Deep Learning-Driven Decision Fusion: Spatio-Spectrogram Features for Inner Speech Recognition From Electroencephalogram Signals

Hussna E. M. Abdalla, Hamidon Basri, Ishak Aris, Abdul Hanif Yusof, Muhammad S. Adha, Hisham Neyaz, Amna Saga, Sadiq H. Abdulhussain, Basheera M. Mahmmod, Nurbek Saparkhojayev, and Syed Abdul Rahman Al-Haddad, *Senior Member, IEEE*

***Abstract*—Inner speech recognition using electroencephalogram (EEG) signals shows strong potential for developing assistive communication technologies. Existing methods often process spatial and temporal features separately, lack interpretability, and are usually tested on a single dataset, limiting their generalization. This study proposes a dual-branch deep learning framework that combines spatial features extracted through common spatial patterns (CSPs) with spectral-temporal features derived from multitaper spectrograms, using convolutional and long short-term memory networks. The model was evaluated on two public datasets, achieving classification accuracies of 89.99% and 92.47% in subject-dependent experiments. Subject-independent evaluation using leave-one-subject-out cross-validation yielded reduced accuracies of 26.20% and 20.47%, reflecting intersubject variability. Interpretability analyses using saliency maps, gradient-weighted class activation mapping, and feature contribution ratios highlighted physiologically meaningful patterns related to model decisions. The proposed method demonstrates strong performance and interpretability for subject-dependent inner speech recognition; while future work will focus on increasing data diversity and improving subject-independent generalization. This study contributes to the development of reliable and explainable EEG-based inner speech decoding for communication applications.**

***Index Terms*—Brain–computer interface (BCI), common spatial pattern (CSP), deep learning, electroencephalogram, fusion technique, inner speech, mental speech, spectrogram features.**

Received 25 July 2025; revised 12 November 2025, 2 March 2026, and 27 April 2026; accepted 1 May 2026. This work was supported in part by the Research Management Centre at Universiti Putra Malaysia under Grant Putra 9695500, in part by The Organization for Women in Science for the Developing World (OWSD), and in part by the Swedish International Development Cooperation Agency (SIDA). This article was recommended by Associate Editor Tiago H. Falk. *(Corresponding authors: Hussna E. M. Abdalla; Syed Abdul Rahman Al-Haddad.)*

Hussna E. M. Abdalla is with the Department of Computer and Communication Systems Engineering, Universiti Putra Malaysia, Serdang 43400, Malaysia, and also with the Department of Biomedical Engineering, Sudan International University, Khartoum 11111, Sudan (e-mail: hussna_elnoor@hotmail.com).

Hamidon Basri and Abdul Hanif Yusof are with the Department of Neurology, Faculty of Medicine and Health Sciences, Universiti Putra Malaysia, Serdang 43400, Malaysia (e-mail: hamidon@upm.edu.my; ahanifkhan@upm.edu.my).

Ishak Aris is with the Department of Electrical and Electronic Engineering, Universiti Putra Malaysia, Serdang 43400, Malaysia (e-mail: ishak_ar@upm.edu.my).

Muhammad S. Adha and Syed Abdul Rahman Al-Haddad are with the Department of Computer and Communication Systems Engineering, Universiti Putra Malaysia, Serdang 43400, Malaysia, and also with the Institut Penyelidikan Matematik (INSPEM), Universiti Putra Malaysia, Serdang 43400, Malaysia (e-mail: shaufiladha@upm.edu.my; sar@upm.edu.my).

Hisham Neyaz is with the Department of Computer and Communication Systems Engineering, Universiti Putra Malaysia, Serdang 43400, Malaysia (e-mail: thisishisham8@gmail.com).

Amna Saga is with the Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Serdang 43400, Malaysia (e-mail: gs63748@student.upm.edu.my).

Sadiq H. Abdulhussain and Basheera M. Mahmmod are with the Department of Computer Engineering, University of Baghdad, Baghdad 10071, Iraq (e-mail: sadiqhabeeb@coeng.uobaghdad.edu.iq; basheera.m@coeng.uobaghdad.edu.iq).

Nurbek Saparkhojayev is with the Rudny Industrial University, Rudny 11150, Kazakhstan (e-mail: nursp81@gmail.com).

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/THMS.2026.3690175.

Digital Object Identifier 10.1109/THMS.2026.3690175

I. Introduction

ELECTROENCEPHALOGRAM (EEG) signals are electrical recordings that reflect intentional neural activity in the human brain. These signals are widely used to study cognitive functions, including inner speech, a covert process in which thoughts are generated as internal auditory representations without vocalization [1], [2]. Decoding inner speech from EEG has gained attention for its potential applications in brain–computer interface (BCI) systems, especially for individuals with severe motor or speech impairments [3]. Inner speech originates in brain regions involved in language processing, mainly Wernicke's area and Broca's area, where neural activity increases during speech-related tasks [4], [5]. Despite its usefulness, EEG faces challenges due to low spatial resolution and overlapping neural signals, limiting decoding accuracy [6]. Reliable recognition requires capturing temporal, spectral, and spatial characteristics of the signal and involves two key steps: 1) extracting meaningful features and 2) applying appropriate classification algorithms [7].

Deep learning has recently become a leading approach for EEG analysis, showing strong performance in imagined speech, emotion recognition, and inner speech decoding [8], [9]. These models handle the high dimensionality of EEG signals by automatically extracting complex and discriminative features [10]. Convolutional neural networks (CNNs) capture spatial dependencies across channels, while the long short-term memory (LSTM) model temporal dynamics. Combining these architectures enables simultaneous learning of spatial, temporal, and

spectral patterns [11]. Various methods have been proposed to enhance inner speech recognition, including wavelet scattering with LSTM- recurrent neural networks (RNNs) [12]; bimodal fusion of functional magnetic resonance (fMRI) and EEG data [13]; temporal convolutional networks integrated with CNNs [14]; and features based on mutual information, phase locking value, and entropy using convolutional and bidirectional LSTMs [15]. Explainable artificial intelligence and attention-based models are increasingly used to reveal neural activity patterns and improve interpretability in clinical EEG classification [16], [17].

Other advanced models like support vector machine (SVM), extreme gradient boosting (XGBoost), LSTM, and bidirectional long short-term memory (BiLSTM) for inner speech recognition use power spectral density (PSD) features and achieve notable accuracy improvements, especially with the BiLSTM model [2]. One study uses gradient and perturbation-based explainability methods to uncover neural differences in major depressive disorder, providing effective spatio-spectral analysis and deeper insights into the disorder's neural underpinnings. These studies highlight the potential of advanced EEG classification techniques for both communication enhancement and medical diagnosis [18]. Moreover, models, such as lightweight convolutional network (LConvNet) and electroencephalography network (EEGNet) have been effective in EEG classification by integrating spatial and temporal feature learning. LConvNet combines CNN and LSTM layers to distinguish pathological EEG patterns, while EEGNet has shown robust performance in subject-independent emotion recognition tasks [19], [20]. Despite recent advancements, key challenges remain in EEG-based inner speech recognition. Many existing models process either spatial or temporal features independently, overlooking the complex interactions between spatial and spectral patterns inherent in EEG signals. Furthermore, a lack of interpretability mechanisms limits transparency in model decisions, which is critical for reliable deployment in real-world BCI systems. Prior studies often evaluate their models using only one type of dataset and lack robust validation strategies. In addition, most existing work fails to explore optimal feature extraction settings, such as the number of spatial filters in CSPs or to justify parameter selection in an interpretable manner. Electrode selection is often arbitrary or overlooked, further constraining the quality of spatial representation and overall model performance.

This study addresses the challenges of EEG-based inner speech decoding by proposing a novel fusion deep learning framework that combines spatial and spectral-temporal EEG features to enhance discriminability and interpretability. In this work, we leverage two complementary, neurobiologically informed feature types. Spatial features are extracted based on electrophysiological evidence that language processing involves a distributed cortical network encompassing the frontal, temporal, and parietal regions [21]. Spectro-temporal features are incorporated to reflect the hierarchical and dynamic organization of neural activity, where brain responses encode information from discrete phonemes to continuous spectro-temporal patterns [22]. The main contributions of this work are outlined as follows.

1) Optimization of the number of CSP components for feature extraction in inner speech decoding, identifying the 32-component configuration as the optimal choice, which led to statistically significant improvements in classification accuracy.
2) Integration of CSP-derived spatial features with multitaper spectrogram-based temporal-spectral features. A dual-branch CNN-LSTM model is utilized to extract complementary spatial and spectral-temporal EEG features, thereby enhancing inner speech decoding performance.
3) Validation on two public datasets ensures the robustness of the proposed method, while interpretability is enhanced using gradient-weighted class activation mapping (Grad-CAM), saliency maps, and contribution ratio analysis for transparent neural insights.

By integrating neurophysiological principles with advanced deep learning, this study offers a reliable, interpretable, and generalizable framework for subject-dependent inner speech decoding, laying the groundwork for future neural communication technologies.

II. Methodology

A. Dataset

This study employed two publicly available EEG datasets for inner speech, as summarized in Table I.

B. Preprocessing

The raw signals from both datasets were filtered to remove noise. A 50-Hz notch filter eliminated power line interference, while a bandpass filter (0.5–80 Hz) addressed external noise. Independent component analysis (ICA) was used to remove artifacts from eye blinks and muscle activity by correlating active electrodes with external ones. The preprocessed datasets had shapes of (2236, 128, 640) for Data_A and (640, 64, 1024) for Data_B, representing (trials, channels, sampling time points). The selection of electrodes varied between the two datasets based on their acquisition protocols and task-specific activation patterns. Data_A, acquired with a 24-b resolution and a 128-channel EEG system, was processed using spatial reduction to manage high data dimensionality. 60 electrodes were chosen from the left upper brain region, particularly Broca's area, which is involved in language processing and inner speech and exhibits heightened activity during inner speech generation [5], [25]. The electrode distribution is shown in Fig. 1. In contrast, Data_B, acquired with a 16-b resolution, retained all 64 electrodes, as the original dataset demonstrated spatially diverse activation patterns across all electrodes for different inner speech tasks. ICA and topographic projections revealed varying brain regions involved in word and semantic processing [24], indicating that reducing spatial coverage may lead to the loss of critical task-relevant information. So, all 64 electrodes were retained to preserve data integrity and capture the full variability in brain activity.

C. Techniques of Feature Extraction

1) Common Spatial Pattern: CSP remains one of the most widely adopted and effective approaches for extracting meaningful features, and was used to compute spatial features, improve the spatial resolution of the EEG [26] and pay more attention to the channels [27]. It converts the signal into a new time series

TABLE I
SUMMARY OF THE PUBLIC EEG DATASETS USED IN THIS STUDY

Category	Data_A Details	Data_B Details
Citation	[23]	[24]
Participant (P)	10 healthy adults, right-handed (P1 – P10)	4 healthy adults, right-handed (P1, P2, P3, P5) P4 was excluded due to EEG artifacts
Age Mean ± Std	34 ± 10 years, range (24 – 56) years	39 ± 8 years, range (33–51) years
Sex	6 Male, 4 Female	2 Male, 2 Female
EEG Hardware	128-channel, active (BioSemi ActiveTwo, 10–20 system); external 8-channel	64-channel, active (BioSemi ActiveTwo, 10–20 system); external 6-channel
EEG Software	256 Hz Sampling rate (SR) / 24-bit Resolution	512 Hz Sampling rate (SR) / 16-bit Resolution
EEG Protocol	Trial structure: concentration (0.5 s) → cue (0.5s) → action (2.5 s) → relax (1.0s) → rest (1.5/2.0 s)	One session per participant; trial structure: fixation (1 s) → inner-speech cue (2 s) → rest (1 s)
Total Words	Spanish words: (”arriba”, ”abajo”, ”derecha”, ”izquierda” mean (”up”, ”down”, ”right”, ”left”, respectively)	4 words / social categories: (”Child”, ”Daughter”, ”Father”, ”Wife”)
Trials per Word	40 trials per word, total 160 trials per participant	40 trials per word, total 160 trials per participant
Total Dataset	2236 trials across all word classes	640 trials across 4 participants

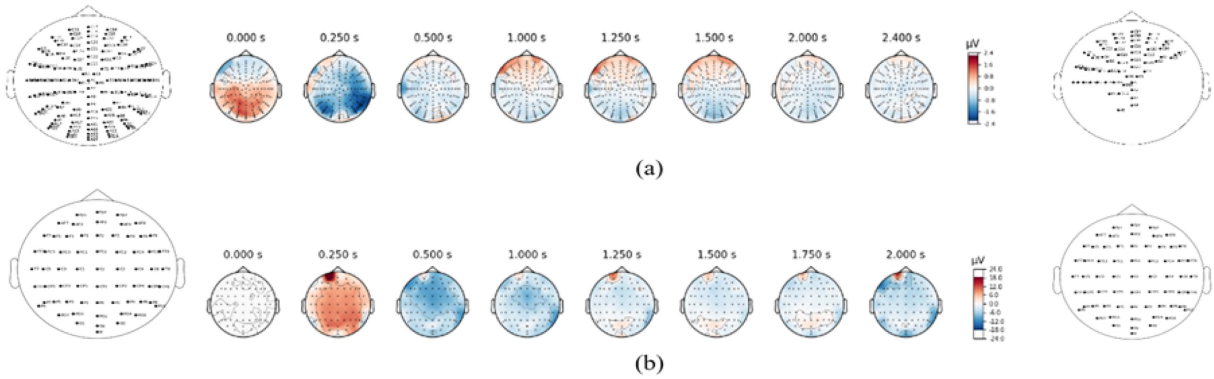

Fig. 1. Scalp topographies showing the distribution of average electrode activity for each dataset (middle plot). (a) Data_A. (b) Data_B. For each dataset, the left plot illustrates the original electrode configuration, while the right plot shows the subset of electrodes used in this study.

by maximizing the variance for one class while simultaneously minimizing it for the other, enhancing class discrimination [28]. The CSP algorithm uses the simultaneous diagonalization of two covariance matrices. In conclusion, for a single-trial EEG, the spatially filtered signal Z is provided as

$$Z = W_{\text{CSP}} X^{C \times T} \tag{1}$$

where X is a matrix of single-trial data, C is the channels' number, and T denotes the number of measurement samples per channel. W_{CSP} refers to the projection matrix used in the CSP method, defined as

$$W_{\text{CSP}} = [w_1, w_2, \ldots, w_n, w_{N-n+1}, w_{N-n}, \ldots, w_N]. \tag{2}$$

In the matrix W, each row represents a stationary spatial filter, and the columns of W^T correspond to the CSPs. N denotes the number of spatial filters, which an equal to EEG channels ($N = C$). The filtered spatial signal Z, as defined in (1), enhances the variance differences between the two EEG measurement classes. Typically, only a limited number of these spatially filtered signals are utilized as classification features. Specifically, the first and last n rows of Z are employed, denoted as Z_P, $p \in \{1 \ldots 2n\}$

$$Z_P = W_{\text{CSP}} X = [Z_1, Z_2, Z_3, \ldots, Z_{2n}]. \tag{3}$$

The number of channels that should be reduced is known as the reduction number. As a result, a select few signals from each class EEG sample matrix were chosen, which are crucial for differentiating between the two classes and multiclass, as in [29]. In this study, n was obtained by 32, so, W size = (C, 32), W^T = (32, C), X = (trials, C, T) and the results of spatial features in a size (trials × 32 × T) matrix, and last transformed of the data after calculating the average values of the number of sampling the size of the feature equal (trials × 32). The choice of 32 components was based on empirical evaluation. Several CSP configurations (16, 24, 32, 40, and 60 components) were tested, and 32 consistently yielded the highest classification accuracy, as presented in Fig. 5.

2) Time-Frequency Presentation (Spectrogram): Spectrograms of EEG signals provide a visual representation of signal power over time and frequency, facilitating detailed analysis of brain activity [30], [31]. In this study, spectral features were extracted from EEG recordings using the multitaper method implemented in the MNE-Python library. After preprocessing, the multitaper approach was applied to compute the time–frequency representation of the EEG signals. By employing multiple orthogonal tapers, this method reduces spectral leakage and improves the stability of PSD estimates across time and frequency. The EEG signals were segmented into overlapping windows and analyzed using the short-time Fourier Transform [32], mathematically expressed as

$$X(T, F) = \int_{-\infty}^{\infty} x(t) w(T - t) e^{-j 2\pi F t} \, dt. \tag{4}$$

Here, $X(T, F)$ represents the time-frequency distribution, $x(t)$ is the time-domain signal, and $w(T\text{-}t)$ denotes a window function centered around time t. The power spectrum P_{PSD} was then calculated as [33]

$$P_{\text{PSD}}(T, F) = \frac{1}{K} \sum_{k=1}^{K} |X(T, F) \cdot w_k(F)|^2. \tag{5}$$

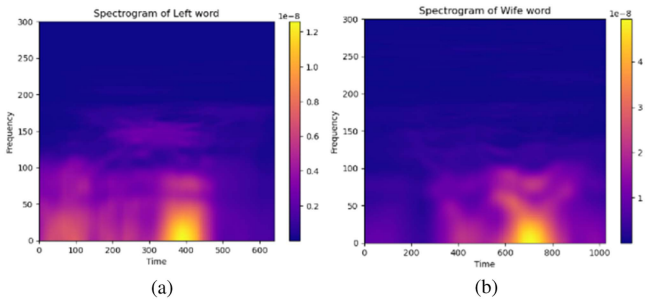

Fig. 2. Spectrogram feature of (a) Data_A ("left") and (b) Data_B ("wife").

This process produced time–frequency representations visualized as spectrograms (Fig. 2), which revealed distinct spectral patterns associated with inner speech activity, particularly in the lower frequency bands. The resulting power spectrum data were structured as (trials, channels, frequencies, time samples) for the full dataset. To emphasize spectral rather than spatial information, channels were averaged across lower frequencies, yielding final spectral features of size (trials, frequencies, time samples). These features served as inputs to the LSTM component of the fusion model.

D. Fusion Model of Classification

1) Model Architecture: This study proposes a dual-branch deep learning framework integrating CNN and LSTM architectures to simultaneously extract spatial and temporal featurs. Although both architectures have been individually applied in EEG-based studies, their neurobiologically informed fusion for inner speech decoding remains underexplored. The model leverages complementary spatial–temporal domains, reflecting the distributed and dynamic nature of language processing in the brain. The architecture, illustrated in Fig. 3, consists of two main branches. *CNN Branch:* receives CSP-derived features with 32 convolutional filters (kernel size = 5), ReLU activation, L2 regularization, 50% dropout, batch normalization, and max pooling (pool size = 1, 5). These parameters were selected through preliminary tuning to balance performance and control overfitting. *LSTM Branch:* processes time–frequency spectrogram features. It comprises three stacked LSTM layers with 64, 32, and 16 units, respectively, determined empirically to capture temporal dependencies effectively without overcomplexity. Outputs from both branches are concatenated and passed to a fully connected layer with 64 ReLU neurons, followed by a softmax output layer performing four-class classification. The model is trained using the Adam optimizer and sparse categorical cross-entropy loss.

2) Model Evaluation: The model's performance was evaluated in a subject-dependent setting using fivefold cross-validation, with each fold split into 80% training and 20% test data using the K-fold method with shuffling and a fixed random seed for reproducibility and class balance. A further 80/20 split within the training set created a validation set to monitor performance. To prevent data leakage and maintain evaluation integrity, we implemented leakage detection by checking for overlapping sample indexes between the training, validation, and test subsets in each fold, as outlined in the supplementary material Algorithm S1. Two callbacks, ReduceLROnPlateau (for adjusting the learning rate) and Early Stopping (to halt training after no improvement), were applied to stabilize training and prevent overfitting. In addition to the subject-dependent evaluation, a subject-independent protocol was implemented using leave-one-subject-out (LOSO) cross-validation [34]. In this setting, data from one subject were entirely held out for testing, while the model was trained on data from the remaining subjects. This process was repeated for all subjects, and the final performance was reported as the average across folds. Unlike the random K-fold scheme, LOSO prevents any subject overlap between training and test sets and therefore provides a stricter assessment of cross-subject generalization. Performance was measured using classification accuracy, confusion matrices, receiver operating characteristic (ROC) curves, and area under the curve (AUC) values. A one-vs-rest strategy was used for multiclass ROC analysis [35]. The confusion matrix and accuracy were computed as follows:

$$M_{i,j} = \sum_{k=1}^{N} \delta(y_k = i) \cdot \delta(\hat{y}_k = j) \tag{6}$$

$$\text{Accuracy} = \frac{\sum_{i=1}^{C} M_{i,i}}{\sum_{i=1}^{C} \sum_{j=1}^{C} M_{i,j}} \times 100 \tag{7}$$

where y_k and $\hat{y}_k$ represent the actual and predicted class labels for the kth sample, respectively. The function $\delta(.)$ refers to the Kronecker delta, C denotes the number of classification categories, and N is the total number of evaluated instances. ROC curves were generated by computing the true positive rate (TPR) and false positive rate (FPR) at varying thresholds as in (8), and the AUC was estimated using trapezoidal rule (9)

$$\text{TPR} = \frac{\text{TP}}{\text{TP+FN}}, \qquad \text{FPR} = \frac{\text{FP}}{\text{FP+TN}} \tag{8}$$

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR}). \tag{9}$$

To measure model complexity and robustness, three key metrics were calculated: 1) floating point operations (FLOPs), derived analytically from the model summary using standard formulas for Conv1D, LSTM, and Dense layers; 2) Inference Latency, averaged over fivefold cross-validation and measured as the mean forward-pass time per trial; and 3) the Parameter Count, extracted from the Keras. summary() output, indicating the total number of trainable parameters. These metrics offer a comprehensive view of the model's computational efficiency and inference performance.

E. Model Interpretability and Explainability

To enhance interpretability, Grad-CAM [36] was used on the first Conv1D layer to visualize class-specific spatial attention over CSP features. A saliency map [37] was computed on the LSTM layer (16 units) to identify temporal–spectral contributions most relevant to each class, based on the gradient magnitude of the class score with respect to each unit's activation. A feature contribution analysis on the Dense layer (64 units) quantified the relative impact of spatial and temporal–spectral features on the final decision. To assess the faithfulness of these explanations, a parameter randomization test [38] was conducted for the CNN, LSTM, and Dense layers. Layer weights

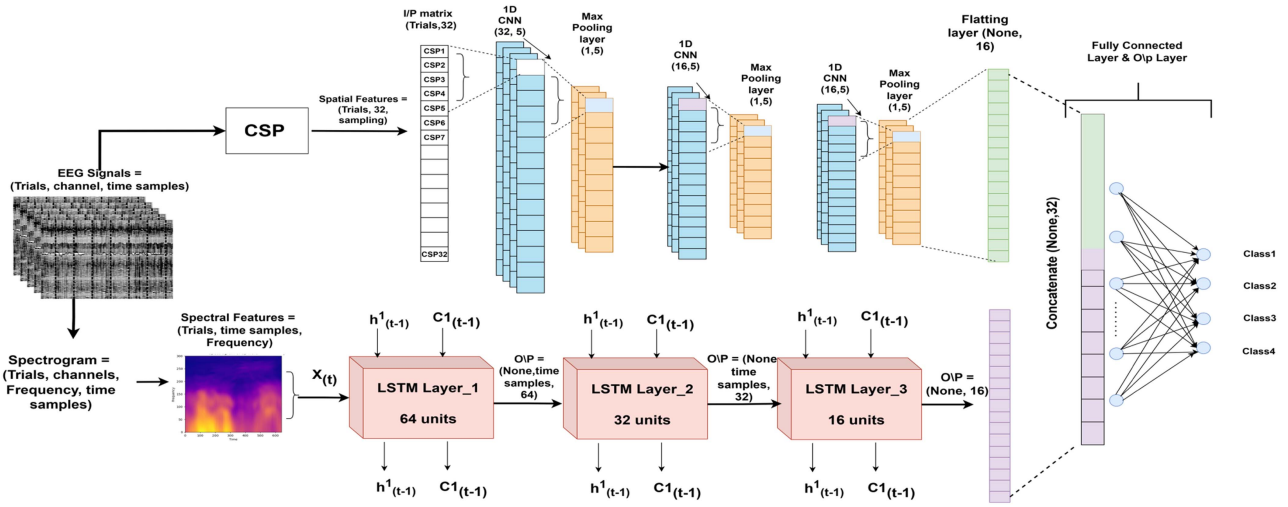

Fig. 3. Architectural diagram of the fusion model, which integrates two branches: Spatial CNNs and Spectro-LSTM, for four-class classification.

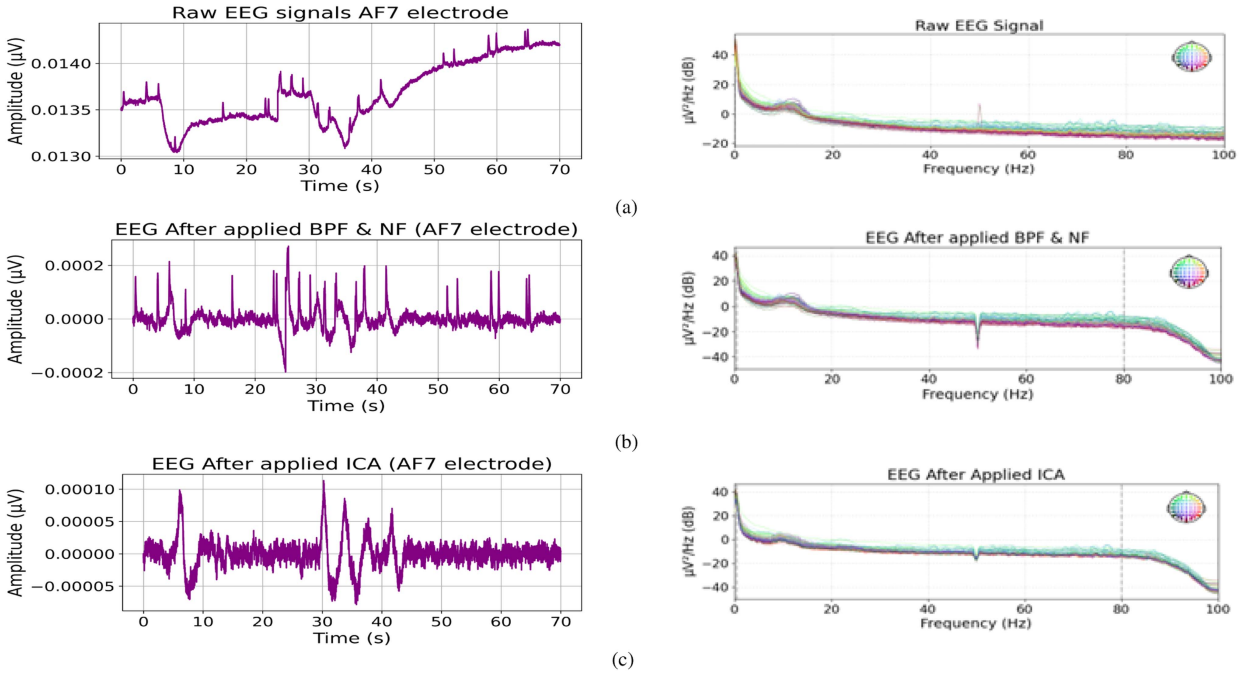

Fig. 4. EEG Preprocessing pipeline and corresponding power spectra. Left panels: EEG signal amplitude (AF7 electrode) in μV. Right panels: corresponding PSD in $\mu V^2/Hz$ at each stage. (a) Raw EEG signal. (b) EEG after band-pass and notch filtering, showing removal of low-frequency drifts and 50-Hz line noise. (c) EEG after ICA, with ocular and muscle artifacts removed.

were randomly permuted while preserving their shape, and the saliency maps from the trained $\tilde{S}^{(L,O)}$ and randomized $\tilde{S}^{(L,r)}$ models were compared using the Pearson correlation coefficient

$$p^{(L)} = \frac{\sum_{i=1}^{N}\left(s_i^{(L,O)} - \bar{s}^{(L,O)}\right)\left(s_i^{(L,r)} - \bar{s}^{(L,r)}\right)}{\sqrt{\sum_{i=1}^{N}\left(s_i^{(L,O)} - \bar{s}^{(L,O)}\right)^2}\sqrt{\sum_{i=1}^{N}\left(s_i^{(L,r)} - \bar{s}^{(L,r)}\right)^2}}. \tag{10}$$

A low correlation ($p^{(L)} \approx 0$) indicates that the saliency maps differ substantially after randomization, confirming that the explanations are sensitive to learned parameters and therefore faithful. Conversely, a high correlation ($p^{(L)} \approx 1$) suggests insensitivity to model learning and potentially unreliable explanations. This multilevel interpretability framework provided consistent insights into how spatial, temporal, and integrated features contribute to the model's decision-making process.

III. Results and Discussion

A. Preprocessing Results

Fig. 4 illustrates the EEG preprocessing pipeline. Band pass and notch filtering effectively attenuated low and high frequencies drifts and suppressed 50-Hz power-line interference. ICA successfully isolated and removed ocular and muscle artifacts.

TABLE II
STATISTICAL COMPARISON OF CSP=32 VS. OTHER CSP CONFIGURATIONS

Comparison	Mean Accuracy (CSP=32) ± std and Mean Accuracy (comparison) ± std	Mean Diff (%) ±95% CI	t(9)	p (one-tailed)	Cohen's dz	Significance
CSP_{32} vs CSP_{16}	91.02 ± 9.92 and 51.75 ± 15.02	39.28 ± 6.23	14.259	$8.749 \times 10^{-8****}$	4.509	Yes
CSP_{32} vs CSP_{24}	91.02 ± 9.92 and 51.36 ± 11.74	39.67 ± 4.56	19.665	$5.268 \times 10^{-9****}$	6.219	Yes
CSP_{32} vs CSP_{40}	91.02 ± 9.92 and 75.82 ± 9.18	15.21 ± 4.72	7.288	$2.311 \times 10^{-5***}$	2.305	Yes
CSP_{32} vs CSP_{60}	91.02 ± 9.92 and 80.05 ± 7.54	10.97± 3.78	6.566	$5.164 \times 10^{-5***}$	2.076	Yes

Mean Accuracy: Classification accuracy (%) averaged across 10 participants. Std: Standard deviation. Mean Diff (%): Average difference between CSP=32 and the comparison method. ± 95% CI: Margin of error (two-tailed) around the mean difference. t(9): t-statistic with 9 degrees of freedom (from paired t-test). p (one-tailed): Statistical significance of whether CSP=32 is better. Asterisks indicate significance level based on one-tailed paired t-test: $*p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001$. Cohen's dz: Standardized effect size for paired samples (≥ 0.8 = large effect). Yes indicates statistically significant improvement ($p < 0.05$). All results show CSP=32 significantly outperforms the compared configuration.

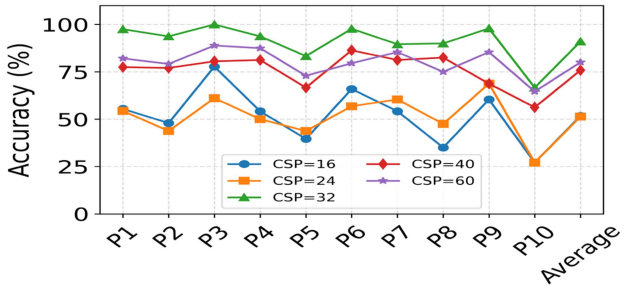

Fig. 5. Classification accuracy for each participant and the overall average using different CSP component configurations (16, 24, 32, 40, and 60).

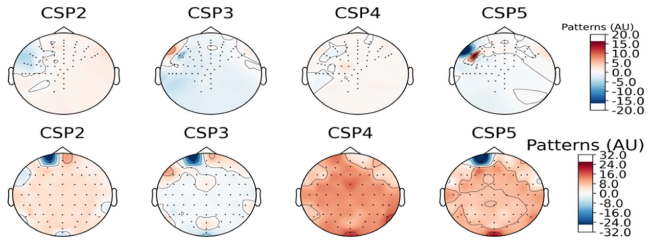

Fig. 6. Topographic distributions of CSP components 2, 3, 4, and 5 from a set of 32 selected CSP components, top row: Data_A; bottom row: Data_B.

The resulting clean EEG signals exhibit stable amplitudes and clearly defined spectral peaks in frequency bands, demonstrating the effectiveness of the preprocessing steps. In Data_A, 60 electrodes were selected from the upper left brain region, focusing on Broca's area, which showed elevated activity during inner speech tasks (Fig. 1). Activity peaked from 0.25 to 1.2 s and then decreased. In contrast, Data_B used all 64 electrodes, as its activation patterns were more diffuse, with only a small region on the upper left showing consistent high activity from one or two electrodes (Fig. 1).

B. CSP Feature Extraction Results

After preprocessing and electrode selection, several CSP component configurations (16, 24, 32, 40, and 64) were evaluated to determine the optimal number of spatial filters. This evaluation was conducted by training the classification model using 80% of the data for training and 20% for testing. The optimization analysis of the CSP component, tested with these configurations, revealed that the 32-component configuration consistently outperformed the others, yielding superior results (Fig. 5). To confirm significance, we performed paired t-tests comparing CSP = 32 with the other configurations and observed statistically significant improvements in all cases, with mean accuracy improvements ranging from 10.97%–39.67%. The large effect sizes (Cohen's dz = 2.076 to 6.219) and significant statistical differences ($p < 0.001$) demonstrate that CSP = 32 captured the most discriminative features for inner speech decoding. The detailed statistical results are presented in Table II, and the corresponding CSP spatial patterns with topographic maps are shown in Fig. 6.

C. Classification Performance Results

The proposed dual-branch CNN–LSTM fusion model demonstrated consistently strong classification performance across both EEG datasets in a subject-dependent setting. Fig. 7 displays the training process results for participant 9 of Data_A (a) and participant 1 of Data_B (b). The model showed effective training, as evidenced by the increase in both training and validation accuracy and the reduction in training and validation loss. For Data_A, the model achieved an average accuracy of 89.99 ± 2.95% with a confidence interval and a mean F1-score of 89.18 ± 2.66%, as summarized in Table III. Participant-level accuracies ranged from 68.75% (P10) to 97.08% (P9), with an average trial latency of 339.05 ± 17.04 ms. For Data_B, the model performed slightly better, with an average accuracy of 92.47 ± 2.25% and a mean F1-score of 0.9297 ± 0.0228, as detailed in Table IV. Participant accuracies ranged from 84.38%–100%, with an average latency of 416.30 ± 18.94 ms. Across both datasets, the model maintained high per-class performance with low variability across folds, demonstrating robustness and reliability for EEG classification tasks.

It is worth noting that LOSO represents a limited form of subject-independent evaluation, whereas fully patient-independent evaluation can be formulated as a leave-n-subjects-out or subject-level K-fold setting, which is more challenging due to increased intersubject variability. To further assess cross-subject generalization, an LOSO protocol was conducted, where one participant was held out for testing while training was performed on the remaining subjects. For Data_A, LOSO yielded an average accuracy of 26.20 ± 7.85% with a mean F1-score of 25.24 ± 8.91% (95% CI). For Data_B, the average accuracy was 20.47 ± 6.8% with a mean F1-score of 18.64 ± 21.46% (95% CI) (Table V and Table S1 in the Supplementary Material). Considering the four-class setting (chance level =

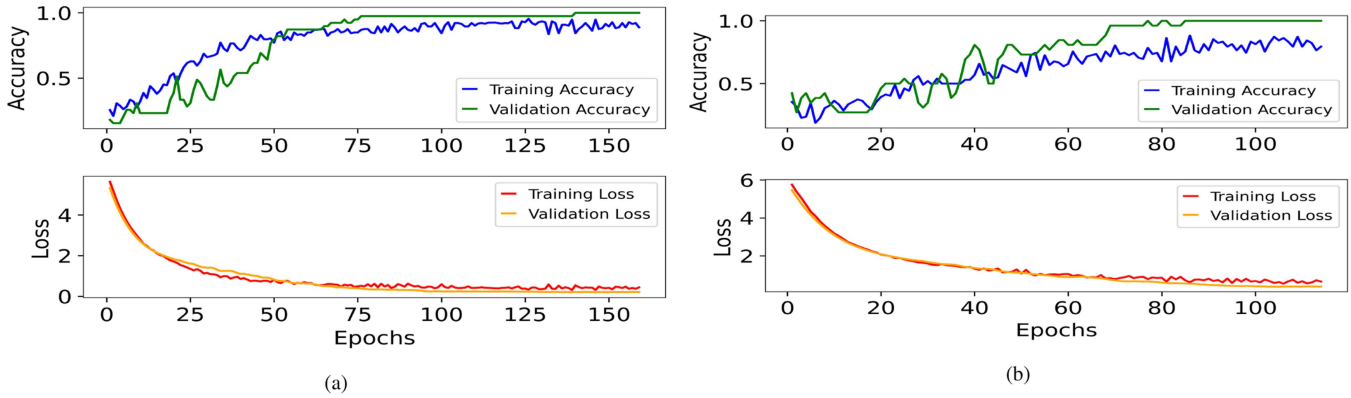

Fig. 7. Epochwise training and validation accuracy and loss for (a) P9 in Data_A and (b) P1 in Data_B datasets.

TABLE III
DATA_A: CLASSIFICATION PERFORMANCE OF THE PROPOSED MODEL ACROSS PARTICIPANTS USING 5-FOLD CROSS-VALIDATION

Part	No. All data	No. test data	Accuracy (%) per fold					Mean Accuracy (%) with 95% CI	Mean F1-Score with 95% CI across folds Per-Class				Latency (ms/trial) ± Std (ms)
			Fold 1	Fold 2	Fold 3	Fold 4	Fold 5		Class0	Class1	Class2	Class3	
P1	200	40	95.0	92.50	90.0	95.0	97.50	94.00 ± 2.55	93.72 ± 5.60	92.91 ± 5.41	93.38 ± 6.88	94.72 ± 4.35	363.22 ± 7.56
P2	240	48	89.58	83.33	89.58	87.50	87.50	87.50 ± 2.28	87.40 ± 6.51	79.59 ± 9.72	90.78 ± 8.80	90.13 ± 8.72	331.69 ± 14.07
P3	180	36	94.44	83.33	94.44	86.11	91.67	90.00 ± 4.51	87.76 ± 8.48	89.17 ± 13.63	94.23 ± 6.57	89.07 ± 10.25	368.59 ± 14.33
P4	240	48	91.67	95.83	95.83	95.83	93.75	94.58 ± 1.67	91.81 ± 5.06	95.15 ± 4.42	93.91 ± 5.40	97.11 ± 3.24	351.72 ± 9.87
P5	240	48	85.42	83.33	83.33	81.25	91.67	85.00 ± 3.58	84.52 ± 12.40	86.64 ± 4.42	88.43 ± 8.50	80.34 ± 9.92	334.71 ± 4.46
P6	216	43	95.45	95.34	97.67	97.67	97.67	96.77 ± 1.11	95.08 ± 3.99	93.21 ± 3.57	100.0 ± 50.00	98.25 ± 2.38	350.58 ± 13.66
P7	240	48	89.58	87.50	89.58	95.83	91.67	90.83 ± 2.83	94.59 ± 5.65	93.02 ± 5.17	89.30 ± 6.02	86.14 ± 7.67	316.75 ± 20.87
P8	200	40	82.50	85.0	97.5	95.0	87.50	89.50 ± 5.79	95.80 ± 6.58	87.75 ± 2.75	86.05 ± 13.39	87.56 ± 11.50	333.82 ± 12.00
P9	240	48	97.92	100.0	97.92	97.92	91.67	97.08 ± 2.83	97.56 ± 2.63	97.38 ± 3.67	98.30 ± 2.36	95.52 ± 5.44	311.93 ± 9.92
P10	240	48	64.58	72.92	70.83	68.75	66.67	68.75 ± 2.95	70.05 ± 14.98	66.97 ± 16.06	63.15 ± 8.03	72.21 ± 11.64	328.47 ± 17.78
Avg								89.99 ± 2.95	89.18 ± 2.66	89.30 ± 2.85	89.47 ± 3.46	88.84 ± 3.24	339.05 ± 17.04

Part: Participants; Avg: Average; mean F1-Scores (±95% CI): denote the average F1-score per class, computed across the five folds, and CI is Confidence Interval. Latency (ms/trial ± Std) represents the mean processing time per trial averaged over 5 folds, with the accompanying standard deviation (Std) indicating temporal variability.

TABLE IV
DATA_B: CLASSIFICATION PERFORMANCE OF THE PROPOSED MODEL ACROSS PARTICIPANTS USING 5-FOLD CROSS-VALIDATION

Part	No. All data	No. test data	Accuracy (%) per fold					Mean Accuracy (%) with 95% CI	Mean F1-Score with 95% CI across folds Per-Class				Latency (ms/trial) ± Std (ms)
			Fold 1	Fold 2	Fold 3	Fold 4	Fold 5		Class0	Class1	Class2	Class3	
P1	160	32	93.75	96.88	90.63	100.0	87.50	93.75 ± 4.42	95.57 ± 5.14	94.63 ± 6.02	90.40 ± 10.96	92.92 ± 7.67	439.48 ± 31.98
P2	160	32	87.50	87.50	87.50	84.38	87.50	86.88 ± 1.25	82.37 ± 4.66	87.37 ± 6.62	89.27 ± 11.23	87.35 ± 8.47	417.06 ± 34.80
P3	160	32	100.0	93.75	96.88	96.88	100	97.50 ± 2.34	98.46 ± 2.90	96.32 ± 5.28	97.89 ± 2.84	97.71 ± 3.13	417.52 ± 16.99
P5	160	32	93.75	96.88	93.75	93.75	90.63	93.75 ± 1.98	94.61 ± 6.04	93.54 ± 2.01	93.31 ± 4.91	94.76 ± 6.28	393.12 ± 11.01
Avg								92.47 ± 2.25	92.97 ± 2.28	92.75 ± 3.35	92.96 ± 1.80	92.72 ± 1.87	416.30 ± 18.94

Part: Participants; Avg: Average; mean F1-Scores (±95% CI): denote the average F1-score per class, computed across the five folds, and CI is Confidence Interval. Latency (ms/trial ± Std) represents the mean processing time per trial averaged over 5 folds, with the accompanying standard deviation (Std) indicating temporal variability.

TABLE V
SUMMARY OF LOSO (SUBJECT-INDEPENDENT) MEAN ACCURACY

Dataset	Mean Accuracy (% ± Standard Deviation)
Data_A	26.20 ± 7.85
Data_B	20.47 ± 6.81

Detailed per-class F1-scores and participant-level results are provided in the supplementary material (Table S1).

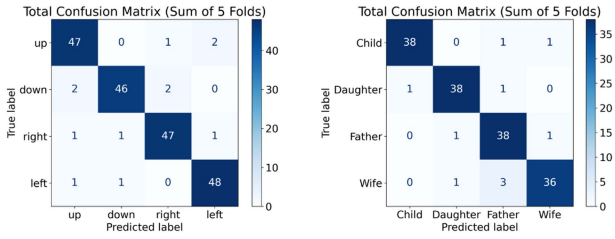

Fig. 8. Confusion matrices for inner speech classification of P1 in (a) Data_A and (b) Data_B using fivefold cross-validation.

25%), performance is close to chance, indicating substantial intersubject variability. This degradation relative to subject-dependent results suggests that the proposed method benefits from subject-specific calibration, and that more general patient-independent settings would likely yield lower performance.

Confusion matrices were computed for each fold and cumulatively across all folds for both datasets. These visualizations are presented in Fig. 8 for P1 in Data_A, and Data_B. The average microaveraged AUC across folds further supported these results, as reported in the Supplementary Material (Table S2 and S3). Most participants achieved high per-class AUC values. In

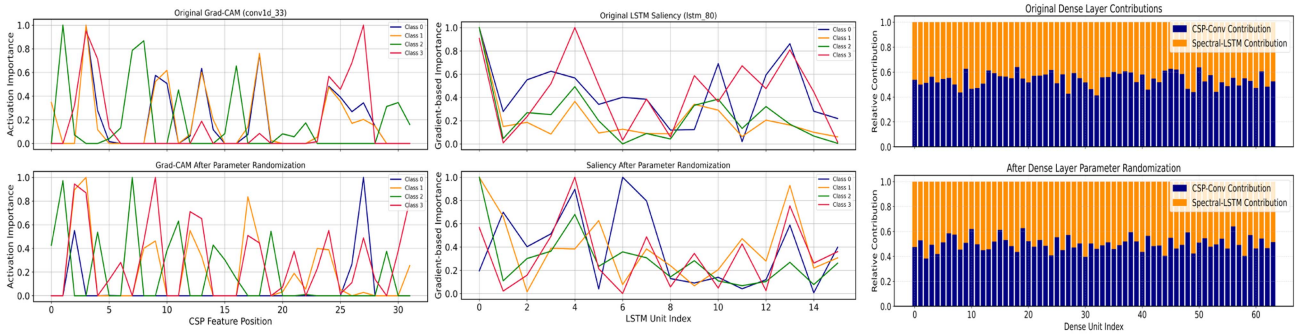

Fig. 9. Interpretability analyses of the model. (a) Grad-CAM activation trends at the conv layer. (b) Gradient-based saliency mapping at the LSTM layer. (c) Relative contributions of CSP-Conv and spectral-LSTM features to dense layer before and after parameter randomization.

TABLE VI
ANOVA AND F-TEST RESULTS FOR ACCURACY AND VARIANCE COMPARISON BETWEEN DATASETS

Test	ANOVA: Single Factor		F-Test Two-Sample (Data_A and Data_B)
	Data_A	Data_B	
F-Statistic (F)	0.1594	0.1714	3.4696
P-Value	0.9577	0.94964	0.1671
F Critical	2.5787	3.0555	8.8123

F: a ratio of variances used to compare groups; p-value: The probability of observing the test statistic by chance; values above 0.05 suggest no significant difference; F-Critical Value: the threshold for rejecting the null hypothesis of equal variances. If the F-statistic is lower, there is no significant difference in variances.

TABLE VII
QUANTITATIVE COMPARISON BETWEEN ORIGINAL AND RANDOMIZED EXPLANATIONS FOR SELECTED LAYERS

Layer	Correlation's values							
	Data_A / Classes (C)				Data_B / Classes (C)			
	C0	C1	C2	C3	C0	C1	C2	C3
CNN_1d (32) Grad-CAM	0.37	0.12	0.47	0.27	0.31	-0.12	0.21	0.48
LSTM (16 units) Saliency Maps	0.09	0.06	-0.11	0.17	-0.14	-0.09	0.21	-0.03
Dense (64 units) Saliency Maps	0.03	-0.25	-0.03	0.15	-0.08	-0.11	-0.09	0.04
Dense Contributions	CSP correlation = -0.091 LSTM correlation = -0.091				CSP correlation = -0.05 LSTM correlation = -0.05			

Data_A, P1 and P6 achieved near-perfect AUCs (P1: 99.76 ± 0.32%, P6: 100 ± 0.00%), while P2 performed lower (97.24 ± 2.93%, C1: 87.04–98.59%). P9 also performed excellently, whereas P10 showed variability, especially for C0 (86.37 ± 3.69%). In Data_B, P1 again reached high accuracy (99.27 ± 1.38%, C0: 96.35–100%), and P3 consistently excelled (C0: 97.50 ± 2.34% to C2: 99.90 ± 0.18%). P5 had weaker results for C0 (93.75–95.65%, average 93.75 ± 1.98%) but maintained strong performance on other classes (99.43–99.90%).

The model showed consistent, high performance with narrow confidence intervals, highlighting its robustness. Some participant variability, particularly for P2 and P10, suggests that subject-specific tuning may be needed, reflecting challenges, such as EEG signal quality or individual differences. Per-class ROC analyses demonstrated excellent discriminative capability, with AUC values consistently exceeding 0.95 for most word classes. The ROC curves for each fold, alongside the averaged ROC plots, are presented in Supplementary Material (Fig. S1). The model showed strong accuracy, stability, and low latency, though variability across participants, to assess whether there are significant differences between the datasets, I applied two statistical tests: 1) ANOVA: Single Factor to evaluate participant-level differences in accuracy within each dataset; and 2) the F-Test two-sample for variances to compare the variances between the two datasets. The results were summarized in Table VI. The ANOVA: Single factor test reveals no significant differences in accuracy within either Data_A ($p = 0.9577$) or Data_B ($p = 0.9496$), as both p-values are much higher than the 0.05 threshold. The F-test two-sample shows an F-statistic of 3.4696 with a p-value of 0.1671, which is also greater than 0.05. Since the F-statistic is lower than the critical value of 8.8123, we conclude that there is no significant difference in variances between Data_A and Data_B.

D. Model Interpretability Results

To improve interpretability, three analyses were performed on the CNN, LSTM, and Dense layers before and after parameter randomization. Grad-CAM activations at the convolutional layer reveal strong, class-specific spatial patterns before randomization, which diminish afterwards, indicating meaningful learned features [Fig. 9(a)]. LSTM saliency analysis revealed that before randomization, unit 4 played a crucial role across all classes, while after randomization, unit 0 became more dominant, emphasizing the model's ability to capture class-specific temporal dependencies [Fig. 9(b)]. The relative contributions of CSP-Conv and spectral-LSTM features to the dense units show balanced, complementary influences before randomization, with unit 9 being influenced by CSP features and unit 32 by spectral inputs. After randomization, contributions became less distinct and more uniform, confirming that the original contributions reflect genuine, task-relevant information used in the model's final predictions [Fig. 9(c)].

Table VII presents the Pearson correlation coefficients between the original and randomized explanation maps for key layers. The results show that CNN_1 d (Grad-CAM) exhibits moderate correlations (ranging from 0.12 to 0.48), indicating sensitivity to learned parameters, with LSTM saliency maps demonstrating the strongest sensitivity to weight randomization and higher faithfulness, especially for both datasets (correlations from −0.14 to 0.21). Dense network saliency maps show weak and often negative correlations (ranging from −0.25 to

TABLE VIII
COMPARISON WITH PREVIOUS INNER SPEECH EEG CLASSIFICATION STUDIES USING THE SAME DATASET

Paper	Dataset	Channels*	Method	Parameters (M)	FLOPs (MFLOPs)	Latency (ms/trial)	Accuracy
[2]	Data_A	Left hemisphere	PSD Features + SVM	-	-	-	26.2%
		128 channels	Most Important Features + LSTM	-	-	-	30.0%
			Raw Data + BiLSTM	-	-	-	36.1%
[39]	Data_A	128 channels	Random Forest classification model	-	-	-	29.7%
			SVM	-	-	-	35.3%
			K-Nearest Neighbor	-	-	-	31.4%
[40]	Data_A	26 left hemisphere	Raw Data + EEGNet	-	-	-	29.7%
[13]	Data_B	64 channels	Baseline EEG + RF Classifier	-	-	-	22.1%
[41]	Data_A	128 channels	Inter-trial Coherence + KNN	-	-	-	58.7%
			Inter-trial Coherence + SVM	-	-	-	67.5%
This study	Data_A	60 channels	Spatio-spectral Features + Fusion Model	115,556 (0.12M)	69.81 MFLOPs	339.85 ± 18.94	89.99%
	Data_B	64 channels	Spatio-spectral Features + Fusion Model	115,556 (0.12M)	111.66 MFLOPs	416.30 ± 18.94	92.47%

∗These are the channels used after dropout: Data_A (original 128 channels) and Data_B (64 channels).

0.15), suggesting that these saliency maps are highly sensitive to learned parameters, reflecting faithfulness. The results show low correlations in all layers, CNN Grad-CAM, LSTM, and Dense saliency maps. These findings highlight the model's alignment with neurophysiological expectations, reinforcing its ability to accurately classify inner speech using both spatial and temporal-spectral features.

Comparing this study to previous work using same dataset, as shown in Table VIII. For Data_A, the model achieved 89.99% accuracy, significantly improving over the previous best result of 67.5% [41], representing a 33.32% improvement. For Data_B, the model achieved 92.47%, an increase of 318.41% compared to the previous best of 22.1% [13]. These improvements are attributed to the integration of optimized spatio-spectral features, region-specific electrode analysis, and deep learning techniques. The model also demonstrated computational efficiency, with 69.81 MFLOPs for Data_A and 111.66 MFLOPs for Data_B, and low latency, making it suitable for real-time applications. These results set a new benchmark in EEG classification, with higher accuracy and efficiency than previous methods (Table VIII). Future work should focus on enhancing artifact resistance and exploring real-time deployment.

IV. CONCLUSION

This study presented a dual-branch deep learning framework that combines spatial features extracted via CSP with spectrogram-based temporal features for inner speech EEG classification. The model, utilizing parallel CNN and LSTM pathways, achieved impressive classification performance in a subject-dependent setting, with accuracies of 89.99% and 92.47% on two distinct EEG datasets. The results highlighted the effectiveness of optimizing CSP components, applying task-specific electrode selection, especially over Broca's area—and combining spatial and spectral representations to enhance decoding performance. Interpretability analyses provided insights into the model's decision-making process, demonstrating that the model relies on physiologically meaningful EEG patterns. These analyses confirmed the transparency of the model's internal workings, aligning its predictions with expected neural activity associated with inner speech. Unlike previous studies that analyzed Data_A or Data_B in isolation, this work integrated both datasets into a unified framework, offering a more robust and comprehensive approach to feature extraction and classification.

Despite these advancements, several limitations must be acknowledged. One primary limitation is the relatively small dataset size, which limits the generalizability of the findings and underscores the need for a more extensive dataset to validate the model's scalability across diverse populations. In addition, a LOSO subject-independent evaluation was conducted to assess cross-subject generalization. The LOSO results yielded average accuracies of 26.20% for Data_A and 20.47% for Data_B, representing a substantial reduction compared to the subject-dependent setting and approaching chance level in the four-class scenario (25%). This degradation reflects pronounced intersubject variability in EEG spatial patterns and the subject-specific characteristics of CSP-based representations. These findings indicate that the proposed framework benefits from user calibration and is currently better suited for subject-centered deployment scenarios.

Future work will focus on expanding the dataset to include a larger and more diverse participant pool to improve generalizability. Strategies, such as domain adaptation, transfer learning, and advanced normalization techniques will be explored to enhance subject-independent robustness and mitigate intersubject variability. Additional comparisons with recent state-of-the-art models will be performed to evaluate relative performance. Efforts will also be directed toward improving artifact resistance and enabling real-time deployment of the framework. These directions aim to increase the scalability, reliability, and practical applicability of EEG-based inner speech decoding systems in real-world scenarios.

REFERENCES

[1] P. S. Pooja, S. Pahuja, and K. Veer, "Recent approaches on classification and feature extraction of EEG signal: A review," *Robotica*, vol. 40, no. 1, pp. 77–101, Jan. 2022, doi: 10.1017/S0263574721000382.

[2] F. Gasparini, E. Cazzaniga, and A. Saibene, "Inner speech recognition through electroencephalographic signals," in *Proc. Italian Workshop Artif. Intell. Hum. Mach. Interac.*, Udine, Italy, Dec. 2022, pp. 1–14.

[3] V. N. Kiroy, O. M. Bakhtin, E. M. Krivko, D. M. Lazurenko, E. V. Aslanyan, and D. G. Shaposhnikov, "Spoken and inner speech-related EEG connectivity in different spatial directions," *Biomed. Signal Process. Control*, vol. 71, Jan. 2022, Art. no. 103224, doi: 10.1016/j.bspc.2021.103224.

[4] T. Schultz, M. Wand, T. Hueber, D. J. Krusienski, C. Herff, and J. S. Brumberg, "Biosignal-based spoken communication: A survey," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 12, pp. 2257–2271, Dec. 2017, doi: 10.1109/TASLP.2017.2752365.

[5] R. Wise, F. Chollet, U. Hadar, K. Friston, E. Hoffner, and R. Frackowiak, "Distribution of cortical neural networks involved in word comprehension and word retrieval," *Brain*, vol. 114, no. 4, pp. 1803–1817, 1991. Available: https://academic.oup.com/brain/article/114/4/1803/387296

[6] H. Ramoser, J. Müller-Gerking, and G. Pfurtscheller, "Optimal spatial filtering of single trial EEG during imagined hand movement," *IEEE Trans. Rehabil. Eng.*, vol. 8, no. 4, pp. 441–446, Dec. 2000, doi: 10.1109/86.895946.

[7] S. H. Lee, M. Lee, and S. W. Lee, "Neural decoding of imagined speech and visual imagery as intuitive paradigms for BCI communication," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 12, pp. 2647–2659, Dec. 2020, doi: 10.1109/TNSRE.2020.3040289.

[8] N. C. Mahapatra and P. Bhuyan, "EEG-based classification of imagined digits using a recurrent neural network," *J. Neural Eng.*, vol. 20, no. 2, Apr. 2023, Art. no. 026040, doi: 10.1088/1741-2552/acc976.

[9] M. Ramzan and S. Dawn, "Fused CNN-LSTM deep learning emotion recognition model using electroencephalography signals," *Int. J. Neurosci.*, vol. 133, no. 6, pp. 587–597, Jun. 2023, doi: 10.1080/00207454.2021.1941947.

[10] S. Paszkiel, *Analysis and Classification of EEG Signals for Brain–Computer Interfaces (Studies in Computational Intelligence)*, vol. 852, Cham, Switzerland: Springer, 2020, doi: 10.1007/978-3-030-30581-9.

[11] M. Ji, C. Huang, and R. Luo, "Automatic modulation recognition based on spatio-temporal features fusion," in *Proc. IEEE 12th Int. Conf. Electron. Inf. Emerg. Commun.*, Jul. 2022, pp. 89–93, doi: 10.1109/ICEIEC54567.2022.9835042.

[12] M. M. Abdulghani, W. L. Walters, and A. K. H. Abed, "Inner speech classification using EEG and deep learning," *Bioengineering*, vol. 10, no. 6, 2023, Art. no. 649, doi: 10.3390/bioengineering10060649.

[13] H. Wilson et al., "Performance of data-driven inner speech decoding with same-task EEG-fMRI data fusion and bimodal models," 2023, *arXiv:2306.10854*.

[14] N. C. Mahapatra and P. Bhuyan, "Multiclass classification of imagined speech vowels and words of electroencephalography signals using deep learning," *Adv. Hum.-Comput. Interact.*, vol. 2022, 2022, Art. no. 1374880, doi: 10.1155/2022/1374880.

[15] U. Goenka, P. Patil, K. Gosalia, and A. Jagetia, "Classification of electroencephalograms during mathematical calculations using deep learning," in *Proc. IEEE 23rd Int. Conf. Inform. Reuse Integrat. Data Sci.*, San Diego, CA, USA, Aug. 2022, doi: 10.1109/IRI54793.2022.00016.

[16] S. K. Prabhakar, H. Rajaguru, C. Kim, and D. O. Won, "A fusion-based technique with hybrid swarm algorithm and deep learning for biosignal classification," *Front. Hum. Neurosci.*, vol. 16, Jun. 2022 Art. no. 895761, doi: 10.3389/fnhum.2022.895761.

[17] C. A. Ellis, A. Sattiraju, R. L. Miller, and V. D. Calhoun, "A framework for systematically evaluating the representations learned by a deep learning classifier from raw multi-channel electroencephalogram data," 2023, *arXiv:2023.03.20.533467*, doi: 10.1101/2023.03.20.533467.

[18] C. A. Ellis, A. Sattiraju, R. L. Miller, and V. D. Calhoun, "Novel approach explains spatio-spectral interactions in raw electroencephalogram deep learning classifiers," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. Workshops*, Rhodes Island, Greece, Jun. 2023, doi: 10.1109/ICASSPW59220.2023.10193605.

[19] S. Omari, M. Kimwele, A. Olowolayemo, and D. Kaburu, "Enhancing EEG signals classification using LSTM-CNN architecture," *Eng. Rep.*, vol. 6, no. 9, Dec. 2023, doi: 10.1002/eng2.12827.

[20] M. Asif, D. Srivastava, A. Gupta, and U. S. Tiwary, "Inter subject emotion recognition using spatio-temporal features from EEG signal," in *Proc. 27th Int. Comput. Sci. Eng. Conf.*, 2023, pp. 1–4, doi: 10.1109/ICSEC59635.2023.10329776.

[21] F. Gasparini, E. Cazzaniga, and A. Saibene, "Inner speech recognition through electroencephalographic signals," 2022, Accessed: Oct. 30, 2025. [Online]. Available: http://ceur-ws.org

[22] S. Martin, I. Iturrate, J. del R. Millán, R. T. Knight, and B. N. Pasley, "Decoding inner speech using electrocorticography: Progress and challenges toward a speech prosthesis," *Front. Neurosci.*, vol. 12, Jun. 2018, Art. no. 367292.

[23] N. Nieto, V. Peterson, H. L. Rufiner, J. E. Kamienkowski, and R. Spies, "Thinking out loud, an open-access EEG-based BCI dataset for inner speech recognition," *Sci. Data*, vol. 9, no. 1, Dec. 2022, Art. no. 52, doi: 10.1038/s41597-022-01147-2.

[24] F. Simistira Liwicki et al., "Bimodal electroencephalography-functional magnetic resonance imaging dataset for inner-speech recognition," *Sci. Data*, vol. 10, no. 1, Dec. 2023, Art. no. 378, doi: 10.1038/s41597-023-02286-w.

[25] A. Rezazadeh Sereshkeh, R. Trott, A. Bricout, and T. Chau, "EEG classification of covert speech using regularized neural networks," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 12, pp. 2292–2300, Dec. 2017, doi: 10.1109/TASLP.2017.2758164.

[26] N. P. Singh, A. K. Gautam, and T. Sharan, "An insight into the hardware and software aspects of a BCI system with focus on ultra-low power bulk driven OTA and GM-C based filter design, and a detailed review of the recent AI/ML techniques," in *Artif. Intell.-Based Brain-Comput. Interface*, V. Bajaj and G. R. Sinha, Eds., London, U.K.: Academic Press, 2022, ch. 13, pp. 283–315, doi: 10.1016/B978-0-323-91197-9.00015-1.

[27] K. Zhu, X. Zhang, J. Wang, N. Cheng, and J. Xiao, "Improving EEG-based emotion recognition by fusing time-frequency and spatial representations," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Rhodes Island, Greece, Jun. 2023, doi: 10.1109/ICASSP49357.2023.10097171.

[28] X. Jiang, L. Meng, X. Chen, Y. Xu, and D. Wu, "CSP-Net: Common spatial pattern empowered neural networks for EEG-based motor imagery classification," *Knowl.-Based Syst.*, vol. 305, 2024, Art. no. 112668.

[29] M. Grosse-Wentrup and M. Buss, "Multiclass common spatial patterns and information theoretic feature extraction," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 8, pp. 1991–2000, Aug. 2008, doi: 10.1109/TBME.2008.921154.

[30] A. Kamble, P. H. Ghare, V. Kumar, A. Kothari, and A. G. Keskar, "Spectral analysis of EEG signals for automatic imagined speech recognition," *IEEE Trans. Instrum. Meas.*, vol. 72, 2023, Art. no. 4009409, doi: 10.1109/TIM.2023.3300473.

[31] V. Bhandage, T. Pokuri, D. Desai, and A. Jeyabose, "Detection of epilepsy disorder using spectrogram images generated from brain EEG signals," *IEEE Access*, vol. 12, pp. 195054–195064, 2024, doi: 10.1109/ACCESS.2024.3520861.

[32] MNE-Python documentation, "mne.time_frequency.tfr_array_multitaper," MNE-Python Documentation, 2026. Accessed: May 14, 2026. [Online]. Available: https://mne.tools/stable/generated/mne.time_frequency.tfr_array_multitaper.html

[33] A. Akan and O. Karabiber Cura, "Time–frequency signal processing: Today and future," *Digit. Signal Process.*, vol. 119, Dec. 2021, Art. no. 103216, doi: 10.1016/j.dsp.2021.103216.

[34] Y. Wang, T. Li, Y. Yan, W. Song, and X. Zhang, "What causes performance degradation in cross-subject EEG classification?," 2026, *arXiv:2410.03057v2*.

[35] J. Brownlee, "One-vs-rest and one-vs-one for multi-class classification," *Mach. Learn. Mastery*, Apr. 2021. [Online]. Available: https://machinelearningmastery.com/one-vs-rest-and-one-vs-one-for-multi-class-classification/

[36] D.-K. Kim, "Grad-CAM: A gradient-based approach to explainability in deep learning," *Medium*, Oct. 2025. Accessed: May 14, 2026. [Online]. Available: Medium.com/@kdk199604/grad-cam-a-gradient-based-approach-to-explainability-in-deep-learning-871b3ab8a6ce

[37] G. Amrani, A. Adadi, and M. Berrada, "An explainable hybrid DNN model for seizure vs. non-seizure classification and seizure localization using multi-dimensional EEG signals," *Biomed. Signal Process. Control*, vol. 95, Sep. 2024, Art. no. 106322, doi: 10.1016/J.BSPC.2024.106322.

[38] A. Hedström, L. Weber, S. Lapuschkin, and M. Höhne, "Sanity checks revisited: An exploration to repair the model parameter randomisation test," Accessed: Oct. 28, 2025. [Online]. Available: https://github.com/understandable-machine-intelligence-lab/Quantus

[39] N. R. Merola, N. G. Venkataswamy, and M. H. Imtiaz, "Can machine learning algorithms classify inner speech from EEG brain signals," in *Proc. IEEE World AI IoT Congr.*, Jun. 2023, pp. 466–470, doi: 10.1109/AIIoT58121.2023.10174580.

[40] B. van den Berg, S. van Donkelaar, and M. Alimardani, "Inner speech classification using EEG signals: A deep learning approach," in *Proc. 2nd IEEE Int. Conf. Human-Mach. Syst.*, Sep. 2021, pp. 1–4, doi: 10.1109/ICHMS53169.2021.9582457.

[41] D. Lopez-Bernal, D. Balderas, P. Ponce, and A. Molina, "Inner speech classification using inter-trial coherence framework for feature extraction," in *Proc. 19th Int. Symp. Med. Inf. Process. Anal.*, 2023, pp. 1–6, doi: 10.1109/SIPAIM56729.2023.10373449.