77**	**85.97**	**85.98**	**82.59**	**10.48**
c	0.0	1	72	2	8	4	87.19	66.43	91.94	83.77	82.25	71.16	91.69	88.92	87.12	83.39	8.94
d	0.0	1	70	2	10	2	87.54	65.72	94.13	83.77	80.07	71.16	90.97	88.19	86.74	83.14	9.33
e	0.0	1	70	2	10	4	87.90	68.55	93.41	82.46	80.80	73.95	90.61	89.30	87.50	83.83	8.21
f	0.0	1	72	2	12	4	89.32	70.31	91.57	82.89	79.71	69.30	91.33	87.72	87.87	83.34	8.57
g	0.0	1	70	3	10	4	87.62	71.02	93.13	85.96	82.22	69.30	90.07	90.41	85.60	83.93	8.43
h	0.0	1	70	4	10	4	88.25	62.54	92.67	85.08	80.79	68.83	90.97	87.82	84.84	82.42	10.23
i	0.0	1	70	4	10	6	88.95	63.95	93.04	86.40	78.98	69.76	91.69	88.19	87.12	83.12	10.13
j	0.0	2	70	2	10	4	85.76	65.72	92.30	85.52	79.71	68.83	90.61	87.82	85.22	82.39	9.31
k	0.0	2	70	2	10	4	85.05	62.54	92.31	85.52	80.43	68.37	90.02	88.19	85.22	81.96	10.05
l	0.0	2	72	3	8	6	85.76	70.31	93.40	85.08	83.69	67.44	91.69	88.19	84.46	83.34	8.86
m	0.0	3	128	2	8	6	83.62	62.54	92.67	84.21	78.98	66.97	89.89	85.97	85.22	81.12	10.10
n	0.0	3	70	2	10	4	87.18	64.31	93.04	82.01	80.07	65.58	90.09	85.23	87.12	81.63	10.22
o	0.1	1	70	1	10	4	84.69	68.10	92.30	83.77	81.52	68.37	90.97	87.45	84.46	82.40	8.74
p	0.1	1	72	2	6	2	83.98	66.07	91.57	86.84	79.34	70.23	89.89	87.45	86.74	82.46	8.88
q	0.1	1	96	2	6	4	86.12	65.37	91.57	81.57	79.34	66.05	90.25	89.67	85.98	81.77	9.93
r	0.1	1	48	2	8	4	88.61	69.61	94.13	85.52	81.52	70.23	90.97	89.29	88.26	84.24	8.83
s	0.1	1	72	2	8	4	86.47	70.67	93.77	86.40	81.52	71.16	91.33	88.56	86.74	84.07	8.20
t	0.1	1	96	2	8	4	86.83	64.66	93.04	85.08	78.62	68.37	90.61	87.82	86.36	82.38	9.86
u	0.1	1	128	2	8	4	84.34	66.43	91.21	84.21	97.71	69.30	90.97	87.45	86.36	84.22	10.19
v	0.1	1	70	2	10	4	87.54	68.90	93.77	84.65	79.38	75.81	89.89	88.93	87.87	84.08	7.90
w	0.1	1	70	2	10	6	88.61	69.61	93.40	82.41	79.71	68.83	92.41	87.45	86.74	83.24	9.03
x	0.1	1	72	2	12	4	88.26	69.25	95.24	84.21	80.43	67.90	88.08	87.45	85.22	82.89	9.04
y	0.1	1	48	2	12	6	88.61	69.25	94.87	83.33	82.60	69.76	90.25	88.19	86.36	83.69	8.83
z	0.1	1	70	3	10	4	87.18	70.67	93.04	87.72	82.25	68.83	90.61	89.29	85.98	83.95	8.60
aa	0.1	1	70	4	10	4	87.54	63.95	93.41	86.40	79.34	71.16	91.69	87.45	85.23	82.91	9.73
ab	0.1	1	96	4	12	6	88.26	64.66	93.04	81.14	82.97	70.23	90.97	88.93	87.12	83.04	9.66
ac	0.1	2	96	2	8	4	84.69	63.60	93.77	80.70	80.79	67.90	90.97	86.34	84.84	81.51	9.95
ad	0.1	2	70	2	10	4	86.47	65.37	92.31	85.53	80.79	66.98	90.61	88.19	85.22	82.39	9.77
ae	0.1	3	128	2	8	6	83.27	65.72	92.67	83.33	78.98	68.37	88.44	85.97	84.46	81.25	8.91
af	0.1	3	70	2	10	4	88.61	64.66	93.41	85.09	81.51	66.04	92.06	84.50	86.36	82.47	10.39
ag	0.2	1	96	2	6	4	86.12	70.31	91.94	82.01	80.07	67.91	89.53	88.92	85.98	82.53	8.46
ah	0.2	1	72	2	8	4	87.54	67.49	93.04	82.46	81.15	72.09	90.61	88.56	87.50	83.38	8.61
ai	0.2	1	70	2	10	4	88.26	66.08	90.84	83.77	80.07	67.91	90.97	87.45	86.36	82.41	9.38
aj	0.2	1	72	2	12	4	88.26	67.49	93.77	84.21	81.88	71.63	92.06	89.67	88.64	84.18	9.10
ak	0.2	1	70	3	10	4	87.90	68.90	93.40	87.71	83.33	70.23	90.97	89.29	85.22	84.11	8.76
al	0.2	1	100	3	10	4	87.90	66.78	93.04	85.96	78.98	68.83	90.61	89.66	87.12	83.21	9.57
am	0.2	1	70	4	10	4	87.19	66.78	93.04	85.53	80.07	68.84	90.97	86.34	84.09	82.54	9.16
an	0.2	1	96	4	12	6	87.54	66.78	94.87	81.57	81.52	67.90	91.33	88.19	87.87	83.06	9.85
ao	0.2	1	96	4	12	6	87.90	66.79	94.14	81.14	81.15	68.38	92.42	87.83	87.88	83.07	9.79
ap	0.2	2	70	2	10	4	87.18	67.13	93.77	85.96	79.71	68.37	91.69	87.45	84.47	82.86	9.46
aq	0.2	2	70	2	10	4	85.76	69.25	95.23	85.96	79.34	68.83	88.08	86.34	84.46	82.58	8.71
ar	0.2	3	128	2	8	6	83.27	64.31	92.67	84.21	80.79	64.18	87.72	84.87	84.84	80.76	9.92
as	0.2	3	70	2	10	4	87.90	67.13	93.04	85.08	81.88	68.83	92.41	85.97	84.46	82.97	9.24
at	0.2	3	120	4	12	6	85.40	62.19	90.84	85.08	80.07	66.97	88.08	87.45	85.22	81.26	9.96
au	0.3	1	30	1	6	2	85.77	66.43	95.24	87.28	82.97	75.81	90.25	88.57	86.74	84.34	8.54
av	**0.3**	**1**	**30**	**1**	**10**	**10**	**88.26**	**72.44**	**96.33**	**84.64**	**83.69**	**73.95**	**85.56**	**88.19**	**87.50**	**84.51**	**7.38**
aw	0.3	1	70	2	10	4	87.90	68.19	94.87	83.77	80.79	73.49	87.36	88.56	88.63	83.73	8.34
ax	0.3	1	72	3	12	6	89.32	70.31	95.23	83.77	79.34	67.44	92.06	85.97	86.36	83.31	9.41

subjects. Fig. 9 subplot (a), STFNet demonstrates a pronounced peak towards the upper end of the Kappa scale, indicating a consistently high agreement between predicted and true labels across the subject pool. Notably, this distribution is not only centered at higher Kappa values but is also narrower compared to other models like C2, C3, and T2C2. The narrowness of this distribution suggests that STFNet achieves consistent performance across subjects, highlighting its resilience to inter-subject variability. In contrast, models such as C2 and C3 exhibit a broader spread with lower peaks, indicating greater variability in performance and reduced reliability in classification across different subjects.

Subplot (c) and (d) further substantiate these findings under modified hyperparameter conditions, specifically increased noise infusion and reduced data augmentation. STFNet continues to display a concentrated distribution near the higher Kappa values, even more prominently than in Fig. 9 subplot (b). This enhanced concentration underlines the model's robustness and stability when subjected to varying experimental conditions. Although models like T1C2 and T3C2 show some degree of improvement, their Kappa distributions remain wider and skewed toward lower scores, underscoring STFNet's superior capacity to generalize across diverse conditions. These results collectively reinforce the notion that STFNet not only excels in classification performance but also exhibits enhanced robustness and stability in its predictive capabilities, particularly when dealing with the complexities and noise inherent in EEG data.

The two noise settings evaluated (Δ=0.1 vs. Δ = 0.2; Table 5 and Table 7) allow us to quantify the impact of the noise layer. Increasing Δ from 0.1 to 0.2 improved average accuracy (0.8869 → 0.9060) and Cohen's Kappa (0.8304 → 0.8590). Importantly, the inter-subject variability was reduced (accuracy SD 0.0725 → 0.0596, Kappa SD 0.1088 → 0.0895), suggesting that higher levels of noise infusion not only increase mean performance but also stabilize performance across a heterogeneous subject pool. This supports the interpretation that the noise layer contributes to robustness by amplifying weak neural signals

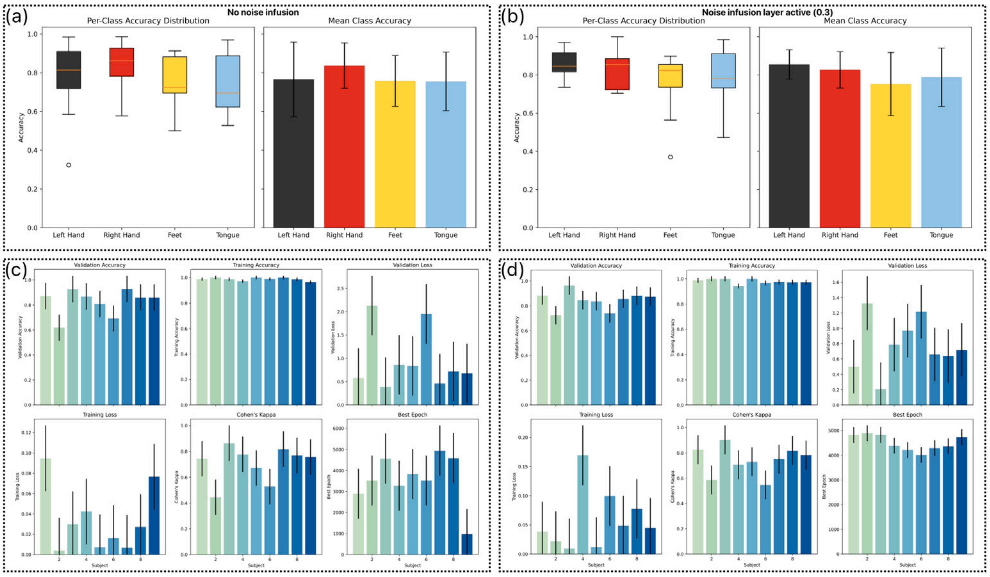

Fig. 6. Performance metrics for the ablation study using Dataset II. (a) and (b) show the distribution of the class-specific accuracy performances across the dataset for both cases without noise infusion and those with noise infusion, respectively. (c) and (d) show per-subject performance metrics for no-noise and noise-infused architectures during the ablation study, respectively.

Table 7
Validation metrics for selected models of experiment C with increased noise infusion and reduced data augmentation (Configuration β).

Parameters	C3	C4	T1C2	T2C2	T3C2	STFNet
Noise [Δ]	0.2	0.2	0.2	0.2	0.2	0.2
J-Aug [Δ]	0.05	0.05	0.05	0.05	0.05	0.05
FD-Aug [Δ]	0.05	0.05	0.05	0.05	0.05	0
MU-Aug [α]	0.1	0.1	0.1	0.1	0.1	0.1
Epochs	200	200	200	200	200	200
Accuracy	0.7231 ± 0.1491	0.5585 ± 0.1646	0.8763 ± 0.1219	0.8059 ± 0.1249	0.8773 ± 0.1281	0.906 ± 0.0596
Loss	0.0137 ± 0.0047	0.0174 ± 0.0036	0.0105 ± 0.0069	0.0144 ± 0.0081	0.0111 ± 0.0084	0.0112 ± 0.008
F1 Score	0.7082 ± 0.17	0.5004 ± 0.2125	0.8717 ± 0.1372	0.8034 ± 0.1269	0.8724 ± 0.1456	0.9045 ± 0.0614
Recall	0.7231 ± 0.1491	0.5585 ± 0.1646	0.8763 ± 0.1219	0.8059 ± 0.1249	0.8773 ± 0.1281	0.906 ± 0.0596
Kappa	0.5846 ± 0.2237	0.3377 ± 0.2468	0.8145 ± 0.1829	0.7088 ± 0.1877	0.816 ± 0.1921	0.859 ± 0.0895

while regularizing against subject-specific variability. These results are further underpinned by the findings of the STFNet hyperparameter and layer ablation study as presented in Table 6.

Overall, the results of Experiment C in conjunction with the results of Experiment B highlight the superiority of transformer-based models, in particular STFNet, which is able to effectively capture nuanced patterns within EEG signals. Its ability to deliver consistently high accuracy, precision, and stability across a large subject pool and different datasets demonstrates its potential as a more robust model for EEG classification tasks, particularly in scenarios involving complex and diverse datasets.

The boxplots in Fig. 8 illustrate the spatial distribution of normalized channel importance across EEG electrodes for the two ablation configurations: (a) without noise infusion $\delta = 0$ and (b) with moderate noise infusion ($\delta = 0.3$), which outlines how the overall topographic distribution of relevant channels remains consistent across both settings, indicating that noise infusion does not substantially alter the model's spatial feature weighting pattern. However, in the $\delta = 0$ condition, stronger emphasis is observed over the central motor regions, particularly C3, Cz, and CP2, reflecting the network's reliance on canonical motor-imagery–related areas [73–75] when no stochastic perturbation is introduced.

When moderate noise $\delta = 0.3$ is infused, the distribution becomes slightly more uniform, with a marginally increased importance of parietal and centro-parietal electrodes such as CP1, CP2, and POz. This suggests that controlled noise promotes more distributed feature utilization and enhances generalization by reducing overreliance on a small subset of dominant motor channels [76,77]. Despite these shifts, no statistically significant differences were observed in the overall channel importance distributions between the two conditions, confirming that the network's learned spatial focus remains physiologically stable while benefiting from improved robustness and feature integration under noise-infused training.

5. Discussion

The results of this study offer key insights into the effectiveness of STFNet in improving EEG-based BCI classification performance, particularly among subjects previously categorized as BCI illiterate.

In Experiment A, the analysis through ANOVA and the subsequent Tukey HSD tests demonstrates that incorporating transformer blocks into the model architecture significantly enhances validation accuracy. However, the data also indicate that merely increasing the number of transformer blocks does not necessarily translate to additional performance improvements. This finding suggests a limit to the benefits of simply adding architectural depth, indicating that the optimal number of transformer blocks for this dataset may be as few as one. Both Table 1 and Table 2 emphasize that while transformer blocks do contribute to enhanced performance, there is a threshold beyond which additional complexity yields diminishing returns.

These results highlight the need for more research into the fine-tuning and optimization of such hybrid systems. Rather than focusing solely on increasing the depth of the architecture, future investigations should explore how to better configure and optimize the interaction

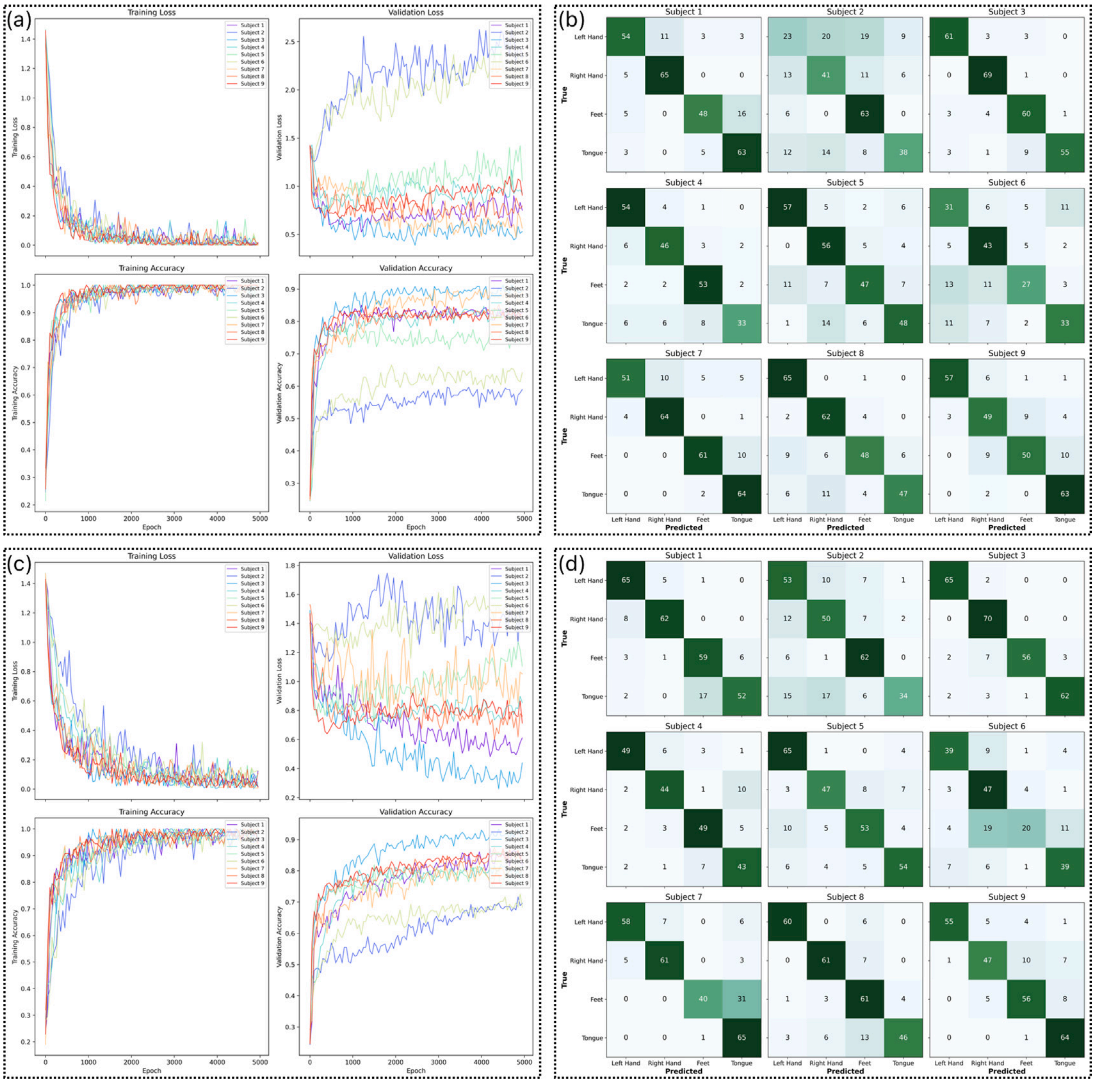

Fig. 7. Training and validation performance metrics for each subject of Dataset II during the STFNet hyperparameter ablation study. (a) and (c) show subject-specific training and validation accuracies and loss curves for the proposed architecture with noise infusion layers $\delta = 0$ and $\delta = 0.3$, respectively. (b) and (d) depict the final (best) validation metrics as confusion matrices for the resulting classifiers with the proposed architecture with noise infusion layers $\delta = 0$ and $\delta = 0.3$, respectively. (c) shows improved continuous learning rate and less stagnation across individuals when compared to (a), with clearer confusion matrices in (d) demonstrating superior classification performance not only for weak learners but across subjects (average) when compared to (b).

between convolutional layers, transformer blocks, and noise-handling strategies. This exploration could involve experimenting with different hyperparameters, layer types, or even self-organizing methods like evolutionary algorithms. A more nuanced understanding of how to optimize these hybrid systems will be crucial in pushing the boundaries of EEG-based BCI performance, particularly in applications where computational efficiency and adaptability are paramount.

As for Experiment B and C, STFNet not only enhanced the classification accuracy of weaker learners, such as Subjects A02 and A05, but also showed substantial improvements across the entire cohort. Fig. 9 (subplot c and d) display t-SNE plots illustrating the clustering of MI tasks for different models. STFNet achieves more distinct clusters for left hand, right hand, and feet tasks, indicating superior feature separation compared to earlier models like C2 and T2C2. This improved task differentiation suggests that STFNet's feature extraction capabilities, coupled with the benefits of stochastic resonance, enhance the model's ability to handle EEG data, particularly for challenging subjects.

For Subject 3 (the weakest learner of Dataset I as per Table 8), STFNet exhibits well-separated and compact clusters for each MI task as shown in Fig. 9, which indicates that STFNet's feature extraction process yields high-dimensional representations that are more discernible when reduced to a two-dimensional space using t-SNE, capturing complex patterns in the EEG data more effectively. In contrast, the t-SNE plots for models like C2 and C3 show significant overlap between task clusters, especially for left-hand and feet-movement tasks. This overlap suggests that these models struggle to extract distinct features, leading to less reliable classifications. Even for weak learners, such as Subject 2, STFNet maintains relatively distinct clusters, albeit with more overlap than seen in Subject 3. This further emphasizes STFNet's robustness and its ability to enhance the performance of subjects previously deemed BCI illiterate.

STFNet's architecture plays a critical role in amplifying subtle neural signals through noise infusion and enhanced feature extraction. Unlike other models that show scattered and intermingled clusters, STFNet's t-SNE plots indicate a higher degree of organization and separation between classes, underscoring the model's capability to manage inter-subject variability. This is particularly important for motor imagery tasks, where neural patterns vary widely among subjects. STFNet's ability to generate well-separated clusters across tasks suggests that it not only captures the relevant features more effectively but also leverages stochastic resonance to enhance the discriminability of these

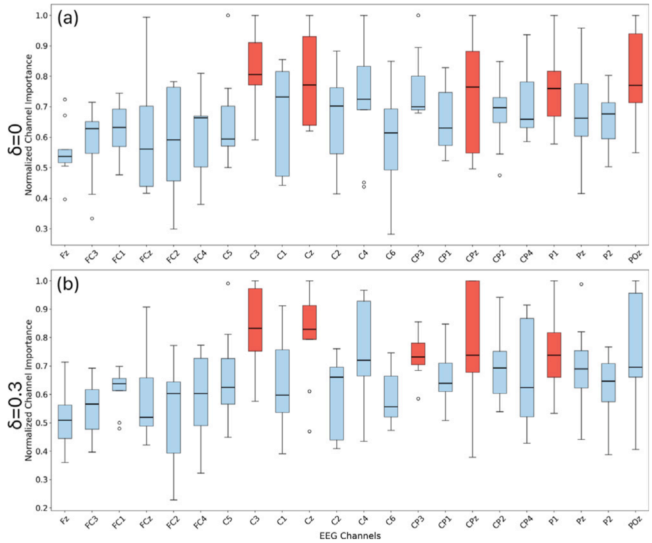

Fig. 8. Normalized channel importances across all subjects for the proposed architecture with noise infusion layers $\delta = 0$ (a) and $\delta = 0.3$ (b). The top 5 most important channels are highlighted in red.

features, leading to improved classification performance. Thus, the t-SNE analysis in Fig. 9 subplots (a,b) provide compelling visual confirmation of STFNet's superior feature extraction capabilities and its potential to improve BCI usability across diverse subject groups.

As shown in Figs. 10 and 11., the representational dynamics of the weak learner (Subject 2) and nono-weak learner (Subject 3)reveal how stochastic noise infusion interacts with the convolutional and transformer stages to restructure the learned feature space and enhance separability between motor imagery classes. Notably, under the $\delta = 0$ condition for the weak learner, activations across layers remain poorly differentiated, with feature trajectories displaying diffuse and overlapping distributions. The temporal convolution layer, though effective in extracting localized temporal features, fails to stabilize the representations sufficiently for downstream separation. Even after the transformer layer, which in principle integrates contextual and cross-channel dependencies, the features remain entangled, indicating that the model has not converged toward distinct task-related manifolds. This behavior is characteristic of weak learners whose EEG signals exhibit low discriminability and unstable spatial–temporal synchronization, leading the network to overfit transient signal noise rather than capturing consistent neural dynamics.

When noise infusion is introduced ($\delta = 0.3$), the network's representational structure evolves in a markedly different manner. The injected stochastic perturbation acts as an implicit regularizer, promoting broader exploration of the feature space during training and preventing early convergence to narrow, non-generalizable filters. The temporal convolution layer now contributes more effectively to shaping temporally coherent activations, as reflected in the greater intra-class compactness and inter-class separation seen in the later layers. Crucially, the transformer stage benefits most from this enhanced input diversity. By attending over richer, noise-stabilized temporal embeddings, the transformer can more effectively capture long-range dependencies and context-sensitive spectral patterns associated with motor imagery. This synergy between noise-conditioned convolutional encoding and transformer-based contextual refinement leads to a more disentangled and geometrically stable latent space in the final embeddings.

The contrast between $\delta = 0$ and $\delta = 0.3$ highlights how stochastic conditioning improves both the internal consistency and class-specific organization of the learned representations. While convolutional layers provide the initial temporal discrimination, the transformer layer refines these representations by aligning them across time and feature dimensions, producing embeddings that are not only less redundant but also more resilient to inter-trial variability. The joint effect is particularly beneficial for weak learners, where signal reliability is low: rather than amplifying noise, the network learns to exploit structured perturbations to reveal latent regularities that remain hidden in unconditioned training. In this context, noise infusion does not merely act as an auxiliary defense against overfitting but becomes a functional component that enhances the transformer's capacity to model and stabilize weak, low-SNR EEG representations.

Table 8
Strongest and weakest learners of Dataset I, across all model runs and architectures as per Experiment C, ranked by average validation accuracy.

	Subject ID	Accuracy [%]	SD σ [%]
Best Learnable	S_55	92.59	5.45
	S_36	91.05	7.59
	S_20	91.05	7.36
	S_76	90.12	8.73
	S_2	90.12	7.69
Least Learnable	S_13	73.46	14.57
	S_91	71.91	22.99
	S_35	70.99	21.95
	S_19	70.37	18.33
	S_3	69.75	22.16

In relation to state-of-the-art methods, Table 4 places STFNet alongside classical pipelines, CNN-based models, and more recent hybrid or attention-driven approaches for a direct comparison of the proposed model performance on this benchmarking dataset. Classical pipelines such as filter-bank common spatial patterns (FBCSP) [78] remain useful due to their interpretability and well-defined spatial–spectral priors, but they degrade considerably under low signal-to-noise conditions and show limited scalability to multiclass paradigms. CNN-based models, including ConvNet [79], EEGNet [80], and FBCNet [81], improved performance by automatically learning spatial and temporal features, yet they often rely on local receptive fields and assumptions of stationarity. These limitations reduce their effectiveness in capturing long-range dependencies, particularly for weak learners. Hybrid and attention-based architectures, such as EEG-TCNet [82], DRDA [83], and Conformer [84], incorporate temporal context and dynamic weighting to improve robustness. Nevertheless, these models generally depend on explicit artifact removal or heavy preprocessing, leaving residual variability across users and sessions. Compared with these approaches, STFNet leverages multi-head self-attention to capture long-range temporal–spatial dependencies while integrating controlled Gaussian noise infusion to exploit stochastic resonance, which enhances subthreshold neural components rather than discarding them. This design yields clear advantages: STFNet achieved the highest mean accuracy and Kappa with reduced standard deviation across subjects (as shown in Table 3), and demonstrated particular improvements in weak learners (e.g., A02 and A05). In Experiment C, involving 106 subjects, STFNet maintained high accuracy and Kappa while also proving robust to variations in noise infusion and data augmentation settings (Tables 5–7). These results indicate that STFNet not only surpasses classical and CNN-based pipelines but also addresses limitations of attention-based hybrids by explicitly enhancing weak neural patterns through stochastic resonance.

An important aspect of STFNet is the complementary interaction between its convolutional and Transformer modules. The convolutional front-end extracts local spatiotemporal features, such as ERD/ERS dynamics within mu, beta, and alpha bands localized to motor-related electrodes, providing robust short-range representations that are less sensitive to noise. The Transformer layers then operate on these representations to capture long-range temporal context and cross-channel dependencies, integrating distributed activity patterns that CNN filters alone cannot resolve. This division of labor explains why CNN-only

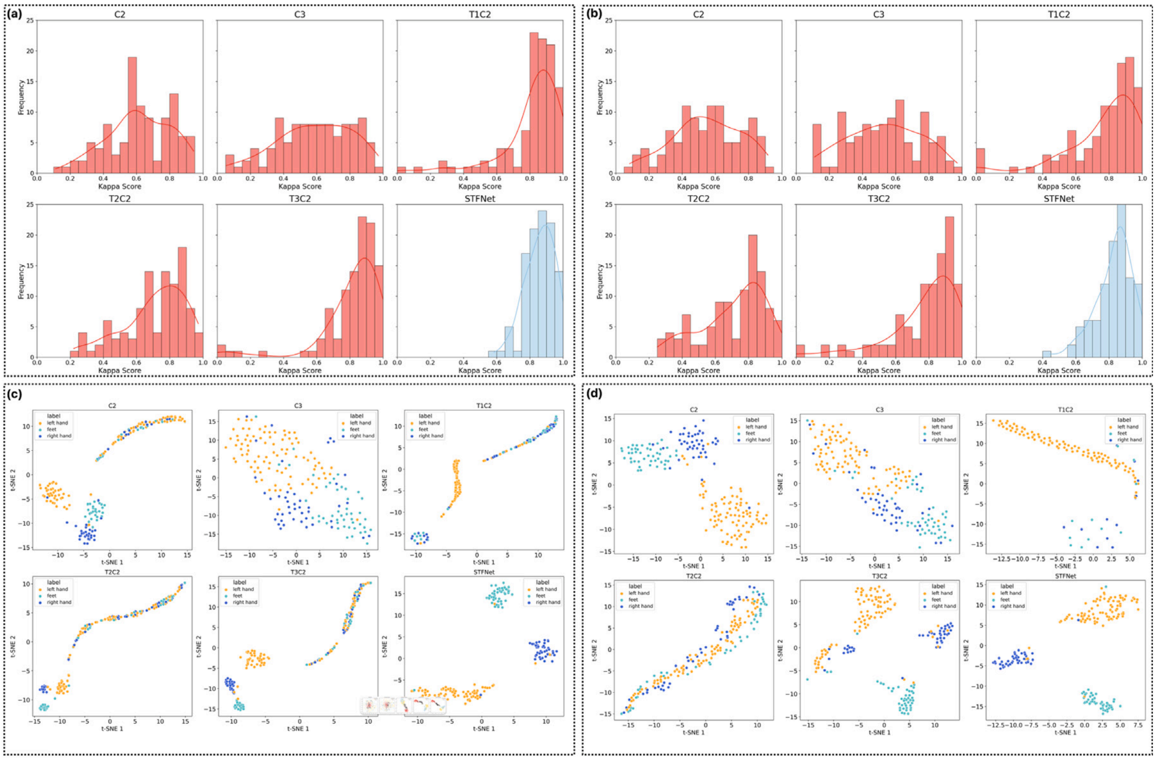

Fig. 9. **(a, b)** Distribution of Cohen's Kappa κ across all individuals of Dataset I (cross-subject ablation study) for configurations as per Tables 5 and 7, respectively. **(c, d)** t-SNE representation for MI-classification ablation study as per Table 7 for Subject 3 of Dataset I for during model training and validation, respectively.

models perform competitively but show higher variance, whereas augmenting them with Transformer blocks significantly improves mean accuracy and stability (ANOVA $F = 12.675$, $p = 0.000967$; Tables 5 and 7). Notably, subjects previously identified as weak learners (e.g., A02, A05) benefit most, indicating that global attention mechanisms can amplify subtle local features extracted by the convolutional layers, leading to more robust classification across heterogeneous populations.

The enhanced performance of STFNet is further supported by the lower standard deviations and higher Kappa scores, pointing to greater robustness across varying subjects. While traditional models like C2 and C3 struggle to consistently differentiate between MI tasks for weaker learners, as seen in the broader spread of points in Fig. 2, STFNet significantly narrows this spread, reflecting more reliable classification results. This robustness, achieved through the combination of convolutional and transformer layers with controlled noise infusion, mitigates the inter-subject variability that has historically been a major challenge in EEG-based BCI systems. The tight clustering of task-specific data points for STFNet indicates that the model is better able to generalize across subjects, including those previously deemed less learnable, supporting its application in real-world BCI settings.

Additionally, the integration of stochastic resonance plays a crucial role in amplifying weak EEG signals, impacting both the overall accuracy and consistency of STFNet's performance. The use of controlled noise, particularly for traditionally challenging subjects, allows the system to enhance the detection of weak or faint signals that other models struggle to classify accurately. This technique is especially useful in scenarios involving low signal-to-noise ratio (SNR) EEG data, which is a common characteristic in BCI illiterate users. The results from Table 7 confirm that STFNet continues to outperform other models, even under modified noise infusion and reduced data augmentation conditions, demonstrating that the model maintains its superior performance in a variety of experimental setups.

Fig. 12 presents layer-wise activation heatmaps for a representative non-weak learner (Subject A03) and a weak learner (Subject A02), illustrating the effect of the stochastic noise infusion mechanism $\delta = 0.3$ on feature representations across the model's hierarchical stages. Each column pair compares activations without noise (left) and with noise infusion (right), spanning the initial input projection layer, the temporal convolution stage, and the final embedding space after the transformer encoder. The visualization provides insight into how noise conditioning affects representational diversity, activation stability, and inter-class separability across different model depths.

For the non-weak learner (A03), activation patterns in the noise-free condition already exhibit moderate class-specific differentiation, particularly at the temporal convolution and transformer levels. However, under the $\delta = 0.3$ condition, a clearer spatial structuring of activations emerges across both early and late layers, with increased contrast and distributed feature engagement across channels and embedding dimensions. This indicates that controlled stochastic perturbation enhances representational richness without degrading class-dependent organization, effectively regularizing the feature space and mitigating over-reliance on dominant spatial sources. The result is a more balanced and discriminative embedding, which aligns with the improved classification stability observed in noise-infused configurations.

In contrast, the weak learner (A02) exhibits more uniform and less distinct activation patterns in the absence of noise, suggesting limited neural feature variability and insufficient separability across motor imagery classes. Following noise infusion, the activations display greater heterogeneity, particularly within the transformer embeddings, where feature contrast between classes becomes more pronounced. This transformation implies that the stochastic noise acts as a representational catalyst, stimulating the network to explore a broader feature manifold rather than converging to narrow, suboptimal activation subspaces. The improvement is most evident in the final embedding layer, where the enhanced dispersion of activations likely contributes to better generalization and reduced overfitting to spurious low-level correlations.

Overall, these findings support the hypothesis that noise infusion promotes more robust and adaptive feature learning across hierarchical

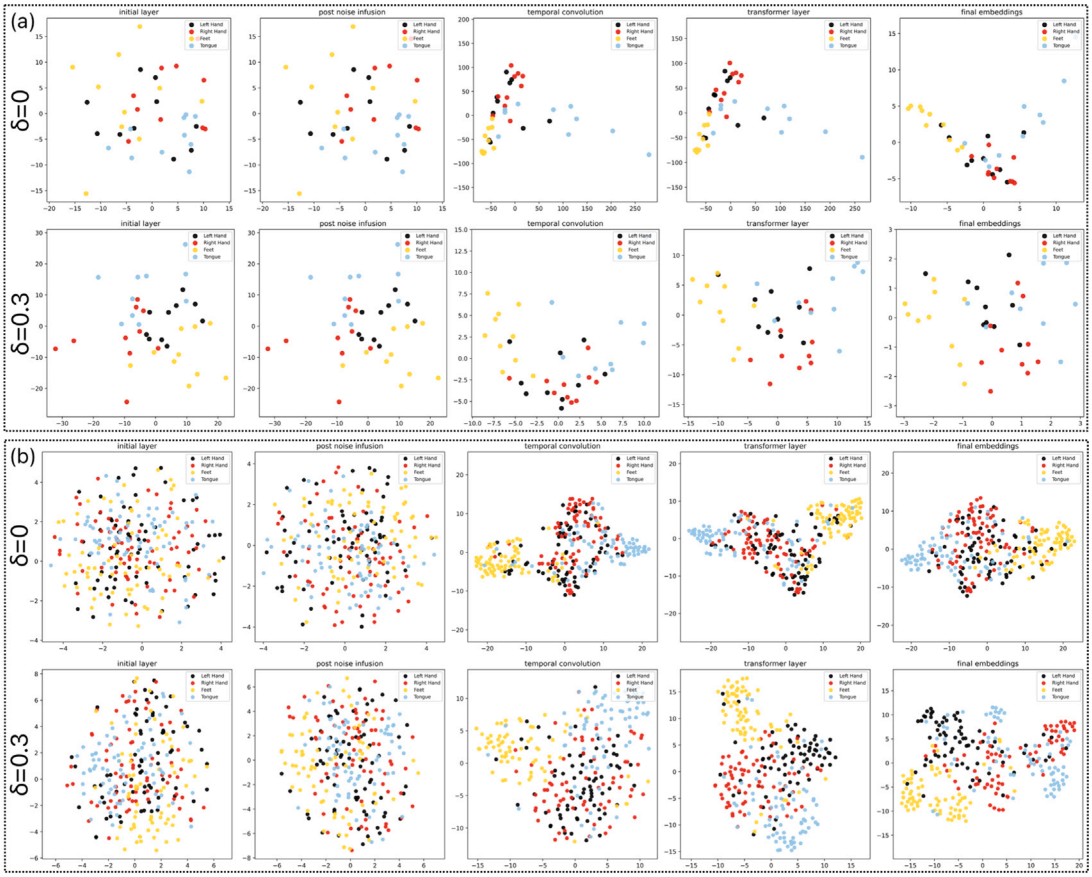

Fig. 10. Feature evolution and t-SNE representations after different network layers with strong discriminative power of STFNet (here: Subject A02, weak learner). (a) Feature evolution for configuration without and with noise infusion layer, respectively. (b) t-SNE evolution over layers for no noise infusion and active noise infusion, respectively.

layers, particularly benefitting weaker subjects whose baseline activations lack discriminative structure. The transformer layer, in particular, appears to capitalize on the noise-induced variability by refining and amplifying relevant task-specific components, thereby improving signal utility for downstream classification.

Overall, the results support the conclusion that STFNet offers a robust and reliable solution for EEG-based BCI systems. The model's ability to consistently outperform traditional approaches, even in subjects considered less learnable, highlights its potential for widespread use in clinical and private applications. Furthermore, its adaptability to various hyperparameter settings and reliance on Apple Silicon chip architecture for efficient processing underscore STFNet's viability for use in environments where computational resources are limited. Using the best inter-subject performances as per the ablation study as presented in Table 6, the proposed method can achieve validation accuracies of up to 0.879 ± 0.085 on the BCI Competition IV 2a dataset, firmly outperforming existing approaches not only in terms of validation accuracy but also kappa metrics and lower standard deviations. Future research could explore further refinements to the STFNet architecture, potentially incorporating evolutionary algorithms as recently introduced [72] to enhance its robustness and performance across an even broader range of users, including those with more severe BCI illiteracy.

6. Conclusion

This study presents a comprehensive evaluation of various neural network architectures for EEG-based BCI systems, emphasizing the significance of integrating transformer blocks and enhanced noise-handling mechanisms. Through systematic experimentation across two datasets with a total of 115 subjects, the proposed hybrid systems and STFNet demonstrated superior classification accuracy and robustness, notably improving the performance of traditionally weak learners or "BCI illiterate" subjects. By leveraging stochastic resonance and advanced temporal processing modules, STFNet achieved higher Kappa scores and reduced variability across subjects, highlighting its capacity to generalize more effectively than existing models like Conformer and EEG-TCNet.

The analysis of the presented results indicates that introducing transformer blocks yields statistically significant improvements in BCI performance, with diminishing returns beyond a certain number of blocks. STFNet, with its optimal integration of transformers, convolutional layers, and noise infusion techniques, consistently exhibited a concentrated distribution of high Kappa values, reinforcing its stability and reliability across varying experimental conditions. Furthermore, the model's resilience under different hyperparameter configurations underscores its adaptability in handling complex, noisy EEG data, particularly in subjects previously considered challenging for BCI applications.

Additionally, the feasibility of implementing the proposed approach in both clinical and private settings where extensive GPU resources may not be available has been successfully demonstrated. By utilizing Apple Silicon chips with Metal programming, we have shown that high-performance BCI systems can be effectively deployed on more accessible hardware platforms, paving the way for broader adoption in various real-world applications.

In summary, the proposed STFNet architecture marks a substantial advancement in EEG-based BCI systems, demonstrating that with the right combination of architectural enhancements and noise-handling

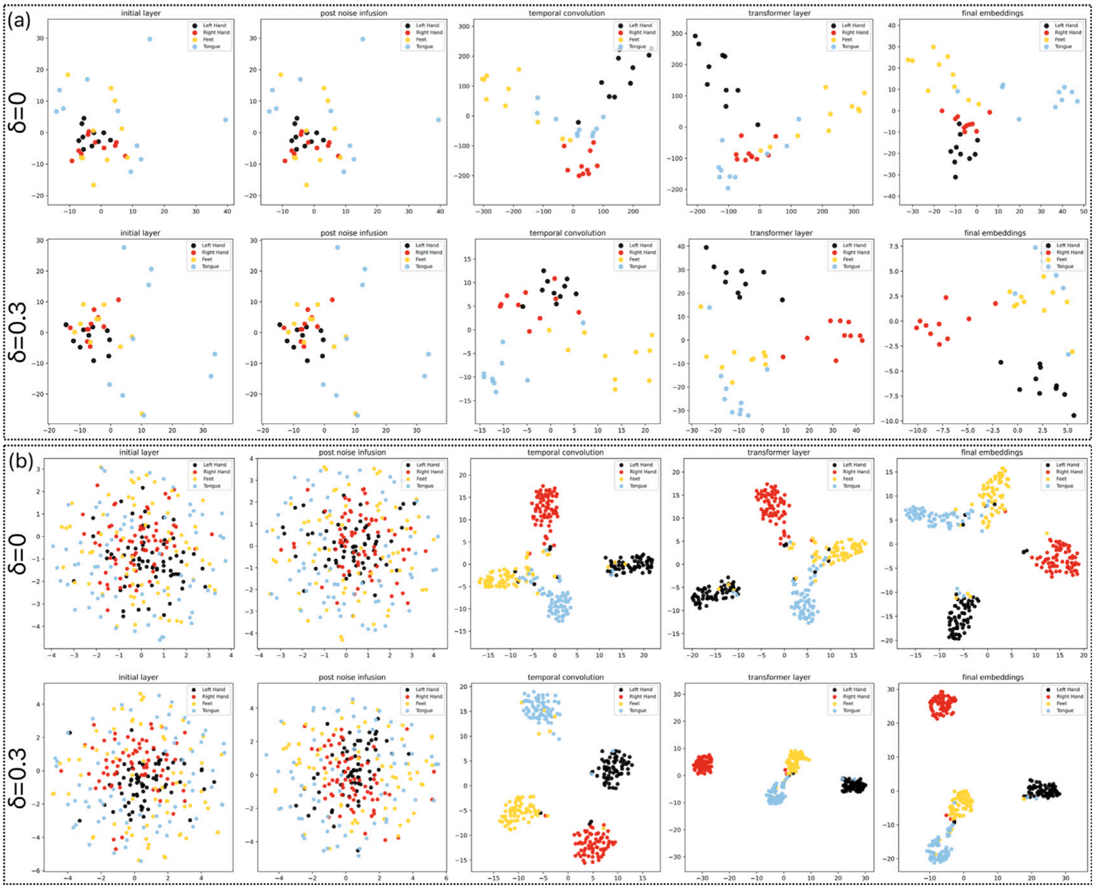

Fig. 11. Feature evolution and t-SNE representations after different network layers with strong discriminative power of STFNet (here: Subject 3 of Dataset II). (a) Feature evolution for configuration without and with noise infusion layer, respectively. (b) t-SNE evolution over layers for no noise infusion and active noise infusion, respectively.

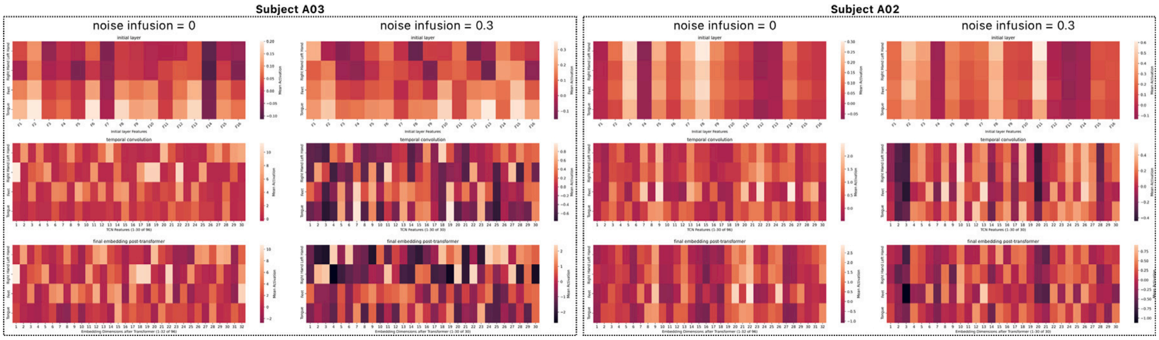

Fig. 12. Class-wise mean activation of features via feature heatmaps during different stages while converging through the proposed network architecture (Dataset II). (a) Ablation study results for A03 (non-weak learner) with different noise level configurations. (b) Same metrics as per (a) for Subject A02 (weak learner). Top row: Initial layer. Center row: post temporal convolution. (c) Bottom row: post transformer layer.

strategies, it is possible to improve the usability of BCIs across a broader range of individuals.

In the future, it is paramount to investigate the settings of such hybrid systems, as there remains room for improvement through the adjustment of layer types, hyperparameters, and other configurations. Future work may involve employing generative or self-organizing methodologies, such as evolutionary algorithms, to enhance the robustness of transformer-based hybrid systems even further, particularly for weak learners or BCI-illiterate subjects.

CRediT authorship contribution statement

Maximilian Achim Pfeffer: Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Johnny Kwok Wai Wong:** Writing – review & editing, Supervision, Investigation. **Sai Ho Ling:** Writing – review & editing, Supervision, Project administration, Investigation, Funding acquisition, Conceptualization.

Ethics approval and consent to participate

Ethics approval and consent of participants for all data recordings were collected by the respective researchers as per their original works [60,61].

Ethics statement

This research did not involve human participants, animal subjects, or sensitive data requiring formal ethical approval.

Declaration of competing interest

The authors declare that there are no conflicts of interest.

Data availability

All datasets analyzed in this study are publicly available. Dataset I (PhysioNet EEG Motor Movement/Imagery Dataset) can be found at https://physionet.org/content/eegmmidb/1.0.0/. Dataset II (Data sets 2a of Berlin BCI Competition IV) can be found at https://www.bbci.de/competition/iv/.

References

[1] B. Basics, The Life and Death of a Neuron, Office of Communications and Public Liaison, National Institute of Neurological Disorders and Stroke, Bethesda, MD, USA, 2002.
[2] F.S. Tyner, J.R. Knott, Fundamentals of EEG Technology: Basic Concepts and Methods, vol. 1, Lippincott Williams & Wilkins, 1983.
[3] S. Moghimi, A. Kushki, A. Marie Guerguerian, T. Chau, A review of EEG-based brain-computer interfaces as access pathways for individuals with severe disabilities, Assist. Technol. 25 (2) (2013) 99–110.
[4] Z. Khademi, F. Ebrahimi, H.M. Kordy, A review of critical challenges in MI-BCI: From conventional to deep learning methods, J. Neurosci. Methods 383 (2023) 109736.
[5] A. Haider, A brief review of signal processing for EEG-based BCI: Approaches and opportunities, in: 2021 IEEE International Conference on Electro Information Technology, EIT, IEEE, 2021, pp. 389–394.
[6] C. Maswanganyi, C. Tu, P. Owolawi, S. Du, Overview of artifacts detection and elimination methods for BCI using EEG, in: 2018 IEEE 3rd International Conference on Image, Vision and Computing, ICIVC, IEEE, 2018, pp. 832–836.
[7] R. Janapati, S.S. Reddy, G. Anitha, R. Shivani, V. Sreya, Challenges exist in translating brain signals into words using brain-computer interfaces (BCIs), in: Medical Robotics and AI-Assisted Diagnostics for a High-Tech Healthcare Industry, IGI Global, 2024, pp. 144–161.
[8] M.F. Mridha, S.C. Das, M.M. Kabir, A.A. Lima, M.R. Islam, Y. Watanobe, Brain-computer interface: Advancement and challenges, Sensors 21 (17) (2021) 5746.
[9] C. Vidaurre, B. Blankertz, Towards a cure for BCI illiteracy, Brain Topogr. 23 (2010) 194–198.
[10] B.Z. Allison, C. Neuper, Could anyone use a BCI? in: Brain-Computer Interfaces: Applying Our Minds to Human-Computer Interaction, Springer, 2010, pp. 35–54.
[11] S. Saha, M. Baumert, Intra-and inter-subject variability in EEG-based sensorimotor brain computer interface: a review, Front. Comput. Neurosci. 13 (2020) 87.
[12] S. Pérez-Velasco, E. Santamaría-Vázquez, V. Martínez-Cagigal, D. Marcos-Martínez, R. Hornero, Eegsym: Overcoming inter-subject variability in motor imagery based BCIs with deep learning, IEEE Trans. Neural Syst. Rehabil. Eng. 30 (2022) 1766–1775.
[13] J. Šťastný, P. Sovka, M. Kostilek, Overcoming inter-subject variability in BCI using EEG-based identification, Radioengineering 23 (1) (2014).
[14] R. Boostani, B. Graimann, M.H. Moradi, G. Pfurtscheller, A comparison approach toward finding the best feature and classifier in cue-based BCI, Med. Biol. Eng. Comput. 45 (2007) 403–412.
[15] A. Vaswani, Attention is all you need, Adv. Neural Inf. Process. Syst. (2017).
[16] A. Hameed, R. Fourati, B. Ammar, A. Ksibi, A.S. Alluhaidan, M.B. Ayed, H.K. Khleaf, Temporal–spatial transformer based motor imagery classification for BCI using independent component analysis, Biomed. Signal Process. Control. 87 (2024) 105359.
[17] J. Luo, Y. Wang, S. Xia, N. Lu, X. Ren, Z. Shi, X. Hei, A shallow mirror transformer for subject-independent motor imagery BCI, Comput. Biol. Med. 164 (2023) 107254.
[18] Y. Song, Q. Zheng, B. Liu, X. Gao, EEG conformer: Convolutional transformer for EEG decoding and visualization, IEEE Trans. Neural Syst. Rehabil. Eng. 31 (2022) 710–719.
[19] C. Zhang, X. Deng, S.H. Ling, Next-gen medical imaging: U-Net evolution and the rise of transformers, Sensors 24 (14) (2024) 4668.
[20] S. Sakhavi, C. Guan, S. Yan, Learning temporal information for brain-computer interface using convolutional neural networks, IEEE Trans. Neural Netw. Learn. Syst. 29 (11) (2018) 5619–5629.
[21] D. Shao, Y. Zhang, G. Xu, EEG-TCNet: Temporal convolutional network for EEG-based brain–computer interfaces, in: Proceedings of the International Conference on Neural Information Processing, 2020.
[22] Z. Huang, R. Wang, Y. Chen, EEG-ITNet: An inception-transformer network for EEG-based brain–computer interfaces, IEEE Trans. Neural Syst. Rehabil. Eng. 29 (2021) 2153–2165.
[23] J. Liu, C. Li, M. Wang, A convolution-transformer hybrid network for EEG-based cognitive load classification, Biomed. Signal Process. Control. 72 (2022) 103339.
[24] H. Zhao, Q. Zheng, K. Ma, H. Li, Y. Zheng, Deep representation-based domain adaptation for nonstationary EEG classification, IEEE Trans. Neural Netw. Learn. Syst. 32 (2) (2020) 535–545.
[25] L. Qiu, J. Li, B. Chen, Time-series transformer for cross-subject EEG decoding, in: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2021.
[26] M.A. Pfeffer, S.S.H. Ling, J.K.W. Wong, Exploring the frontier: Transformer-based models in EEG signal analysis for brain-computer interfaces, Comput. Biol. Med. (2024) 108705.
[27] B. Abibullaev, A. Keutayeva, A. Zollanvari, Deep learning in EEG-based BCIs: a comprehensive review of transformer models, advantages, challenges, and applications, IEEE Access (2023).
[28] A. Keutayeva, B. Abibullaev, Data constraints and performance optimization for transformer-based models in EEG-based brain-computer interfaces: A survey, IEEE Access (2024).
[29] L. Bianchi, S. Sami, A. Hillebrand, I.P. Fawcett, L.R. Quitadamo, S. Seri, Which physiological components are more suitable for visual ERP based brain–computer interface? A preliminary MEG/EEG study, Brain Topogr. 23 (2010) 180–185.
[30] R. Fazel-Rezai, W. Ahmad, P300-based brain-computer interface paradigm design, Recent. Adv. Brain-Comput. Interface Syst. (2011) 83–98.
[31] G. Pfurtscheller, F.L. Da Silva, Event-related EEG/MEG synchronization and desynchronization: basic principles, Clin. Neurophysiol. 110 (11) (1999) 1842–1857.
[32] E.W. Sellers, E. Donchin, A P300-based brain–computer interface: initial tests by ALS patients, Clin. Neurophysiol. 117 (3) (2006) 538–548.
[33] G. Pfurtscheller, C. Neuper, W. Mohl, Event-related desynchronization (ERD) during visual processing, Int. J. Psychophysiol. 16 (2–3) (1994) 147–153.
[34] A. Furdea, S. Halder, D. Krusienski, D. Bross, F. Nijboer, N. Birbaumer, A. Kübler, An auditory oddball (p300) spelling system for brain-computer interfaces, Psychophysiology 46 (3) (2009) 617–625.
[35] B. Blankertz, S. Lemm, M. Treder, S. Haufe, K.-R. Müller, Single-trial analysis and classification of ERP components—a tutorial, NeuroImage 56 (2) (2011) 814–825.
[36] E.V. Friedrich, D.J. McFarland, C. Neuper, T.M. Vaughan, P. Brunner, J.R. Wolpaw, A scanning protocol for a sensorimotor rhythm-based brain–computer interface, Biol. Psychol. 80 (2) (2009) 169–175.
[37] M. Ahn, H. Cho, S. Ahn, S.C. Jun, High theta and low alpha powers may be indicative of BCI-illiteracy in motor imagery, PloS One 8 (11) (2013) e80886.
[38] M.C. Thompson, Critiquing the concept of BCI illiteracy, Sci. Eng. Ethics 25 (4) (2019) 1217–1233.
[39] G. Edlinger, B.Z. Allison, C. Guger, How many people can use a BCI system? Clin. Syst. Neurosci. (2015) 33–66.
[40] T. Wang, S. Du, E. Dong, A novel method to reduce the motor imagery BCI illiteracy, Med. Biol. Eng. Comput. (2021) 1–13.
[41] S. Halder, C.A. Ruf, A. Furdea, E. Pasqualotto, D. De Massari, L. van der Heiden, M. Bogdan, W. Rosenstiel, N. Birbaumer, A. Kübler, et al., Prediction of P300 BCI aptitude in severe motor impairment, PloS One 8 (10) (2013) e76148.
[42] T. Halder, S. Talwar, A.K. Jaiswal, A. Banerjee, Quantitative evaluation in estimating sources underlying brain oscillations using current source density methods and beamformer approaches, Eneuro 6 (4) (2019).
[43] A. Myrden, T. Chau, Effects of user mental state on EEG-BCI performance, Front. Hum. Neurosci. 9 (2015) 308.
[44] S. Li, J. Duan, Y. Sun, X. Sheng, X. Zhu, J. Meng, Exploring fatigue effects on performance variation of intensive brain–computer interface practice, Front. Neurosci. 15 (2021) 773790.
[45] W. Skrandies, Global field power and topographic similarity, Brain Topogr. 3 (1990) 137–141.
[46] M. Fatourechi, A. Bashashati, R.K. Ward, G.E. Birch, EMG and EOG artifacts in brain computer interface systems: A survey, Clin. Neurophysiol. 118 (3) (2007) 480–494.
[47] C.Y. Jung, S.S. Saikiran, A review on EEG artifacts and its different removal technique, Asia-Pacific J. Converg. Res. Interchang. 2 (4) (2016) 43–60.
[48] R.J. Croft, R.J. Barry, Removal of ocular artifact from the EEG: a review, Neurophysiol. Clinique/Clinical Neurophysiol. 30 (1) (2000) 5–19.

[49] I. Winkler, S. Haufe, M. Tangermann, Automatic classification of artifactual ICA-components for artifact removal in EEG signals, Behav. Brain Funct. 7 (2011) 1–15.

[50] F. Lotte, C. Jeunet, Towards improved BCI based on human learning principles, in: The 3rd International Winter Conference on Brain-Computer Interface, IEEE, 2015, pp. 1–4.

[51] G. Huang, Z. Zhao, S. Zhang, Z. Hu, J. Fan, M. Fu, J. Chen, Y. Xiao, J. Wang, G. Dan, Discrepancy between inter-and intra-subject variability in EEG-based motor imagery brain-computer interface: Evidence from multiple perspectives, Front. Neurosci. 17 (2023) 1122661.

[52] L. Gammaitoni, P. Hänggi, P. Jung, F. Marchesoni, Stochastic resonance, Rev. Modern Phys. 70 (1) (1998) 223.

[53] K. Hayashi, S. De Lorenzo, M. Manosas, J. Huguet, F. Ritort, Single-molecule stochastic resonance, Phys. Rev. X 2 (3) (2012) 031012.

[54] B. Andò, S. Graziani, Stochastic Resonance: Theory and Applications, Springer Science & Business Media, 2000.

[55] F. Chapeau-Blondeau, D. Rousseau, Noise-enhanced performance for an optimal Bayesian estimator, IEEE Trans. Signal Process. 52 (5) (2004) 1327–1334.

[56] G. Winterer, M. Ziller, H. Dorn, K. Frick, C. Mulert, N. Dahhan, W. Herrmann, R. Coppola, Cortical activation, signal-to-noise ratio and stochastic resonance during information processing in man, Clin. Neurophysiol. 110 (7) (1999) 1193–1203.

[57] J.J. Collins, T.T. Imhoff, P. Grigg, Noise-mediated enhancements and decrements in human tactile sensation, Phys. Rev. E 56 (1) (1997) 923.

[58] P.E. Greenwood, M.D. McDonnell, L.M. Ward, Dynamics of gamma bursts in local field potentials, Neural Comput. 27 (1) (2014) 74–103.

[59] Y. Song, X. Jia, L. Yang, L. Xie, Transformer-based spatial-temporal feature learning for EEG decoding, 2021, arXiv preprint arXiv:2106.11170.

[60] A.L. Goldberger, L.A. Amaral, L. Glass, J.M. Hausdorff, P.C. Ivanov, R.G. Mark, J.E. Mietus, G.B. Moody, C.-K. Peng, H.E. Stanley, PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals, Circulation 101 (23) (2000) e215–e220.

[61] B. Blankertz, G. Dornhege, M. Krauledat, K.-R. Muller, V. Kunzmann, F. Losch, G. Curio, The Berlin Brain-Computer Interface: EEG-based communication without subject training, IEEE Trans. Neural Syst. Rehabil. Eng. 14 (2) (2006) 147–152.

[62] C. Kenyon, C. Capano, Apple silicon performance in scientific computing, in: 2022 IEEE High Performance Extreme Computing Conference, HPEC, IEEE, 2022, pp. 1–10.

[63] University of Technology Sydney, Faculty of Engineering and Information Technology, iHPC resources, 2025.

[64] H. Zhang, M. Cisse, Y.N. Dauphin, D. Lopez-Paz, Mixup: Beyond empirical risk minimization, in: International Conference on Learning Representations, 2018.

[65] S. Sakhavi, C. Guan, S. Yan, Learning temporal information for brain-computer interface using convolutional neural networks, IEEE Trans. Neural Netw. Learn. Syst. 29 (11) (2018) 5619–5629.

[66] R.T. Schirrmeister, J.T. Springenberg, L.D.J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, T. Ball, Deep learning with convolutional neural networks for EEG decoding and visualization, Hum. Brain Mapp. 38 (11) (2017) 5391–5420.

[67] V.J. Lawhern, A.J. Solon, N.R. Waytowich, S.M. Gordon, C.P. Hung, B.J. Lance, EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces, J. Neural Eng. 15 (5) (2018) 056013.

[68] T.M. Ingolfsson, M. Hersche, X. Wang, N. Kobayashi, L. Cavigelli, L. Benini, EEG-TCNet: An accurate temporal convolutional network for embedded motor-imagery brain–machine interfaces, in: 2020 IEEE International Conference on Systems, Man, and Cybernetics, SMC, IEEE, 2020, pp. 2958–2965.

[69] K.K. Ang, Z.Y. Chin, C. Wang, C. Guan, H. Zhang, Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b, Front. Neurosci. 6 (2012) 39.

[70] R. Mane, E. Chew, K. Chua, K.K. Ang, N. Robinson, A.P. Vinod, S.-W. Lee, C. Guan, Fbcnet: A multi-view convolutional neural network for brain-computer interface, 2021, arXiv preprint arXiv:2104.01233.

[71] M.A. Pfeffer, S.H. Ling, Evolving optimised convolutional neural networks for lung cancer classification, Signals 3 (2) (2022) 284–295.

[72] M.A. Pfeffer, A.H.P. Nguyen, K. Kim, J.K.W. Wong, S.H. Ling, Evolving optimized transformer-hybrid systems for robust BCI signal processing using genetic algorithms, Biomed. Signal Process. Control. 108 (2025) 107883.

[73] N. Robinson, K.P. Thomas, A.P. Vinod, Canonical correlation analysis of EEG for classification of motor imagery, in: 2017 IEEE International Conference on Systems, Man, and Cybernetics, SMC, IEEE, 2017, pp. 2317–2321.

[74] M. Bakker, F.P. De Lange, R.C. Helmich, R. Scheeringa, B.R. Bloem, I. Toni, Cerebral correlates of motor imagery of normal and precision gait, Neuroimage 41 (3) (2008) 998–1010.

[75] Siuly, Y. Li, P. Wen, Comparisons between motor area EEG and all-channels EEG for two algorithms in motor imagery task classification, Biomed. Eng.: Appl. Basis Commun. 26 (03) (2014) 1450040.

[76] J. Herding, S. Ludwig, A. von Lautz, B. Spitzer, F. Blankenburg, Centro-parietal EEG potentials index subjective evidence and confidence during perceptual decision making, NeuroImage 201 (2019) 116011.

[77] A.S. Aghaei, M.S. Mahanta, K.N. Plataniotis, Separable common spatio-spectral patterns for motor imagery BCI systems, IEEE Trans. Biomed. Eng. 63 (1) (2015) 15–29.

[78] K.K. Ang, Z.Y. Chin, H. Zhang, C. Guan, Filter bank common spatial pattern (FBCSP) in brain-computer interface, in: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), IEEE, 2008, pp. 2390–2397.

[79] R.T. Schirrmeister, J.T. Springenberg, L.D.J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, T. Ball, Deep learning with convolutional neural networks for EEG decoding and visualization, Hum. Brain Mapp. 38 (11) (2017) 5391–5420.

[80] V.J. Lawhern, A. Solon, N.R. Waytowich, S.M. Gordon, C.P. Hung, B.J. Lance, EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces, J. Neural Eng. 15 (5) (2018) 056013.

[81] Y. Zhang, Y. Liu, Z. Zhou, N. Zhang, X. Jiang, E. Yin, Y. Zhang, FBCNet: A multi-view convolutional neural network for brain-computer interface, IEEE Trans. Neural Syst. Rehabil. Eng. 29 (2021) 302–310.

[82] T.M. Ingolfsson, M. Hersche, X. Wang, N. Kobayashi, L. Cavigelli, L. Benini, EEG-TCNet: An accurate temporal convolutional network for embedded motor-imagery brain–machine interfaces, in: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society, EMBC, IEEE, 2020, pp. 1272–1275.

[83] Z. Li, L. Cui, Z. Li, Z. Ma, J. Chen, Z. Zhang, Dynamic recurrent dual-attention network for EEG-based emotion recognition, IEEE Trans. Affect. Comput. (2022).

[84] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, R. Pang, Conformer: Convolution-augmented transformer for speech recognition, in: Proc. Interspeech, 2020.