Jurnal Kejuruteraan 38(7) 2026: 3543-3554
https://doi.org/10.17576/jkukm-2026-38(7)-30

SE-MDTCN: A Spatially Enhanced Temporal Convolutional Network for Imagined Speech Classification from Electroencephalography Signals

Fujie Qiu[a], Tian Swee Tan[a*] & Sadeem Nabeel Saleem Kbah[a,b]

[a]*Department of Biomedical Engineering & Health Sciences, Faculty of Electrical Engineering,*

Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia,

[b]*Department of Biomedical Engineering, Al-Khwarizmi College of Engineering,*

University of Baghdad, 10071 Baghdad, Iraq

[*]*Corresponding author: tantswee@utm.my*

Received 29 April 2026, Received in revised form 9 July 2026
Accepted 9 August 2026, Available online 30 October 2026

ABSTRACT

Electroencephalography (EEG)-based imagined speech decoding is a promising approach for non-invasive neural speech prostheses and related brain-computer interface (BCI) applications. However, this task remains challenging because EEG signals have a low signal-to-noise ratio (SNR), limited spatial resolution, and significant individual variability. To address these challenges, this study proposes a Spatially Enhanced Multi-branch Depthwise Temporal Convolutional Network (SE-MDTCN), an end-to-end spatiotemporal network designed to enhance spatial representation learning and temporal feature extraction for imagined speech decoding. The proposed model integrates electrode normalization, a modified squeeze-and-excitation (SE)-based spatial filter, a multi-branch depthwise temporal convolutional network (MDTCN), and a multilayer perceptron (MLP) classifier. SE-MDTCN was evaluated on the 2020 International BCI Competition Track 3 dataset under a participant-dependent 5-class setting, using both raw and independent component analysis (ICA)-cleaned EEG. The model achieved average 5-class accuracies of 74.4% on raw EEG and 69.1% on ICA-cleaned EEG, outperforming several representative deep-learning baselines for EEG-based BCI decoding. Ablation studies and feature analyses further demonstrated the consistent contributions of electrode normalization and spatial attention, while temporal features extracted with smaller kernels were found to be more robust than those extracted with larger kernels. These findings suggest that SE-MDTCN improves imagined-speech EEG decoding by jointly modelling spatial structure and temporal dynamics, while providing interpretable spatiotemporal representations. This approach may contribute to the development of practical non-invasive speech BCIs. The code is publicly available to facilitate reproducibility.

Keywords: Brain-computer interface (BCI); imagined speech, electrode normalization; squeeze-and-excitation (SE); multi-branch depthwise temporal convolutional network (MDTCN)

INTRODUCTION

BCIs acquire brain signals directly to generate control commands, bypassing peripheral nerves and muscles to operate external devices. Scalp EEG offers high temporal resolution and is non-invasive, making brain-signal decoding both safe and cost-effective. The predominant paradigms in current EEG-BCI studies include Steady-State Visual Evoked Potential (SSVEP), P300, and Motor Imagery (MI) (Edelman et al. 2025). BCIs based on the SSVEP and P300 paradigms, as exogenous BCIs, require external stimuli to evoke responses that drive the execution system. The requirement for external stimuli may result in visual or auditory fatigue during experiments. In contrast, the MI paradigm, as an endogenous BCI, does not rely on any external stimuli. However, because it relies on motor imagery of limb movements, the number of available commands is limited.

Human speech encodes and transmits information at approximately 4-5 syllables per second (Coupé et al. 2019). Because speech production is naturally controllable, EEG-

based decoding of overt or imagined speech has attracted increasing attention. Although "imagined speech" is used broadly in the literature, we define it here as speech MI without activation of articulatory muscles (Schultz et al. 2017). Speech decoding based on electrocorticography (ECoG) has been validated in multiple clinical studies. Luo et al. (2023) achieved 90.5% accuracy for six-word decoding over three months in a patient with a speech disorder, and Metzger et al. (2023) decoded a large vocabulary at 78 words/min using a 253-electrode high-density ECoG array. However, scalp EEG-based imagined-speech decoding remains challenging due to lower spatial resolution, fewer channels, and artifact-induced low SNR.

To enhance the SNR, filtering techniques have been applied in the temporal, spectral, and spatial domains to identify the most discriminative EEG components for classification. In the temporal domain, Qiu et al. (2025) demonstrated that the short-term preparation phase preceding the imagined-speech period contains speech-related information. In the spectral domain, multiple studies have demonstrated that the gamma band (> 30 Hz) plays an important role in decoding imagined-speech commands (Lee S.-H. et al. 2020; Qiu et al. 2025; Shah et al. 2024; Zhao et al. 2025). Lee S.-H. et al. (2020) further showed that the discriminative components were dispersed throughout the entire gamma band. In the spatial domain, under the common average reference (CAR), EEG electrodes associated with Broca's area, Wernicke's area, and the temporal lobe showed stronger contributions to classification performance (Lee S.-H. et al. 2020). Because full-scalp EEG recordings contain substantial redundancy, optimized channel-selection methods can reduce the number of required electrodes to fewer than 32 with minimal loss in decoding performance (Rekrut et al. 2025; Shah et al. 2024). However, the specific importance of electrode channels was found to vary across individuals (Zhao et al. 2025).

Deep neural networks can automatically extract features, enabling classifiers to use more informative representations and achieve better decoding performance. Lawhern et al. (2018) first proposed EEGNet, a convolutional neural network (CNN)-based architecture, which provided promising performance in multiple EEG-BCI decoding tasks at the time. This model rapidly became the baseline architecture and benchmark for subsequent EEG deep learning models. Lee D.-Y. et al. (2021) proposed a CNN-based Siamese neural network combined with K-Nearest Neighbors (KNN) to compare feature similarities, achieving a 6-class accuracy of 45.0% and a 5-class accuracy of 48.1%. Li F. et al. (2021) proposed a Hybrid-Scale Spatial-Temporal Dilated Convolution Network (HS-STDCN) based on EEGNet, which achieved an average 8-class accuracy of 54.3%. Kim et al. (2023) integrated denoising diffusion probabilistic models with a conditional autoencoder to propose the Diff-E model, which achieved a 5-class accuracy of 60.6%. However, Diff-E lacks feature-level interpretability, which limits its usefulness for guiding further investigation. Despite many studies reporting strong performance, there is no consensus on the optimal model architecture for multi-class imagined speech decoding, highlighting the need for further research.

Building on recent advances in CNNs and BCI research, this study proposes a novel end-to-end spatiotemporal network for decoding imagined speech from EEG signals. The main contributions of our work are summarized as follows:

1. We propose SE-MDTCN, an end-to-end spatiotemporal network for imagined speech decoding from EEG signals. The model consists of an electrode normalization module, a modified SE-based spatial filter, an MDTCN for temporal feature extraction, and an MLP classifier, enabling joint modelling of spatial and temporal EEG patterns in a parameter-efficient manner.
2. We develop task-oriented spatial and temporal modules for EEG-based imagined speech decoding. The proposed normalization scheme is designed to preserve EEG topographical structure better than conventional within-electrode z-score normalization, while the SE-based spatial filter and MDTCN enhance electrode selection and multi-scale temporal feature extraction.
3. Comprehensive experiments were conducted on the 2020 International BCI Competition Track 3 dataset. Results showed that SE-MDTCN outperformed several representative BCI baselines in the participant-dependent classification setting. Extensive ablation studies and feature visualisations were conducted to quantify the contribution of each module and to better understand the model's behaviour. The results support the effectiveness of the proposed components and provide further insight into the limitations of conventional EEGNet-style modules (Lawhern et al. 2018) for imagined speech decoding.
4. The code is publicly available to facilitate reproducibility at https://github.com/Fujie-UTM/SE_MDTCN.

METHODOLOGY

MODEL ARCHITECTURE

Figure 1(a) illustrates the overall architecture of SE-MDTCN. Let $X \in \mathrm{R}^{B \times E \times T}$ denote a batch of EEG trials, where B, E, and T denote the batch size, number of electrodes, and number of time points in each trial.

Electrode Normalization: Traditional preprocessing, such as baseline correction or per-electrode z-score normalization across time points and trials, may disrupt inter-electrode correlations and distort scalp topography (Alday 2019). Here, we propose an electrode normalization method based on CAR and layer normalization (Ba et al. 2016). In practice, the average potential across electrodes approaches a small but non-zero bias, which decreases as the number of electrodes increases to 64 or more (Wolpaw & Wolpaw 2012). Therefore, EEG signals are normalized across electrodes at each time point with learnable affine parameters. Let $x_{e,t}$ denote the measured potential of electrode e at time t for $e = 1, \cdots, E$ and $t = 1, \cdots, T$ in one trial X_i. The normalized potential $x'_{e,t}$ is defined as follows:

$$x'_{e,t} = \frac{x_{e,t} - \bar{x}_t}{\sqrt{s_t^2 + \varepsilon}} \gamma_e + \beta_e \tag{1}$$

where $\bar{x}_t$ and s_t^2 are the cross-electrode average potential and variance at time t, respectively; γ_e and β_e are the learnable affine parameters; and ε is a small constant added for numerical stability. This approach can preserve temporal dynamics and stabilize cross-channel distributions, thereby maintaining the EEG topography and improving inter-subject generalization.

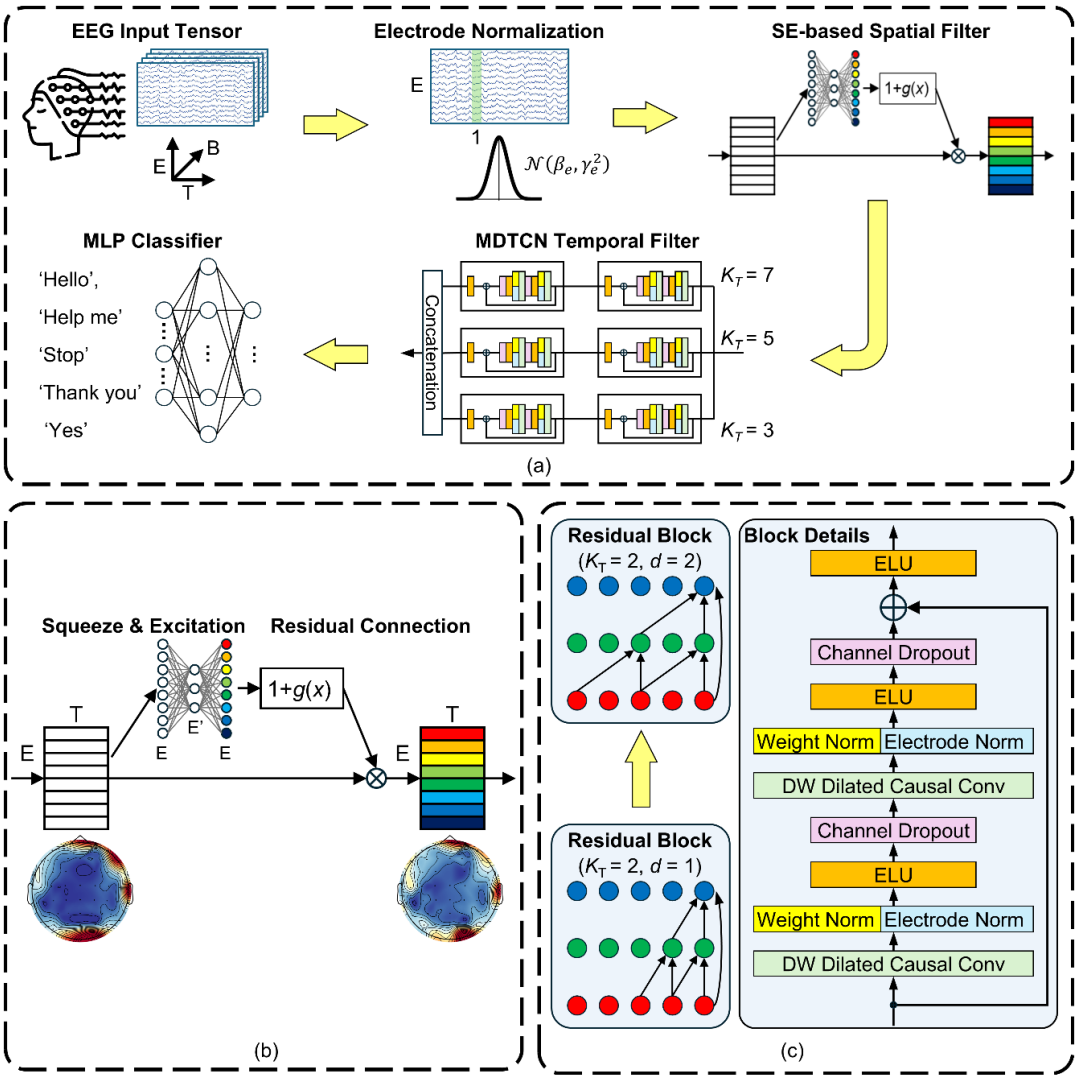


FIGURE 1. Architecture of SE-MDTCN: (a) Overall cascading architecture; (b) SE-based spatial filter; (c) Depthwise TCN temporal filter with kernel size $K_T = 2$ and dilation factor $d_n = \{1, 2\}$

SE Attention and Spatial Filter: The SE attention mechanism performs dynamic recalibration of channel-wise feature weights and typically comprises two stages: squeeze and excitation (Hu et al. 2018). In its 1D variant, the input tensor is $X_i \in \mathrm{R}^{C\times L}$, where C denotes the feature channels, and L represents the sequential length (Wang et al. 2024). In the squeeze stage, a channel descriptor is obtained by applying 1D global average pooling (GAP) along the sequence dimension. In the excitation stage, fully connected layers with nonlinear activations learn inter-channel dependencies and produce channel weights. The output is obtained by scaling each channel with its corresponding weight, as follows:

$$Y_i = X_i \cdot \sigma\{W_2 \cdot \delta[W_1 \cdot \mathrm{GAP}(X_i)]\} \tag{2}$$

where W_1 and W_2 denote the weight matrices of the first and second fully connected layers, respectively. The first layer $W_1 \in \mathrm{R}^{C'\times C}$ reduces the dimensionality from C to C', while the second layer $W_2 \in \mathrm{R}^{C\times C'}$ restores it back to C. The operator σ denotes the sigmoid function, and δ denotes the exponential linear unit (ELU) activation function (Clevert et al. 2016).

To apply the SE module as a spatial filter for EEG electrodes, the feature-channel dimension C is set to the number of electrodes E and the sequence length L is set to the trial length T. The adopted parameters and modifications are shown in Figure 1(b). First, the reduced dimension E' is chosen to be 32 to limit performance degradation (Rekrut et al. 2025; Shah et al. 2024). Second, the GAP in Equation (2) is replaced with global variance pooling (GVP) of the EEG signals, yielding a more interpretable spatial filter: task-relevant electrodes tend to exhibit significantly higher power or stronger task-related power fluctuations, while variance pooling also suppresses uninformative DC offsets in GAP. Third, a residual connection with an amplification factor is added to promote stable gradient propagation and enhance the SNR of task-related electrodes. It is computed as follows:

$$Y_i = X_i \cdot \langle 1 + r_a \cdot \sigma\{W_2 \cdot \delta[W_1 \cdot \mathrm{GVP}(X_i)]\}\rangle \tag{3}$$

$$r_a = 2 \cdot \sqrt{\frac{\max \mathrm{var}(X_i)}{\min \mathrm{var}(X_i)}} \tag{4}$$

where r_a denotes the amplification factor, which equals twice the square root of the maximum-to-minimum variance ratio across all electrodes.

Variance-based pooling is adopted because imagined-speech EEG discriminability may be reflected not only in mean-amplitude changes but also in task-related power fluctuations across electrodes. In contrast to global average pooling, which may be affected by near-zero-mean oscillatory EEG activity and DC offsets, global variance pooling provides a compact descriptor of electrode-wise signal energy within each trial. This makes the spatial attention weights more closely related to electrode-wise task modulation. The residual scaling term is introduced to avoid completely suppressing electrodes that may contain weak but useful information and to improve gradient propagation through the spatial filtering module. In this design, the SE module acts as an adaptive electrode-recalibration mechanism rather than a hard channel-selection procedure.

MDTCN and Temporal Filter: Temporal Convolutional Networks (TCNs) (Bai et al. 2018) are one-dimensional convolutional architectures for time-series modelling. They support parallel computation, flexible receptive fields (RFs), and stable gradients, which have shown advantages over traditional sequence models. As illustrated in Figure 1(c), a TCN comprises three principal components: causal convolutions, dilated convolutions, and residual connections. Causal convolutions ensure that the output at time t is computed using only inputs from time t and earlier, therefore preserving temporal causality. Dilated convolutions expand the RFs without a proportional increase in depth, allowing the model to capture long-range dependencies. Residual connections alleviate the vanishing-gradient problem in deep networks. Given a TCN with N temporal convolutional layers, each layer's kernel size is K_T, and the dilation factor $d_n = 2^n - 1$ ($n = 1, 2, \cdots, N$), the RF of layer n is determined by $(K_T - 1)\cdot d_n$. The final RF of the TCN is determined by:

$$\mathrm{RF} = 1 + 2 \cdot (K_T - 1)(2^N - 1) \tag{5}$$

Compared with the original TCN (Bai et al. 2018) and its BCI variants (Ingolfsson et al. 2020; Salami et al. 2022), the proposed TCN incorporates the following modifications. First, weight normalization and spatial dropout are retained as in Bai et al. (2018), instead of employing batch normalization and element-wise dropout as in Ingolfsson et al. (2020) and Salami et al. (2022), because our model operates on raw EEG rather than pre-extracted features. Second, the rectified linear unit (ReLU) is replaced with the ELU to alleviate vanishing gradients (Lawhern et al. 2018). Third, the original standard convolutions are replaced with depthwise convolutions, where each input channel is filtered by its own kernel, thereby avoiding cross-channel mixing and preserving inter-electrode spatial structure. An additional electrode normalization layer is added to further constrain electrode-wise relationships and accelerate model convergence. Fourth, following InceptionTime (Ismail Fawaz et al. 2020), multiple

depthwise TCN branches with different kernel sizes are used to capture multi-scale features and enhance robustness.

Specifically, $K_T \in \{3, 5, 7\}$ and $N = 4$ are used, corresponding to approximate receptive fields of 238, 473, and 707 ms, respectively, at a sampling frequency of 256 Hz. These kernel sizes are selected to capture short-, intermediate-, and long-range temporal patterns following the multi-scale principle of Inception-type models, while keeping the filters compact. Because dilation already expands the effective receptive field, larger kernels are unnecessary and would increase the number of parameters, computational cost, and early-stage optimization difficulty with limited additional benefit.

MLP Classifier: The features extracted by the TCN are subsequently fed into a 2-layer MLP classifier. First, the features are normalized using Layer Normalization, then projected into a higher-dimensional space with an expansion factor of 4, followed by the application of the Gaussian Error Linear Unit (GELU) activation function (Hendrycks & Gimpel 2016). The projection to a high-dimensional space facilitates better clustering and separability of the features (Gorban et al. 2021). The GELU function provides a smoother, probabilistic approach to modelling activations than other activation functions. Finally, the features are projected onto output neurons corresponding to the target classes and passed to the loss function for training.

DATASET AND PREPROCESSING

To validate SE-MDTCN, the public dataset from the 2020 International BCI Competition Track 3 was used in this study (https://osf.io/pq7vb/overview). The experiments were conducted in accordance with the Declaration of Helsinki and were approved by the Institutional Review Board at Korea University (KUIRB-2019-0143-01). Participants were instructed to imagine pronouncing the given prompts as if speaking, without articulatory movement or vocalization. Thus, imagined speech was assumed to involve the same functional processes as overt articulation, except for the absence of final muscle activations. The dataset included five prompts ("hello", "help me", "stop", "thank you", and "yes") and recordings from 15 healthy participants, each completing 400 trials. The original split comprised 60, 10, and 10 trials per class for training, validation, and testing, respectively (Jeong et al. 2022). In this study, the original validation and test sets were merged into a single test set, and the original training set was used for model training and validation.

The preprocessing pipeline was performed as follows. First, the segmented EEG signals were downsampled to 256 Hz. Signals were re-referenced using the CAR. ICA was then performed using the MNE (Gramfort et al. 2013) and MNE-ICALabel packages (Li A. et al. 2022), and artifactual components were automatically rejected at a probability threshold of $p = 0.9$. The rejected artifacts included muscle artifacts, eye-blink, heartbeat, line noise and channel noise. In this study, ICA-cleaned EEG denotes EEG after ICA-based artifact reduction, whereas the unprocessed data are referred to as "raw" EEG. EEG epochs were extracted from −500 to 2,000 ms relative to cue onset, as the pre-onset standby period has been demonstrated to contain speech-related information (Qiu et al. 2025). A 60 Hz notch filter was applied to remove the power line noise. Finally, a fourth-order zero-phase Butterworth band-pass filter was used to extract gamma-band activity (30-125 Hz), consistent with previous studies (Lee S.-H. et al. 2020; Qiu et al. 2025; Shah et al. 2024).

TRAINING CONFIGURATION

All experiments were implemented using Python and PyTorch and performed on a single NVIDIA RTX 3080 GPU. The original training set was divided into five folds for checkpoint selection. In each cross-validation iteration, four folds were used for training and the remaining fold was used for validation and early stopping, resulting in five candidate checkpoints. The checkpoint with the best validation performance among these five candidates was selected as the final model. The original validation and test sets were merged as a larger held-out evaluation set to provide a more stable estimate of final performance. This held-out set was evaluated only once and was never used during training, checkpoint selection, or early stopping. The Mixup method (Zhang et al. 2018) was employed for data augmentation. The models were optimized with AdamW using a learning-rate scheduler, which included linear warm-up and cosine decay. The peak learning rate was set to 0.01, with a warm-up phase of 50 epochs and a total of 500 epochs for the experiments. The batch size was set to 64. Cross-entropy loss with label smoothing was used, with a smoothing factor of 0.1.

PERFORMANCE METRICS

Balanced accuracy (Acc) and Cohen's Kappa (κ) were employed to evaluate the performance. Balanced accuracy represents the macro-average of the proportion of correctly classified instances per class and is defined as follows:

$$\mathrm{Acc} = \frac{1}{M}\sum_{i=1}^{M}\frac{TP_i}{TP_i+FN_i} \tag{6}$$

where TP_i and FN_i represent the number of true positives and false negatives for class i, respectively, and

M is the total number of classes. Cohen's Kappa measures the agreement between the tested classifier and an ideal perfect classifier, which is defined as follows:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{7}$$

where P_o is the proportion of agreement, equivalent to the classification accuracy of the tested model, and P_e is the proportion of chance agreement, equivalent to the chance-level accuracy. Cohen's Kappa provides a metric for comparing multi-class BCI paradigms.

MODEL COMPARISON

To evaluate SE-MDTCN, the following six representative BCI models were selected as baselines and evaluated using conventional within-electrode z-score normalization.

1. EEGNet-8,2 (Lawhern et al. 2018) is a compact CNN for EEG decoding. It employs depthwise and separable convolutions to decouple temporal and spatial filtering and to fuse features, providing a general backbone for multiple BCI paradigms. The numbers 8 and 2 denote the numbers of temporal and spatial filters, respectively.
2. EEG-TCNet (Ingolfsson et al. 2020) extends EEGNet by appending a TCN module after EEGNet's feature-extraction layer, thereby expanding the RF with minimal additional parameters and improving time-series feature extraction.
3. EEG-Conformer (Song Y. et al. 2023) combines EEGNet-style CNN feature extractors with a Transformer encoder. CNNs capture local features, while self-attention models global dependencies among temporal features.
4. HS-STDCN (Li F. et al. 2021) is an end-to-end model that integrates hybrid-scale temporal convolutions, depthwise spatial convolutions, and dilated blocks. This network is designed for imagined speech classification, achieving an average accuracy of 54.3% on a Chinese 8-word classification task.
5. TSception (Ding et al. 2023) combines dynamic temporal layers with multi-scale 1D convolutions, asymmetric spatial filtering, and advanced fusion. Although originally proposed for emotion recognition, it has recently shown effectiveness for imagined-speech classification (Ahn et al. 2023).
6. Dynamical Graph Convolutional Neural Network (DGCNN) (Song T. et al. 2020) formulates multichannel EEG decoding as a graph-based representation learning problem, where features extracted from each electrode are represented as graph nodes and inter-electrode dependencies are modelled by a dynamically learned adjacency matrix. Originally proposed for EEG-based emotion recognition, DGCNN has shown strong participant-dependent EEG decoding performance across multiple datasets (Liu et al. 2025). In this study, differential entropy features were used as the input features because of its established effectiveness in EEG-based emotion recognition.

All baseline models used the same dataset split, preprocessing protocol, training procedure, and data augmentation as the proposed model. The only exception was normalization: SE-MDTCN performs electrode-wise normalization through its built-in normalization module, whereas the baseline models followed their original settings and used conventional within-electrode z-score normalization as an external preprocessing step. For these baselines, z-score parameters were fitted only on the training/validation data within cross-validation and then applied to the held-out test data, ensuring that no test-set information was used during preprocessing or tuning.

All models were trained using AdamW, and optimizer-related hyperparameters were optimized through grid-searched to obtain the best baseline performance. For EEGNet, EEG-TCNet, HS-STDCN, TSception, and DGCNN, learning rates of {0.01, 0.001} and weight decay values of {0.01, 0} were searched. For EEG-Conformer, learning rates of {0.001, 0.0001} and weight decay values of {0.01, 0} were searched because higher learning rates caused instability or non-convergence. Architecture-related settings were kept consistent with the original configurations, except for necessary adjustments to temporal kernel sizes and electrode numbers. All paired performance differences between the proposed model and the optimally tuned baseline models were evaluated using a one-sided Wilcoxon signed-rank test.

ABLATION EXPERIMENTS

Ablation experiments were performed to evaluate the contributions of electrode normalization, the SE-based spatial filter, and the MDTCN temporal filter. To further investigate the effects of normalization and the primary spatiotemporal feature-extraction modules in EEGNet and its variants, three additional experiments were conducted: (i) replacing electrode normalization in the SE-MDTCN with conventional within-electrode z-score normalization; (ii) inserting a 2D temporal convolution with kernel size $(1, T_0)$ before the SE-based spatial filter, where T_0 was set to half the sampling frequency to capture frequency ≥ 2 Hz; and (iii) replacing the SE-based spatial filter with depthwise spatial convolutions of kernel size $(E, 1)$, where E is the number of electrodes. All paired performance differences between the full and ablated models were further evaluated using a one-sided Wilcoxon signed-rank test.

RESULTS AND DISCUSSION

MODEL PERFORMANCE

Figure 2 presents the participant-wise classification accuracy of the proposed SE-MDTCN on raw EEG data, together with the corresponding confusion matrix. Figure 2(a) shows that the proposed model achieved an average accuracy of 74.4% ± 8.5% across participants, substantially exceeding the chance level of 20%. Most participants obtained accuracies within the range of 70%–80%. Although moderate inter-participant variability was observed, the overall performance remains stable. Furthermore, the confusion matrix in Figure 2(b) shows relatively balanced class-wise accuracies ranging from 72.0% to 75.3% and no evident performance degradation for any specific prompt. This suggests that model performance is not strongly affected by prompt-specific factors such as prompt length.

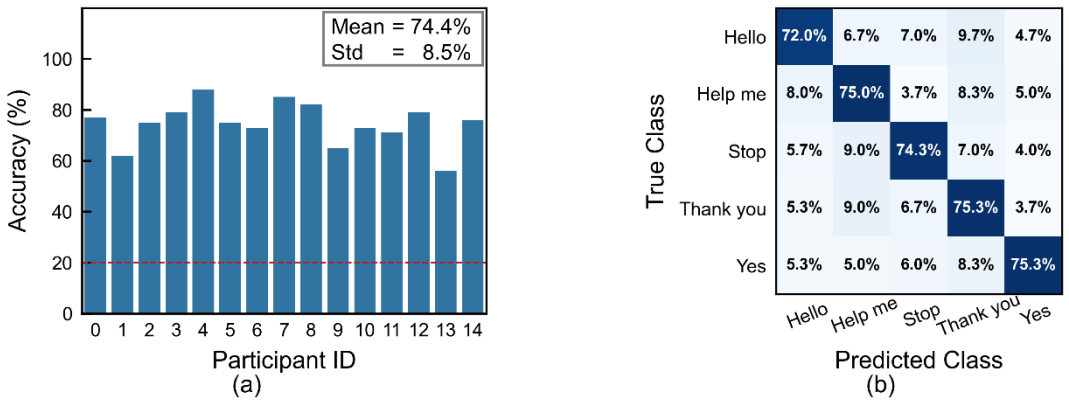


FIGURE 2. Classification performance of SE-MDTCN: (a) Accuracy across 15 participants, with the red dashed line indicating chance level. (b) Confusion matrix of 5 imagined-speech prompts

MODEL COMPARISON

Table 1 compares the performance of the proposed SE-MDTCN with representative baseline models, with statistically significant differences marked by asterisks. The performance of SE-MDTCN on ICA-cleaned EEG is also included to illustrate the impact of artifact reduction. As shown in Table 1, most of the baseline models exhibited lower performance, with HS-STDCN performing near chance level. The proposed SE-MDTCN achieved the highest average accuracy on raw EEG data and the second-highest accuracy on ICA-cleaned EEG.

ABLATION RESULTS

Table 2 reports the ablation results with statistically significant differences marked by asterisks. Electrode normalization had the largest impact on model performance. The removal of this module resulted in a 9.6% accuracy decrease, while replacing it with within-electrode z-score normalization led to a gentler decrease of approximately 3.9% on the test set. The contributions of the SE-based spatial filter and the MDTCN temporal filter were similar. The results support the effectiveness of each module to overall performance. In addition, both replacing the 2D spatial filter and adding the 2D temporal filter from EEGNet led to performance degradation, suggesting that the EEGNet-based BCI model may not be suitable for the present dataset.

TABLE 1. Model comparison results. The best performance is highlighted in bold type

Model	Train/Validation	Test	
	Mean (Std) Acc (%)	Mean (Std) Acc (%)	κ
EEGNet-8,2 (Lawhern et al. 2018)	39.0 (3.3)*	28.8 (5.6)*	0.11
EEG-TCNet (Ingolfsson et al. 2020)	38.2 (3.7)*	30.5 (5.4)*	0.13
EEG-Conformer (Song Y. et al. 2023)	43.6 (3.1)*	36.9 (5.2)*	0.21
HS-STDCN (Li F. et al. 2021)	36.6 (3.8)*	26.7 (6.2)*	0.08
TSception (Ding et al. 2023)	59.2 (5.8)*	50.5 (8.6)*	0.38
DGCNN (Song T. et al. 2020)	55.5 (5.9)*	50.4 (9.2)*	0.38
Proposed model (ICA-cleaned EEG)	75.5 (6.4)*	69.1 (8.5)*	0.61
Proposed model (raw EEG)	**78.1 (6.8)**	**74.4 (8.5)**	**0.68**

* $p < 0.05$

TABLE 2. Ablation results of SE-MDTCN model on the raw EEG data. The best performance is highlighted in bold type

Model	Train/Validation	Test	
	Mean (Std) Acc (%)	Mean (Std) Acc (%)	κ
Full SE-MDTCN model	**78.1 (6.8)**	**74.4 (8.5)**	**0.68**
Model w/o electrode normalization	72.0 (7.8)*	64.8 (8.4)*	0.56
Model w/o SE spatial filter	75.0 (7.2)*	69.5 (10.8)*	0.62
Model w/o MDTCN temporal filter	76.6 (7.1)*	70.3 (9.7)*	0.63
Model w/ within-electrode normalization	75.6 (7.4)*	70.5 (8.8)*	0.63
Model w/ 2D temporal convolution	66.3 (10.2)*	58.7 (11.4)*	0.48
Model w/ 2D spatial convolution	65.6 (10.3)*	57.8 (10.9)*	0.47

* $p < 0.05$

DISCUSSION

Both the comparison and ablation experiments support the effectiveness of SE-MDTCN for imagined-speech decoding. As shown in Table 1, automatic artifact removal reduced the accuracy by 5.3% on the test set, suggesting that the higher raw-EEG performance may result from the combined contributions of neural EEG activity and residual artifacts. Among possible non-neural components, muscle-related activity associated with unconscious or subtle articulatory movements may be a plausible task-related source. Because artifact-removal methods cannot perfectly separate neural from non-neural signals, ICA and ICLabel were used as a conservative artifact-reduction procedure. In this procedure, components classified as artifacts with a probability greater than 0.9 were removed to reduce contamination while avoiding excessive removal of EEG activity. The ICA-cleaned EEG may therefore still contain residual artifacts, and the raw-to-clean comparison should be interpreted as supporting evidence rather than definitive proof of purely neural decoding. Nevertheless, the modest decrease from raw to ICA-cleaned EEG indicates that SE-MDTCN remains effective after artifact reduction.

Under the present participant-dependent protocol using the imagined-speech EEG dataset, all EEGNet-derived baselines reported in Table 1, including EEGNet-8,2, EEG-TCNet, EEG-Conformer, and HS-STDCN, exhibited substantially weaker generalization performance than the proposed SE-MDTCN. A possible explanation is that EEGNet-style models rely heavily on early temporal and spatial convolutional operations for primary EEG feature extraction. These models typically use a 2D temporal convolution with a kernel size of (1, T_0), followed by a depthwise 2D spatial convolution with a kernel size of (E, 1). Under the current preprocessing pipeline and data split, such large early-stage temporal kernels may increase optimization difficulty and reduce the reliability of learning stable task-related temporal representations, particularly for non-stationary and low-SNR EEG signals. Moreover, full-channel spatial kernels may aggregate electrode-level information prematurely, thereby weakening the intrinsic scalp-topographical organization of EEG activity.

In cascaded architectures that lack sufficient residual pathways, these early-stage suboptimal representations may further hinder the optimization of subsequent modules. By contrast, SE-MDTCN integrates electrode-wise normalization, residual spatial recalibration, and multi-scale temporal modelling, providing a potentially more stable and robust feature-enhancement framework for imagined-speech EEG decoding.

This interpretation is supported by the ablation results in Table 2. Replacing the proposed SE-based spatial filter with a 2D depthwise spatial convolution reduced the test accuracy by 16.6%, while adding a two-dimensional temporal convolution reduced the test accuracy by 15.7%. These results suggest that the performance gain of SE-MDTCN does not simply come from increasing model depth or stacking existing modules. Instead, the proposed architecture is designed to balance feature-extraction capacity with optimization stability. Extensive residual connections are introduced throughout the network to preserve low-level electrode-wise information, reduce the risk of suppressing useful EEG components through excessive early-stage transformations, and facilitate gradient propagation during back-propagation.

Furthermore, SE-MDTCN explicitly decouples early-stage spatial recalibration from subsequent temporal modelling through its modular architecture. The variance-guided SE-based spatial filter adaptively enhances electrode-level representations according to task-related spatial variation, whereas the MDTCN captures temporal dependencies at multiple receptive-field scales. This design differs from EEGNet-related baselines by emphasizing the preservation of electrode topographic structure, adaptive information recalibration, and temporal feature enhancement. Therefore, the proposed architecture provides a more task-specific solution for imagined-speech EEG classification by improving trainability, preserving electrode topographic structure, and extracting multi-scale temporal dynamics.

Regarding DGCNN, graph-based modelling is useful for capturing inter-channel EEG dependencies, but several factors limit its application to imagined-speech decoding. Differential entropy is a handcrafted statistical feature and may not fully capture nonlinear and time-varying EEG dynamics. In addition, DGCNN mainly emphasizes spatial dependencies, whereas temporal evolution across EEG segments is not explicitly modelled. Its Chebyshev polynomial-based graph convolution can also introduce excessive feature mixing or over-smoothing, potentially weakening electrode-specific discriminative patterns (Jin & Zhu 2025). These limitations highlight the complementary motivation of SE-MDTCN, which integrates adaptive spatial recalibration, multi-scale temporal modelling, and residual feature preservation.

FEATURE VISUALISATION

To investigate the mechanism of the SE-based spatial filter, participants 4 (Acc: 88.0%) and 7 (Acc: 85.0%) were selected. Figure 3 shows their topographies before and after application of the spatial filter, computed from normalized mean power. The most active regions for the same imagined-speech prompt differ between the two participants. Moreover, due to the residual connection, the pre- and post-filter topographies exhibit similar overall patterns. After filtering, activation increased for most electrodes, particularly in the temporal lobe, whereas the initial focal high-activation regions become less dominant.

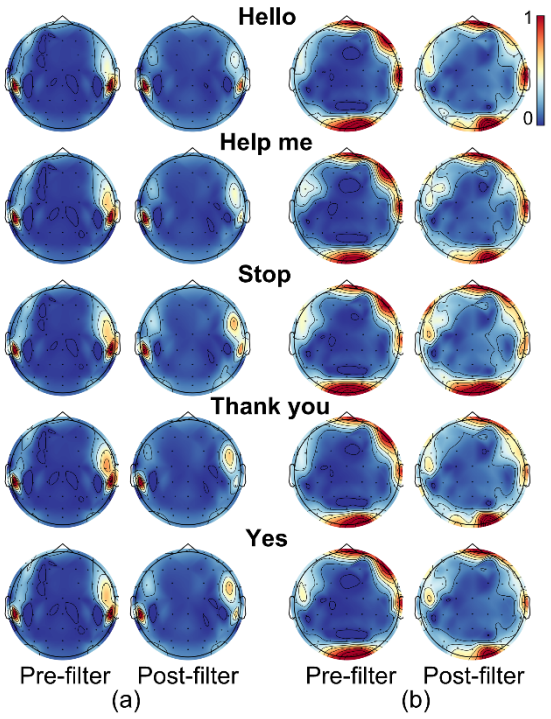


FIGURE 3. Topographical maps before and after the SE-based spatial filter using normalized average power: (a) Participant 4; (b) Participant 7

Figure 3 suggests that the most active electrodes vary across participants, which is consistent with previous findings on participant-specific imagined-speech EEG patterns (Zhao et al. 2025). This differs from the more spatially localized and lateralized topographical patterns commonly observed in limb motor imagery, especially the event-related desynchronization/synchronization (ERD/ERS) over the contralateral sensorimotor cortex during left- or right-hand imagination (Edelman et al. 2025). Imagined speech may engage a dual-stream speech network, involving ventral-stream semantic retrieval and dorsal-stream auditory-to-articulatory mapping associated with articulation preparation and motor planning (Panachakel & Ramakrishnan 2021). These processes may vary substantially across participants because internal speech strategies, phonological imagery, and covert articulatory tendencies are not directly observable. Consequently, imagined-speech EEG may show less spatially consistent topographies than limb motor imagery,

making participant-independent decoding more challenging. This difference further explains why adaptive spatial recalibration and participant-dependent modelling may be particularly important for imagined-speech EEG classification.

To further explore the feature extraction capabilities qualitatively, the normalized *t*-Distributed Stochastic Neighbor Embedding (*t*-SNE) method (Van der Maaten & Hinton 2008) was employed to visualize the learned features. Given that the proposed model achieved high performance for participant 7, the analysis focused on this participant. All features were projected into a two-dimensional embedding space, as shown in Figure 4.

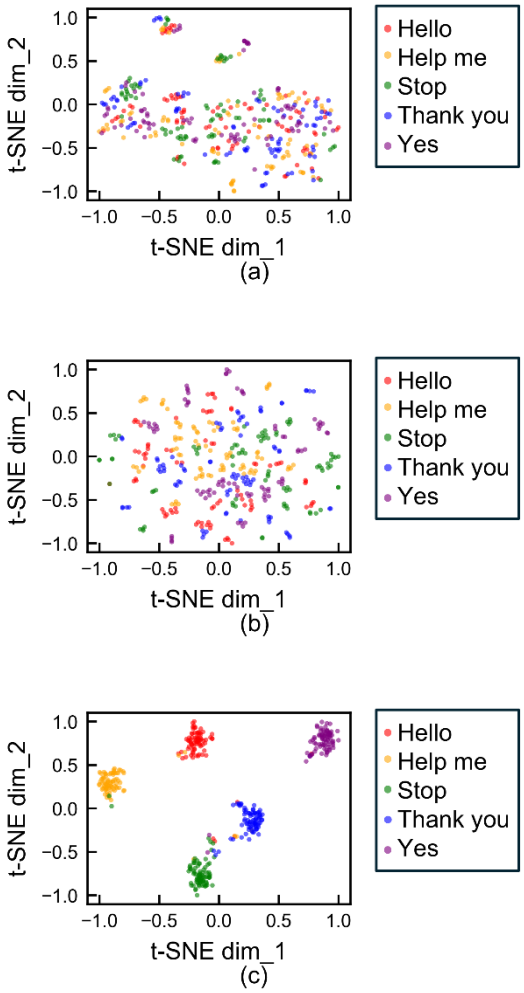


FIGURE 4. Normalized *t*-SNE feature distributions in participant 7: (a) Variance features of raw EEG data; (b) Features from the expansion layer of the classifier; (c) Classifier outputs.

The *t*-SNE results depicted in Figure 4 show that variance-based features are highly overlapping (Figure 4(a)), suggesting limited class separability. After spatial and temporal filtering and dimensional expansion, the embeddings become more dispersed with minimal overlap (Figure 4(b)). After projection by the final classifier, features form well-separated clusters with large margins (Figure 4(c)), indicating that SE-MDTCN progressively extracts discriminative representations across processing stages.

CONCLUSION

In this study, inspired by EEG signal characteristics, we proposed an interpretable SE-MDTCN model for imagined-speech decoding. The proposed framework improves training stability and preserves electrode-wise EEG structure through built-in electrode normalization, extensive residual connections, adaptive spatial recalibration, and multi-scale depthwise temporal modelling. Under the participant-dependent protocol of the 2020 International BCI Competition Track 3 dataset, SE-MDTCN achieved higher decoding accuracy than the baseline models compared in this study. The ablation results support the contribution of the proposed components, while the feature visualizations provide qualitative insight into participant-specific spatial patterns and learned representations. In addition, large early-stage temporal kernels and full-channel spatial kernels in baseline models were less robust than the proposed residual spatial-recalibration and multi-scale temporal-modelling design in this study. These findings suggest that SE-MDTCN is an effective participant-dependent decoding framework for this imagined-speech EEG benchmark.

Despite the encouraging results, this study has several limitations. First, SE-MDTCN was evaluated under a participant-dependent setting, which is common in imagined-speech EEG studies but may overestimate performance for real-world cross-user BCI deployment. Future work should investigate participant-independent protocols and use transfer learning or shared latent representation learning to improve robustness across users. Second, although ICA-cleaned EEG was used to assess decoding after artifact reduction, the present experiments cannot fully separate neural and non-neural contributions. Stricter artifact-control procedures and auxiliary physiological recordings are needed to better identify the origin of discriminative signals. Third, the dataset contains only five predefined prompts. A larger vocabulary would be needed to examine the effects of word length, syllabic structure, phonological complexity, and semantic factors. Future work should also investigate why spatial-module perturbations caused greater performance degradation than temporal-module perturbations, despite the inherently temporal nature of speech. Addressing these issues may

provide deeper insight into imagined-speech EEG representations and support more generalizable EEG-based decoding models.

ACKNOWLEDGEMENT

This research was supported by the Universiti Teknologi Malaysia Fundamental Research (UTMFR) under Cost Centre No. Q.J130000.3823.23H68. The authors gratefully acknowledge the support provided by UTM.

DECLARATION OF COMPETING INTEREST

None.

REFERENCES

Ahn, H.-J., Lee, D.-H., Jeong, J.-H. & Lee, S.-W. 2023. Multiscale convolutional transformer for EEG classification of mental imagery in different modalities. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 31: 646-656.

Alday, P.M. 2019. How much baseline correction do we need in ERP research? extended GLM model can replace baseline correction while lifting its limits. *Psychophysiology* 56(12): e13451.

Ba, J.L., Kiros, J.R. & Hinton, G.E. 2016. Layer normalization. *ArXiv Preprint* arXiv:1607.06450. https://arxiv.org/abs/1607.06450 [29 April 2026].

Bai, S., Kolter, J.Z. & Koltun, V. 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *ArXiv Preprint* arXiv:1803.01271. https://arxiv.org/abs/1803.01271 [29 April 2026].

Clevert, D.-A., Unterthiner, T. & Hochreiter, S. 2016. Fast and accurate deep network learning by exponential linear units (elus). *ArXiv Preprint* arXiv:1511.07289. https://arxiv.org/abs/1511.07289 [29 April 2026].

Coupé, C., Oh, Y.M., Dediu, D. & Pellegrino, F. 2019. Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Science Advances* 5(9): eaaw2594.

Ding, Y., Robinson, N., Zhang, S., Zeng, Q. & Guan, C. 2023. TSception: capturing temporal dynamics and spatial asymmetry from EEG for emotion recognition. *IEEE Transactions on Affective Computing* 14(3): 2238-2250.

Edelman, B.J., Zhang, S., Schalk, G., Brunner, P., Müller-Putz, G., Guan, C. & He, B. 2025. Non-invasive brain-computer interfaces: state of the art and trends. *IEEE Reviews in Biomedical Engineering* 18: 26-49.

Gorban, A.N., Grechuk, B., Mirkes, E.M., Stasenko, S.V. & Tyukin, I.Y. 2021. High-dimensional separability for one- and few-shot learning. *Entropy* 23(8): 1090.

Gramfort, A., Luessi, M., Larson, E., Engemann, D.A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L. & Hämäläinen, M. 2013. MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience* 7: 267.

Hendrycks, D. & Gimpel, K. 2016. Gaussian error linear units (gelus). ArXiv Preprint arXiv:1606.08415. https://arxiv.org/abs/1606.08415 [29 April 2026].

Hu, J., Shen, L. & Sun, G. 2018. Squeeze-and-excitation networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2018)* 7132-7141.

Ingolfsson, T.M., Hersche, M., Wang, X., Kobayashi, N., Cavigelli, L. & Benini, L. 2020. EEG-TCNet: an accurate temporal convolutional network for embedded motor-imagery brain–machine interfaces. *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC 2020)* 2958-2965.

Ismail Fawaz, H., Lucas, B., Forestier, G., Pelletier, C., Schmidt, D.F., Weber, J., Webb, G.I., Idoumghar, L., Muller, P.-A. & Petitjean, F. 2020. InceptionTime: finding AlexNet for time series classification. *Data Mining and Knowledge Discovery* 34(6): 1936-1962.

Jeong, J.-H., Cho, J.-H., Lee, Y.-E., Lee, S.-H., Shin, G.-H., Kweon, Y.-S., Millán, J.d.R., Müller, K.-R. & Lee, S.-W. 2022. 2020 International brain–computer interface competition: a review. *Frontiers in Human Neuroscience* 16: 898300.

Jin, Y. & Zhu, X. 2025. Oversmoothing alleviation in graph neural networks: a survey and unified view. *Knowledge and Information Systems* 67(12): 11259-11285.

Kim, S., Lee, Y.-E., Lee, S.-H. & Lee, S.-W. 2023. Diff-E: diffusion-based learning for decoding imagined speech EEG. *24th Annual Conference of the International Speech Communication Association (INTERSPEECH 2023)* 1159-1163.

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

Lee, D.-Y., Lee, M. & Lee, S.-W. 2021. Decoding imagined speech based on deep metric learning for intuitive BCI communication. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 29: 1363-1374.

Lee, S.-H., Lee, M. & Lee, S.-W. 2020. Neural decoding of imagined speech and visual imagery as intuitive paradigms for BCI communication. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 28(12): 2647-2659.

Li, A., Feitelberg, J., Saini, A.P., Höchenberger, R. & Scheltienne, M. 2022. MNE-ICALabel: automatically annotating ICA components with ICLabel in Python. *Journal of Open Source Software* 7(76): 4484.

Li, F., Chao, W., Li, Y., Fu, B., Ji, Y., Wu, H. & Shi, G. 2021. Decoding imagined speech from EEG signals using hybrid-scale spatial-temporal dilated convolution network. *Journal of Neural Engineering* 18(4): 0460c4.

Liu, H., Yang, S., Zhang, Y., Wang, M., Gong, F., Xie, C., Liu, G., Liu, Z., Liu, Y.-J., Lu, B.-L. & Zhang, D. 2025. LibEER: a comprehensive benchmark and algorithm library for EEG-based emotion recognition. *IEEE Transactions on Affective Computing* 16(4): 3596-3613.

Luo, S., Angrick, M., Coogan, C., Candrea, D.N., Wyse-Sookoo, K., Shah, S., Rabbani, Q., Milsap, G.W., Weiss, A.R., Anderson, W.S., Tippett, D.C., Maragakis, N.J., Clawson, L.L., Vansteensel, M.J., Wester, B.A., Tenore, F.V., Hermansky, H., Fifer, M.S., Ramsey, N.F. & Crone, N.E. 2023. Stable decoding from a speech BCI enables control for an individual with ALS without recalibration for 3 months. *Advanced Science* 10(35): 2304853.

Metzger, S.L., Littlejohn, K.T., Silva, A.B., Moses, D.A., Seaton, M.P., Wang, R., Dougherty, M.E., Liu, J.R., Wu, P., Berger, M.A., Zhuravleva, I., Tu-Chan, A., Ganguly, K., Anumanchipalli, G.K. & Chang, E.F. 2023. A high-performance neuroprosthesis for speech decoding and avatar control. *Nature* 620(7976): 1037-1046.

Panachakel, J.T. & Ramakrishnan, A.G. 2021. Decoding Covert Speech From EEG-A Comprehensive Review. Frontiers in Neuroscience 15: 642251.

Qiu, F., Tan, T.S., Tiong, M.F.T. & Sheikh, J. 2025. Optimizing temporal and spectral interval selection for imagined speech decoding from EEG. *Emerging Science and Technology for Human Well-Being, the International Human-Centered Technology Conference 2024 (iHumEnTech 2024)* 215-223.

Rekrut, M., Ihl, J., Jungbluth, T. & Krüger, A. 2025. How low can you go: evaluating electrode reduction methods for EEG-based speech imagery BCIs. *Frontiers in Neuroergonomics* 6: 1578586.

Salami, A., Andreu-Perez, J. & Gillmeister, H. 2022. EEG-ITNet: an explainable inception temporal convolutional network for motor imagery classification. *IEEE Access* 10: 36672-36685.

Schultz, T., Wand, M., Hueber, T., Krusienski, D.J., Herff, C. & Brumberg, J.S. 2017. Biosignal-based spoken communication: a survey. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25(12): 2257-2271.

Shah, U., Alzubaidi, M., Mohsen, F., Alam, T. & Househ, M. 2024. Ensemble-based feature engineering mechanism to decode imagined speech from brain signals. *Informatics in Medicine Unlocked* 47: 101491.

Song, T., Zheng, W., Song, P. & Cui, Z. 2020. EEG emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing* 11(3): 532-541.

Song, Y., Zheng, Q., Liu, B. & Gao, X. 2023. EEG conformer: convolutional transformer for EEG decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 31: 710-719.

Van der Maaten, L. & Hinton, G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9: 2579-2605.

Wang, W., Li, B., Wang, H., Wang, X., Qin, Y., Shi, X. & Liu, S. 2024. EEG-FMCNN: a fusion multi-branch 1D convolutional neural network for EEG-based motor imagery classification. *Medical & Biological Engineering & Computing* 62(1): 107-120.

Wolpaw, J.R. & Wolpaw, E.W. (eds). 2012. *Brain–computer interfaces: principles and practice*. New York: Oxford University Press.

Zhang, H., Cisse, M., Dauphin, Y.N. & Lopez-Paz, D. 2018. Mixup: beyond empirical risk minimization. *6th International Conference on Learning Representations (ICLR 2018)* 2866-2878.

Zhao, Z., Li, Y., Peng, Y., Camilleri, K. & Kong, W. 2025. Multi-view graph fusion of self-weighted EEG feature representations for speech imagery decoding. *Journal of Neuroscience Methods* 418: 110413.