Feature-Level Analysis and Robust Baselines for EEG-Based Imagined Speech Recognition on the ASU Dataset

Hatem T M Duhair[1], Masrullizam Mat Ibrahim[2]*,
Jamil Abedalrahim Jamil Alsayaydeh[3]*, Mazen Farid[4]*, Safarudin Gazali Herawan[5]
Department of Engineering Technology-Fakulti Teknologi Dan Kejuruteraan Elektronik Dan Komputer (FTKEK), Universiti Teknikal Malaysia Melaka, Melaka, Malaysia[1, 2, 3]
Faculty of Information Science and Technology (FIST), Multimedia University, Melaka 75450, Malaysia[4]
Centre for Intelligent Cloud Computing, COE for Advanced Cloud, Multimedia University, Melaka 75450, Malaysia[4]
Faculty of Engineering-Industrial Engineering Department, Bina Nusantara University, Jakarta, Indonesia[5]

***Abstract*—Imagined speech decoding from non-invasive electroencephalography remains a challenging problem, especially when moving beyond small vocabularies and optimistic evaluation protocols. This work revisits the Arizona State University (ASU) imagined speech dataset and treats it as a rigorous ten-class benchmark, with a focus on offline, corpus-level analysis rather than real-time deployment. After unifying all recordings into 5 s epochs at 256 Hz, 6,520 trials with 60 EEG channels were preprocessed using bandpass filtering, baseline correction, z-score normalization, and trial-wise ICA for artifact attenuation. On top of this pipeline, a comprehensive feature representation was constructed that combines common spatial patterns, discrete wavelet statistics, time-domain moments, autocorrelation coefficients, power spectral density band powers, and Hjorth parameters into a single 5,120-dimensional vector. A block-wise ablation indicates that autocorrelation, CSP, PSD, and Hjorth features carry most of the discriminative information in this setting, while wavelet and simple statistical descriptors contribute little and can be removed without harming performance. Using only the informative blocks (3,440 features), a multinomial logistic regression classifier reaches about 0.41 accuracy and 0.42 macro F1 on the ten-class task, roughly four times chance level. A multi-layer perceptron and a CNN–LSTM model, trained under the same splits and with class weighting, do not outperform this linear baseline and exhibit stronger overfitting. Within the evaluated protocol, these findings suggest that carefully engineered features capture most of the discriminative structure accessible on this corpus, and that deeper models add complexity without clear benefit. The study provides a transparent baseline and a feature-level analysis that can serve as a reference point for future work on imagined speech recognition and transfer learning across EEG corpora.**

Keywords*—EEG-based imagined speech; ASU imagined speech dataset; brain–computer interface; feature extraction; logistic regression

I. Introduction

Electroencephalography (EEG) based imagined speech recognition has emerged as a promising route toward silent speech interfaces for users who cannot rely on the vocal tract, for example people with advanced neuromuscular disease or severe dysarthria. By decoding internal speech representations directly from cortical activity, such systems aim to restore basic communication with minimal motor effort and without external articulation. Recent surveys on silent speech interfaces and EEG-based brain computer interfaces consistently highlight imagined speech as one of the most challenging yet societally important application domains [1], [2].

Despite rapid advances in deep learning, the performance of noninvasive imagined speech decoders remains modest when evaluated in realistic conditions. Most published studies operate on highly constrained tasks, such as binary vowel classification or discrimination between a few short words, and often use subject-specific models with carefully tuned pipelines. Reported accuracies in these settings typically lie around 70 to 80 percent, but they tend to fall toward 20 to 50 percent once the vocabulary is expanded, the class structure becomes more heterogeneous, or evaluation protocols adopt stricter cross-validation schemes [3], [4]. This gap between optimistic benchmark numbers and robust multiclass performance has motivated calls for more systematic baselines and ablation studies that clarify what can realistically be achieved with current sensors, preprocessing strategies, and feature representations.

The Arizona State University (ASU) imagined speech corpus is one of the few publicly available datasets that support such analysis [5]. It contains multi-session EEG recordings of short words, longer phrases and isolated vowels, acquired at 256 Hz with a high-density montage and segmented around the final auditory cue [6]. Compared with smaller imagined speech datasets, ASU introduces several sources of difficulty that are frequently encountered in real use: multiple lexical categories, variable trial lengths, and imbalanced class distributions that reflect differences in the number of repetitions per condition. Previous work on ASU has mainly focused on end-to-end architectures that combine time frequency representations with convolutional or recurrent neural networks, usually reporting single model accuracies without detailed inspection of the underlying feature space [7].

In parallel, the broader EEG community has developed a rich toolbox of handcrafted feature extractors, such as common spatial patterns, wavelet coefficients, spectral band powers and Hjorth parameters, that have proved effective in motor imagery

*Corresponding authors

and other BCI paradigms [2]. However, their behaviour in the context of imagined speech is less well documented. Many studies adopt a fixed subset of features without quantifying how each block contributes to classification, whether different descriptors are redundant or complementary, or how simple linear classifiers compare with more elaborate deep architectures when both are applied to the same feature space. This lack of systematic feature-level analysis makes it difficult to interpret negative results, to compare methods across papers, or to design compact models that might eventually run in real time on portable hardware.

The present work addresses these gaps through a deliberately transparent study of EEG-based imagined speech recognition on the ASU dataset. A reproducible preprocessing pipeline is first constructed that performs channel selection, bandpass filtering, baseline correction, trial length normalization and artifact removal using independent component analysis. The cleaned signals are then transformed into a large unified feature set... On top of this representation, several strong but interpretable classifiers are evaluated, including multinomial logistic regression, linear support vector machines, multilayer perceptrons and a lightweight CNN LSTM network, using consistent train validation test splits and macro-averaged metrics to account for label imbalance. The intended deployment scenario is explicitly offline: the goal is to characterise what handcrafted EEG features can achieve on the ASU corpus under a fixed evaluation protocol, rather than to prototype a real-time BCI. The resulting baselines are framed as a diagnostic reference that any subsequent real-time system, with its tighter latency and subject-adaptation constraints, would have to match or exceed.

Beyond reporting headline accuracies, particular emphasis is placed on understanding which feature families actually support discrimination in this ten-class imagined speech task. A leave-one-block-out analysis and reduced feature experiments quantify the marginal contribution of each feature group and show, perhaps counterintuitively, that a smaller subset that combines spatial filters, autocorrelation, band power, and Hjorth parameters can match or slightly improve the performance of the full feature set. These results provide a nuanced picture: overall macro F1 scores remain in the range of 0.40 to 0.44, which confirms the intrinsic difficulty of the problem on this dataset, yet the systematic ablations reveal meaningful structure that can inform subsequent model design.

In summary, this study makes three main contributions. First, it offers a carefully documented preprocessing and feature extraction pipeline for the ASU imagined speech corpus that other groups can reuse as a common baseline. Second, it provides a comparative evaluation of classical and deep classifiers on a unified feature space, highlighting where additional model complexity does and does not translate into improved generalization. Third, it delivers a detailed feature-level analysis that clarifies which descriptors carry the most useful information for ten-class imagined speech recognition. Taken together, these contributions position transparent baselines as an essential complement to more ambitious end-to-end architectures, helping the field move toward more realistic expectations and more robust future systems.

The remainder of this study is organised as follows: Section II reviews related work on EEG-based imagined speech recognition, with particular attention to the ASU corpus and to feature engineering practices in non-invasive brain–computer interfaces. Section III describes the proposed methodology, including the preprocessing pipeline, the multi block feature extraction scheme, and the classification models and evaluation protocols. Section IV presents the experimental results on the ten-class ASU task, reports the feature ablation and model comparison findings, and discusses their implications for future imagined speech systems. Finally, Section V summarises the main conclusions, highlights the limitations of the current study, and outlines several directions for extending the work through cross-dataset transfer learning and more compact real-time implementations.

II. Related Works

Research on EEG-based imagined speech recognition has expanded notably in the last five years, but performance remains far below that of motor imagery or overt speech decoding. Recent systematic reviews show that most EEG speech imagery systems are evaluated on small, highly constrained vocabularies and often under subject-dependent protocols, where models are trained and tested on the same individual. Under such conditions, reported accuracies can exceed 80 %, yet they typically fall to the 40–60 % range when the vocabulary grows, when multi-class settings are used, or when cross-subject generalisation is considered [8]. This gap between optimistic, laboratory-scale results and more realistic, multi-class performance is central to the present work.

A first line of recent studies still relies heavily on hand-crafted features combined with relatively shallow classifiers. Before surveying individual contributions, it is useful to fix the three axes along which they differ most and that determine whether a reported number is directly comparable to another: 1) the scale and paradigm of the dataset, including the number of participants, the number of imagined classes, and whether the corpus uses short vowels, isolated words, phrases or continuous material; 2) the evaluation protocol, spanning subject-dependent versus cross-subject splits, stratified trial-level versus session-level partitioning, and whether macro-averaged metrics are used in the presence of class imbalance; and 3) the model class, ranging from linear projections on engineered features to convolutional, recurrent or hybrid deep architectures operating on raw or minimally processed EEG. Headline accuracies that look superficially similar can correspond to very different tasks along these axes, and most of the apparent disagreement in the literature dissolves once studies are stacked against this grid. The studies surveyed below are ordered accordingly, moving from small-vocabulary subject-dependent pipelines with shallow classifiers toward multi-class, cross-subject deep models, and are referenced back to this frame whenever they are introduced. These works exploit time-frequency decompositions, power spectral descriptors, and spatial projections such as common spatial patterns in order to enhance signal-to-noise ratio before classification. For example, deep metric learning has been used to map EEG trials for five vowels and six words into a latent space where distances reflect phonetic

similarity, achieving around 45 % accuracy in a six-word imagined speech task despite strong inter-subject variability [9]. Other studies report comparable performance when combining wavelet-based features with traditional machine learning models, particularly when the number of commands is limited, and the recording protocol is tightly controlled [6]. These results suggest that carefully engineered features can extract some discriminative structure from EEG speech imagery, but they do not fully resolve the problem of class overlap in larger label spaces.

More recent work has shifted towards deep learning architectures that operate directly on raw or minimally processed EEG. Abdulghani et al. used wavelet scattering representations of 8-channel EEG and an LSTM network to classify four imagined directional commands, reporting an accuracy of about 92.5 % under a subject-specific setting, which illustrates how recurrent models can capture temporal dynamics when the task is well constrained [10]. In contrast, when the vocabulary increases and multi-class decoding is attempted, accuracy tends to decline. Alharbi et al. proposed hybrid 3D CNN-StackLSTM and 3D CNN-BiLSTM models that operate on sequences of EEG topographic maps for five imagined words in the BCI2020 dataset. Their best model reached an average accuracy of roughly 45 %, with substantial variability across subjects, even though the architecture explicitly encodes both spatial and temporal information [11]. Similar hybrid CNN-BiLSTM systems, often combined with wavelet-based front-ends or metaheuristic feature selection, have shown strong single-subject performance but less convincing cross-subject generalisation [12].

Another strand of work explores transfer learning and representation learning to mitigate the scarcity of labelled imagined speech data. Lee et al. applied deep metric learning to jointly optimise a feature space for vowels and words, enabling reuse of representations across tasks and datasets and achieving mid-40 % accuracy for six-class problems [9]. Other authors have investigated transfer learning from overt speech or from related EEG paradigms. One recent study proposed a transfer learning framework that jointly extracts temporal and spectral features, then fine-tunes a classifier on limited imagined speech trials, showing consistent but modest gains over training from scratch [13], [14]. These approaches underline that cross-task and cross-corpus transfer is feasible, yet the improvements are incremental and far from closing the gap to practical communication rates.

The limitations of current performance are closely tied to data availability. For many years, researchers relied on small institutional datasets with a handful of commands or vowels, which restricted the diversity of linguistic material and impeded robust evaluation. The recent release of the Chinese Imagined Speech Corpus (Chisco), which contains more than 20,000 sentences of high-density EEG per subject, is a major step forward and offers a much richer testbed for neural language decoding from EEG [15]. However, most published work with Chisco and similar corpora is still at an early stage and often focuses on feasibility demonstrations rather than systematic baselines across feature families and model classes [8].

Against this backdrop, the present study takes a deliberately conservative and diagnostic perspective. Instead of aiming for a complex, highly tuned end-to-end model, the analysis is performed on a medium-scale imagined speech dataset (ASU) that contains ten classes and 6,520 trials, using a rich but interpretable feature space that combines common spatial patterns, discrete wavelet features, statistical descriptors, autocorrelation, band-power estimates, and Hjorth parameters. The discriminative contribution of each feature family is then quantified by analysing block-wise performance with linear models and by comparing these baselines to compact neural architectures such as multilayer perceptrons and CNN-LSTM hybrids. The results, which cluster around 40–44 % macro F1 for ten-class classification, are consistent with the more demanding multi-class findings in the recent literature and provide a transparent reference point for future work on transfer learning and model compression in EEG-based imagined speech recognition.

III. Materials and Methods

This section describes the complete processing pipeline used in this study, from raw electroencephalographic (EEG) recordings to the classification of imagined speech commands. The workflow is organised into three tiers. Tier 1 covers data acquisition from the ASU imagined speech corpus and the preprocessing applied to the 60-channel EEG signals, including filtering, baseline correction, artefact rejection, and quality control. Tier 2 details the feature engineering stage, in which six complementary feature blocks are extracted and combined into a reduced 3,440-dimensional representation. Tier 3 introduces the discriminative models and evaluation procedures that operate on this compact feature space. An overview of these stages and their dependencies is summarised in Fig. 1.

A. Dataset and Labelling

The ASU imagined speech dataset was collected by the Human Oriented Robotics and Controls (HORC) laboratory at Arizona State University using a 64-channel BrainProducts ActiCHamp amplifier. Recordings were acquired from 15 healthy volunteers (11 males and 4 females), who were instructed to silently rehearse visually and auditorily cued prompts without overt articulation or muscle movements. The vocabulary comprised four groups of stimuli: short commands (“in”, “out”, “up”), long words (“cooperate”, “independent”), short–long combinations (“in”, “cooperate”), and sustained vowels (/a/, /i/, /u/). Each trial followed a structured cueing protocol that aligned the onset of imagined speech relative to a sequence of auditory beeps, a design that is now standard in imagined speech paradigms [5], [16].

This study focused on the "last-beep" interval, that is, the segment immediately preceding the final cue, which has been reported to concentrate the most stable imagined speech activity while limiting contamination from anticipatory and post-movement processes [16]. After aggregating all usable files and retaining only last-beep epochs, a total of $N = 6520$ trials distributed across ten imagined classes. These classes correspond to three short commands (“in”, “up”, “out”), two long words (“cooperate”, “independent”), two short–long phrase conditions (“cooperate”, “in”), and three vowels (“a”, “i”, “u”). For supervised learning, each semantic class was mapped to an

integer label $y \in \{0, \ldots, 9\}$, following a consistent mapping used throughout the subsequent analysis and model training.

The original montage contained dedicated electrooculography (EOG) channels that are useful for artefact detection but not for decoding imagined speech. For each trial, the designated EOG indices (0, 9, 32, and 63 in the original layout) were removed, and the first $C = 60$ EEG channels as the working montage. All signals were sampled at $f_s = 256$ Hz, and trial lengths varied slightly across recordings because of minor differences in cue timing and file boundaries. Before feature extraction and modelling, these variable-length epochs were transformed into a fixed-size tensor through the preprocessing pipeline described in Section III.B. The chosen epoch length of approximately 5 s corresponds to the full last-beep interval defined by the ASU acquisition protocol and used in previous ASU-based analyses; it encompasses the imagined articulation itself together with a short post-cue window during which sustained speech-motor activity has been reported. Shorter windows would truncate this sustained activity and eliminate the lag range exploited by the autocorrelation and Hjorth descriptors introduced later, whereas substantially longer windows would reintroduce pre-cue anticipation and post-trial drift that the last-beep alignment was designed to avoid.

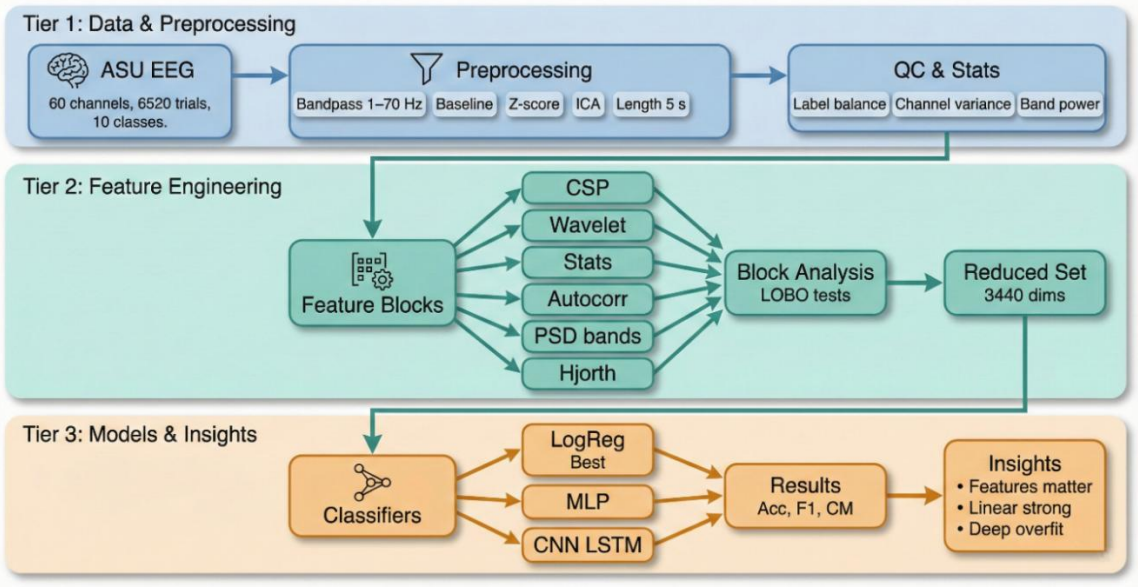

Fig. 1. Overview of the proposed EEG imagined speech recognition pipeline on the ASU dataset

B. Preprocessing Pipeline

All preprocessing was implemented in Python using MNE-Python together with standard scientific libraries (NumPy, SciPy, scikit-learn). The design goal was to obtain artefact-reduced and temporally aligned trials while preserving the spectral content that is typically informative for imagined speech and BCI tasks [17]. The same sequence of operations was applied to every trial: temporal alignment by zero-padding or truncation, band-pass filtering, baseline correction, trial-wise z-score normalization, and independent component analysis (ICA) for artefact attenuation. Fig. 2 displays the raw versus the preprocessed data for the dataset using our pipeline.

Let $x_c(t)$ denote the raw signal from channel $c \in \{1, \ldots, C\}$ at discrete time index t, with trial-specific length T. First, a common temporal length $T_{\max}$, defined as the maximum number of samples across all trials in the dataset. For trials shorter than $T_{\max}$ we applied zero-padding at the end, and for trials longer than $T_{\max}$ we truncated the surplus samples:

$$\bar{x}_c(t) = \begin{cases} x_c(t), & 0 \le t < T, \\ 0, & T \le t < T_{max}, \end{cases} \quad (1)$$

This operation produces a temporally aligned representation without interpolation, which helps preserve the phase relationships of oscillatory activity that are often exploited in EEG decoding [18]. At this stage, each trial has shape $C \times T_{\max}$.

To suppress slow drifts and high-frequency sensor noise while retaining the classical delta to low-gamma range relevant for cognitive and speech-related processes, a zero-phase fifth-order Butterworth band-pass filter with passband $[f_{\text{low}}, f_{\text{high}}] = [1,70]$ Hz. The filter was implemented using forward-backward IIR filtering (filtfilt) to avoid phase distortion [19]. For each channel, the band-passed signal is:

$$y_c(t) = \mathcal{F}_{\text{BP}}\left(\bar{x}_c(t); f_{\lambda o \omega} = 1 \text{ Hz}, f_{\eta \iota \gamma \eta} = 70 \text{ Hz}\right), \quad (2)$$

where, F_{BP} denotes the zero-phase band-pass operator applied along the time axis.

Slow baseline shifts were then removed at the trial level. For the ASU dataset, a dedicated baseline interval of 200 ms was available at the beginning of each epoch. Let B be the set of baseline time indices, corresponding to [0,0.2]s at 256 Hz. The baseline mean for the channel c is:

$$\mu_c^{\text{base}} = \frac{1}{|B|} \sum_{t \in B} y_c(t). \quad (3)$$

The baseline-corrected signal is then:

$$z_c(t) = y_c(t) - \mu_c^{\beta \alpha \sigma \varepsilon}. \quad (4)$$

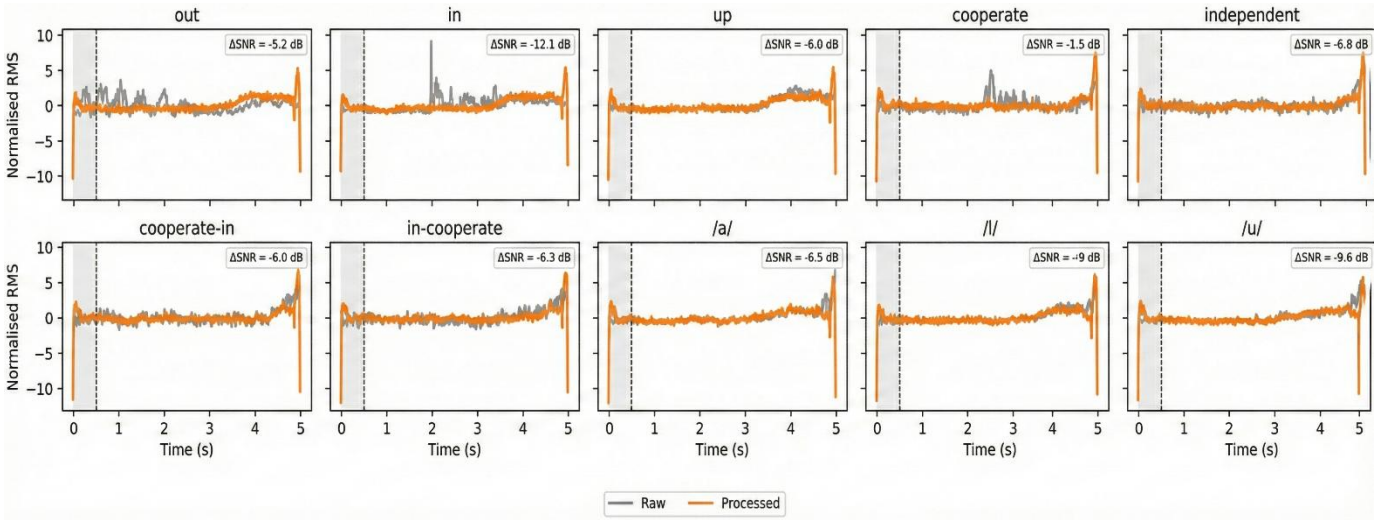

Fig. 2. Raw versus processed classes of the ASU dataset.

When a dedicated baseline window is not explicitly defined, the first 5% of the trial can be used as an approximate baseline, which is a common practical choice in continuous EEG paradigms [20].

To stabilise subsequent covariance-based methods and ensure that channels contribute on comparable scales, trial-wise z-score normalization was performed per channel. For each channel c, the mean and standard deviation of the baseline-corrected signal were computed over the entire trial,

$$\mu_c = \frac{1}{T_{max}} \sum_{t=0}^{T_{max}-1} z_c(t), \tag{5}$$

$$\sigma_c = \sqrt{\frac{1}{T_{max}} \sum_{t=0}^{T_{max}-1} (z_c(t) - \mu_c)^2}, \tag{6}$$

and obtained the normalized time series:

$$\hat{x}_c(t) = \frac{z_c(t) - \mu_c}{\sigma_c + \varepsilon} \tag{7}$$

where, ε is a small constant used for numerical stability. This per-trial normalization mimics the standardisation strategies widely adopted in deep EEG models and large-scale BCI datasets [21], [22].

Residual eye movements and muscle artefacts were attenuated using ICA. For each trial, an MNE RawArray was constructed with $C = 60$ channels and sampling rate $f_s = 256\ Hz$, and estimated an ICA decomposition with as many components as channels. Let $X \in \mathbb{R}^{C \times T_{\max}}$ denote the normalized multichannel signal. ICA models X as a linear mixture of statistically independent sources:

$$X = AS \tag{8}$$

where, $A \in \mathbb{R}^{C \times C}$ is the mixing matrix and $S \in \mathbb{R}^{C \times T_{max}}$ contains the independent components. Components whose time courses and topographies resembled canonical ocular or myogenic artefacts were excluded heuristically, following established guidelines for EEG artefact handling [23]. The cleaned signal is then reconstructed as:

$$\hat{X} = A\hat{S} \tag{9}$$

where, $\hat{S}$ equals S with the identified artefact components set to zero. ICA-based correction remains one of the most effective approaches for denoising EEG in both clinical and BCI contexts [24].

After preprocessing, each trial is represented by a cleaned tensor $\hat{X} \in \mathbb{R}^{C \times T_{\max}}$. Stacking all $N = 6520$ trials yields the final preprocessed dataset:

$$\hat{\mathcal{X}} \in \mathbb{R}^{N \times C \times T_{max}}, \tag{11}$$

which serves as the input to the subsequent feature extraction and classification stages described in the following subsections.

C. Feature Extraction

We designed a unified feature extraction block that compresses each preprocessed trial into a 5120-dimensional vector. The block combines six complementary descriptors: spatial filters from Common Spatial Patterns (CSP), time frequency statistics from a discrete wavelet transform, low-order time domain moments, lagged autocorrelation, band power from power spectral density (PSD), and Hjorth mobility and complexity. Similar hybrid designs have been shown to outperform single-family features in motor imagery and speech imagery BCIs, since they capture both transient and sustained dynamics across multiple scales. The resulting representation is deliberately broad rather than minimal: the intent is to expose redundancy between feature families so that the block-wise ablation in Section IV-B can quantify which descriptors genuinely contribute and which can be removed. All features were computed per trial and concatenated, then normalised once at the end using a single StandardScaler fitted on the training partition, avoiding information leakage into the test set. The trade-off between feature breadth and redundancy is examined directly in the ablation results, and the reduced 3,440-dimensional subset is presented there as the operational representation used by the classifiers [25].

1) Common spatial patterns: To capture spatially discriminative activity across electrodes, CSP was applied in a one versus rest setting for each imagined class. Let denote a single trial with channels and time samples, and let and be the

class averaged covariance matrices for the target class and the pooled background, respectively. For each binary problem CSP solves the generalised eigenvalue problem $X \in \mathbb{R}^{C\times T}$ $C T \Sigma_1 \Sigma_2$:

$$\Sigma_1 \mathbf{w}_k = \lambda_k \Sigma_2 \mathbf{w}_k, \quad (12)$$

where, $\mathbf{w}_k$ is the k-th spatial filter and λ_k the associated eigenvalue. Projecting the trial as $\mathbf{z}_k = \mathbf{w}_k^T X$ and computing the log-normalised variance:

$$f_k^{\mathrm{CSP}} = \log \frac{\mathrm{var}(\mathbf{z}_k)}{\sum_j \mathrm{var}(\mathbf{z}_j)} \quad (13)$$

yields CSP features that emphasise spatial patterns with maximal class-specific variance [26]. The implementation allocates n_{comp} filters per class, chosen adaptively from the number of channels and classes, which in the ASU configuration results in 60 CSP features per trial.

2) Wavelet-based time frequency statistics: Fine-grained time frequency structure was represented using a Level 3 Discrete Wavelet Transform (DWT) with a Daubechies 4 Mother Wavelet. Db4 is frequently adopted in EEG work because it provides compact support and smooth oscillatory atoms that align well with dominant EEG rhythms in the alpha and beta ranges [6]. For each channel, the preprocessed time series is decomposed into one approximation and three detail coefficient vectors. For each coefficient vector of length six, statistics are computed: $c, x_c(t)\{\mathbf{c}_\ell\}_{\ell=0}^{3} n$

$$\mu = \frac{1}{n}\sum_{i=1}^{n} c_i, \quad (14)$$

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n} (c_i - \mu)^2 \quad (15)$$

$$\mathrm{skew} = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{c_i-\mu}{\sigma}\right)^3, \quad (16)$$

$$\mathrm{kurt} = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{c_i-\mu}{\sigma}\right)^4, \quad (17)$$

along with the minimum and maximum of c. Concatenating these statistics across all levels and 60 channels yields a 1440-dimensional DWT feature block that captures transient, band-limited changes associated with imagined speech cues [27].

3) Time domain statistics: To summarise coarse distributional properties in the time domain, four low-order moments are computed directly from the baseline corrected and normalised signal. For each channel, we estimate the mean, standard deviation, skewness, and kurtosis over time and concatenate them across channels. The resulting 240-dimensional block provides a compact characterisation of overall amplitude, variability, and waveform asymmetry. Although simple, such statistics have been reported to add complementary information when combined with more structured features in EEG classification tasks [28] $cx_c(t)$.

4) Autocorrelation features: Imagined speech tends to recruit sustained, quasi-rhythmic activity that extends over several hundred milliseconds, and articulatory-motor planning signals have been reported to unfold over comparable timescales in overt and covert speech paradigms. To capture these temporal dependencies, the one-sided autocorrelation is computed for each channel up to a maximum lag samples, which corresponds to approximately 0.2 s at 256 Hz. This horizon is short enough to avoid washing out the moment-to-moment dynamics of a 5 s trial, yet long enough to span the dominant alpha/beta cycles (roughly 50–400 Ms) that are implicated in speech-motor and sensorimotor integration, and it is comparable to the autocorrelation ranges adopted in earlier imagined-speech and motor-imagery feature studies. For channel, $\tau_{\max} = 50c$:

$$r_c(\tau) = \sum_{t=0}^{T_{\max}-1-\tau} x_c(t)x_c(t+\tau), \tau = 0, \ldots, \tau_{\max} \quad (18)$$

The first 50 autocorrelation coefficients per channel are concatenated to form a 3000-dimensional block across all 60 channels. These features encode the decay and periodicity of channel-wise dynamics without imposing stationarity assumptions and have been shown to benefit EEG-based decoding of cognitive and motor states [29].

5) Power spectral density and band power: Frequency-specific information is extracted using Welch's Method for PSD Estimation. Each channel is segmented into overlapping windows, tapered with a Hamming window, and the periodograms are averaged to obtain. The PSD is then integrated within five canonical bands $PSD_c(f)$:

$$P_c^{\theta} = \sum_{f\in[4,8)} PSD_c(f), \quad (19)$$

$$P_c^{\beta} = \sum_{f\in[13,30)} PSD_c(f), \quad (20)$$

$$P_c^{\delta} = \sum_{f\in[0.5,4)} PSD_c(f), \quad (21)$$

$$P_c^{\alpha} = \sum_{f\in[8,13)} PSD_c(f), \quad (22)$$

$$P_c^{\gamma} = \sum_{f\geq 30} PSD_c(f), \quad (23)$$

yielding five band powers per channel and 300 PSD-based features per trial. Band power in these ranges remains a workhorse descriptor in motor imagery and speech imagery BCIs and offers a robust, interpretable link to underlying oscillatory processes [30].

6) Hjorth parameters: Finally, Hjorth mobility and complexity parameters are computed, which summarise the spectral content of a signal using simple variance ratios in the time domain. For channel, let denote the time series, its first discrete derivative, and its second derivative. Mobility and complexity are defined as $cx_c(t)\dot{x}_c(t)\ddot{x}_c(t)$:

$$\mathrm{mobility}_c = \sqrt{\frac{\mathrm{var}(\dot{x}_c(t))}{\mathrm{var}(x_c(t))}}, \quad (24)$$

$$\mathrm{complexity}_c = \sqrt{\frac{\mathrm{var}(\ddot{x}_c(t))}{\mathrm{var}(\dot{x}_c(t))}} / \mathrm{mobility}_c. \quad (25)$$

These quantities reflect the dominant frequency and the change in frequency over time and have been repeatedly validated as efficient EEG features for BCI and clinical applications [31]. Concatenating mobility and complexity

across the 60 channels produces an 80-dimensional Hjorth feature block.

7) Feature normalisation and reduced subset: All six blocks are concatenated to form a feature matrix, where trials. A single StandardScaler is then fitted on to obtain normalised features with zero mean and unit variance per dimension, which are used by both linear and deep models. To reduce redundancy and computational cost, we also define a reduced subset by retaining only the CSP, Autocorrelation, PSD and Hjorth Blocks, resulting in a 3440-dimensional representation. As discussed later in the results, this subset preserves most of the discriminative structure in a simple Multinomial Logistic Regression Analysis, while substantially lowering the input dimensionality for the neural classifiers. $F \in \mathbb{R}^{N\times 5120} N = 6520 F\tilde{F}$.

D. Model Architectures and Training Procedure

All classification experiments on the ASU imagined speech dataset were carried out on the reduced feature representation described in Section III-C. Each trial is represented by a 3440-dimensional vector $\mathbf{x} \in \mathbb{R}^{3440}$, obtained by concatenating spatial (CSP), temporal (autocorrelation), spectral (PSD), and dynamical (Hjorth) descriptors from the preprocessed EEG. The ASU corpus contributes 6520 labelled trials with ten classes ($y \in \{0, \ldots, 9\}$), corresponding to short, long, short-long, and vowel commands. This setting allows us to decouple model behaviour from low-level signal processing and to focus on how different architectures exploit the same feature space, in line with recent work on deep learning for EEG-based BCIs.

1) Baseline and deep architecture: Three discriminative models were considered: a Multinomial Logistic Regression Baseline, a Multilayer Perceptron (MLP), and a Deeper CNN-BiLSTM-FCNN Hybrid. The motivation for this choice was comparative rather than architectural optimisation. The linear model establishes an interpretable reference in the same feature space. The MLP tests whether a modest amount of non-linear recombination of the engineered features improves discrimination. The CNN-BiLSTM-FCNN is included as a deliberate stress test: although the 3,440-dimensional feature vector is a heterogeneous concatenation of spatial, spectral and temporal descriptors and therefore lacks the strict local ordering that convolutions typically exploit in raw EEG decoding, this architecture has been reported to improve performance on pre-engineered EEG features in several imagined-speech studies. Including it on the same representation allows an empirical test of whether any residual sequential structure along the feature axis can be recovered by a deeper sequence model, and if so, at what computational cost. Alternative architectures that are better suited to tabular inputs – for instance, gradient-boosted ensembles or attention-based tabular encoders – are considered promising directions for follow-up work but are not evaluated here, so as to preserve a controlled three-way comparison on a common feature space. All three implement the same probabilistic mapping via a softmax output layer, but differ in how they transform the input features. The detailed architectures of the three classifiers are illustrated in Fig. 3: $p_\theta(y \mid \mathbf{x})$.

For the logistic regression baseline, the conditional class probabilities are given by:

$$p_\theta(y = k \mid \mathbf{x}) = \frac{\exp\left(\mathrm{w}_k^\top \mathrm{x} + b_k\right)}{\sum_{j=0}^{C-1} \exp\left(\mathrm{w}_j^\top \mathrm{x} + b_j\right)}, k = 0, \ldots, C-1, \quad (26)$$

with $C = 10$ and $\theta = \{\mathrm{w}_k, b_k\}_{k=0}^{C-1}$. This model provides a well-understood linear baseline to contextualise the gains from non-linear feature transformations.

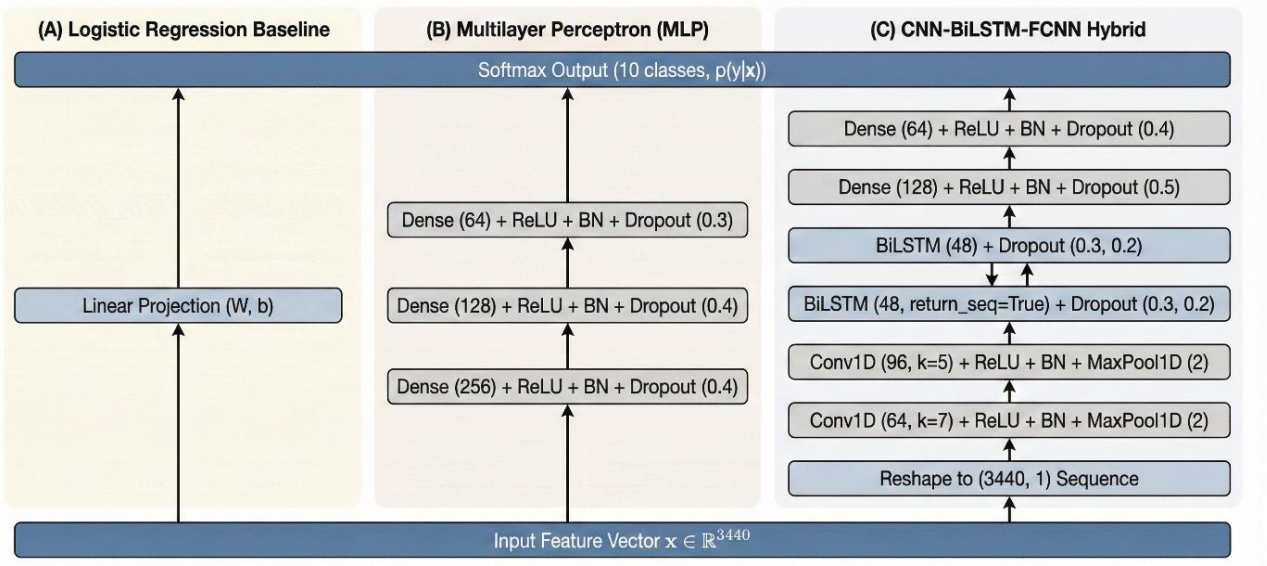

Fig. 3. Model architectures used in the ASU experiments: (A) Logistic regression baseline, (B) Multilayer perceptron with three hidden layers, and (C) CNN–BiLSTM–FCNN hybrid operating on the 3 440-dimensional feature sequence.

The improved MLP operates directly on x and stacks three fully connected hidden layers with batch normalisation and dropout. If $\mathrm{h}^{(0)} = \mathrm{x}$, each hidden layer l computes

$$\mathrm{h}^{(l)} = \phi\left(\mathrm{W}^{(l)}\mathrm{h}^{(l-1)} + \mathrm{b}^{(l)}\right), \tag{27}$$

Where, $\phi(\cdot)$ is the rectified linear unit (ReLU). The final softmax layer produces $\hat{\mathbf{y}} = \text{softmax}\left(\mathrm{W}^{(L)}\mathrm{h}^{(L-1)} + \mathrm{b}^{(L)}\right)$. All dense layers are regularised with an ℓ_2 penalty on the weights to reduce overfitting in this relatively high-dimensional space. Such compact MLPs have been found competitive on engineered EEG features when the sample size is modest.

The CNN-BiLSTM-FCNN model treats each feature vector as a 1D sequence of length 3440 with a single channel and therefore seeks local patterns along the feature axis. Two 1D convolutional blocks first extract local motifs:

$$\mathbf{z}_t^{(c)} = \phi\left(\sum_j \sum_\tau w_{j,\tau}^{(c)} x_{j,t+\tau} + b^{(c)}\right), \tag{28}$$

followed by batch normalisation and max pooling to build increasingly abstract representations. The resulting sequence of feature maps is fed to stacked bidirectional LSTM layers that capture longer range dependencies by processing the sequence in both forward and backward directions. Bidirectional recurrent architectures are known to improve sequence modelling in speech and EEG applications by exploiting past and future context. The final temporal embedding is passed through a small fully connected head with ReLU, batch normalisation, and dropout before the softmax classifier.

Table I summarises the main architectural choices. All hyperparameters are fixed across experiments on the ASU dataset in order to attribute performance differences to the model families rather than to extensive tuning.

TABLE I. TRAINING AND OPTIMIZATION SETTINGS FOR THE THREE CLASSIFIERS ON THE REDUCED ASU FEATURE SET

Model	Input shape	Optimizer	Initial LR	Max epochs	Batch size	Early stopping	LR schedule	Regularisation/class weighting
Logistic regression	(3440)	L2-regularised multinomial solver	–	–	–	–	–	ℓ_2 penalty on **W** class_weight = balanced
MLP (deep baseline)	(3440)	Adam	3×10−4	150	64	val_loss, 25	factor 0.5, patience 8, min_lr 10^{-5}	ℓ_2 on all dense layers, batch norm, dropout 0.4–0.3
CNN-BiLSTM-FCNN hybrid	(3440,1)	Adam	3×10−4	120	64	val_loss, 20	factor 0.5, patience 6, min_lr 10^{-5}	ℓ_2 on FC layers, batch norm, dropout 0.5–0.4, class_weight = balanced

2) Loss function, class weighting, and optimization: All models are trained with a weighted categorical cross-entropy loss. Given a training set, the objective is $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$:

$$\mathcal{L}(\theta) = -\frac{1}{N}\sum_{i=1}^N w_{y_i} \log\, p_\theta(y_i \mid \mathbf{x}_i), \tag{29}$$

where, w_{y_i} is the class weight associated with label y_i. Class weights are computed inversely proportional to class frequencies in the training partition, which mitigates the moderate label imbalance observed in the ASU corpus and avoids the need for synthetic oversampling. Similar class reweighting strategies are standard practice in EEG classification when some commands appear less frequently than others.

As illustrated in Table I, all neural models use the Adam optimiser with an initial learning rate of 3×10^{-4}. Training proceeds for a maximum of 150 epochs for the MLP and 120 epochs for the CNN-BiLSTM-FCNN, using mini-batches of size 64. To reduce manual tuning, a Reduce-on-Plateau scheduler halves the learning rate whenever the validation loss fails to improve for several epochs (8 epochs for the MLP, 6 for the CNN-BiLSTM-FCNN), with a floor at 10^{-5}. Early stopping with patience of 25 epochs (MLP) and 20 epochs (CNN-BiLSTM-FCNN) restores the weights that achieve the lowest validation loss, which helps to prevent overfitting and stabilises the reported results.

3) Data partitioning and evaluation metrics: For each model, the 6520 ASU Trials are split with stratification into training, validation, and held-out test data. Stratification preserves the empirical label distribution across splits and ensures that rare classes are represented in both validation and test sets. No subject-specific information is used in this stage, since the goal of this experiment is to characterise how well different architectures exploit the engineered feature space rather than to probe cross-subject generalisation $60\%, 20\%, 20\%$.

The data were split at the trial level into train, validation, and test partitions of 60 %, 20 %, and 20 %, respectively, preserving the mild class imbalance described in Section III. To report performance in a way that is robust to this imbalance, overall accuracy and macro-averaged F1 are the primary metrics of interest. All models were trained with a single fixed random seed to keep the three-way comparison on strictly identical splits; the resulting point estimates should therefore be read as single-run performance on the held-out test set rather than as means of a sampling distribution. A formal repeated-runs analysis with confidence intervals over multiple seeds and resampled stratified splits is a natural next step and is discussed among the study's limitations. Accuracy is defined as:

$$\text{Acc} = \frac{1}{N}\sum_{i=1}^N \mathbf{1}\{\hat{y}_i = y_i\} \tag{30}$$

while macro-F1 averages per-class F1 scores,

$$F_1^{\text{macro}} = \frac{1}{K}\sum_{k=1}^K \frac{2\Pi\rho\varepsilon\chi\iota\sigma\iota\sigma\nu_k \text{Recall}_k}{\text{Precision}_k + \text{Recall}_k}, \tag{31}$$

with $K = 10$ classes. Macro-F1 is particularly informative for imagined speech where some commands are intrinsically harder to discriminate than others. This metric treats all classes equally, which is crucial in imagined speech tasks where misclassifying a rarely used command can be more problematic

than misclassifying a frequent one. Confusion matrices and full per class precision, recall, and F1 scores are additionally reported to reveal systematic confusions between short, long, and vowel commands and to support later qualitative analysis of error patterns.

Taken together, this design yields a controlled comparison between a linear baseline, a compact deep MLP, and a more expressive CNN–BiLSTM–FCNN architecture on exactly the same ASU feature representation.

IV. Results and Discussion

A. Overall Classification Performance

After feature engineering and block-wise reduction, the ASU imagined speech dataset yielded a compact representation of 6,520 trials in a 3,440-dimensional feature space with ten classes (three vowels, three phrase variants, and four directional commands). Table II summarises the test performance of the three classifiers evaluated on the reduced feature set.

The numbers are consistent with the bar plot in Fig. 4, which juxtaposes accuracy and macro-F1 for each model.

The first observation is that the linear logistic baseline already reaches a macro-F1 of about 0.42, essentially matching the deeper MLP. This confirms that once the EEG is projected into a carefully designed feature space, simple discriminative models can remain highly competitive, a pattern also reported for CSP-based motor imagery and imagined speech pipelines [32], [33]. The MLP does not significantly improve aggregate metrics, although it offers slightly more stable learning dynamics and provides a natural stepping stone toward later transfer learning stages in our broader programme.

The CNN–BiLSTM–FCNN hybrid, in contrast, underperforms slightly despite its far higher computational cost. For the MLP, both training and validation loss decrease rapidly during the first 30 epochs, after which validation loss plateaus while training loss continues to fall. Validation accuracy stabilises around 0.42, whereas training accuracy climbs above 0.60, indicating mild overfitting but no catastrophic divergence. A similar pattern is present, but more pronounced, for the CNN–BiLSTM model, whose validation loss flattens early while the network continues to specialise on the training set. These dynamics suggest that the reduced feature set does not provide enough independent evidence for the deeper temporal model to exploit, a finding in line with recent reports that end-to-end deep networks for non-invasive imagined speech often saturate near 30–40 % accuracy on vocabularies of 6–11 words [34], [35].

The confusion matrix in Fig. 5 further clarifies where errors arise. Vowel classes /a/, /i/, and /u/ show relatively high diagonal terms, with many trials correctly localised, while the directional commands “out”, “in”, and “up” are frequently confused with one another. The social-interaction phrases “cooperate”, “independent”, “cooperate-in”, and “in-cooperate” occupy an intermediate regime, with good separation between “cooperate-in” and the rest, but noticeable cross-talk between “cooperate” and “independent”. This structure mirrors previous findings that imagined vowels tend to elicit more stereotyped sensorimotor and auditory patterns than semantically richer phrases.

TABLE II. Test Accuracy, Macro-F1, and Qualitative Computational Cost for Three Classifiers Trained on the Reduced ASU Feature Set

Model	Test Accuracy (%)	Test macro-F1 (%)	Qualitative cost profile*
Logistic regression	41	42	Very low, single linear layer.
MLP (three-layer network)	41	42	Moderate, ≈ 60 epochs, fast per epoch.
CNN–BiLSTM–FCNN hybrid	40	38	High, long epochs and large parameter count.
*Cost profile is based on empirical training times and model size, not on formal complexity analysis.			

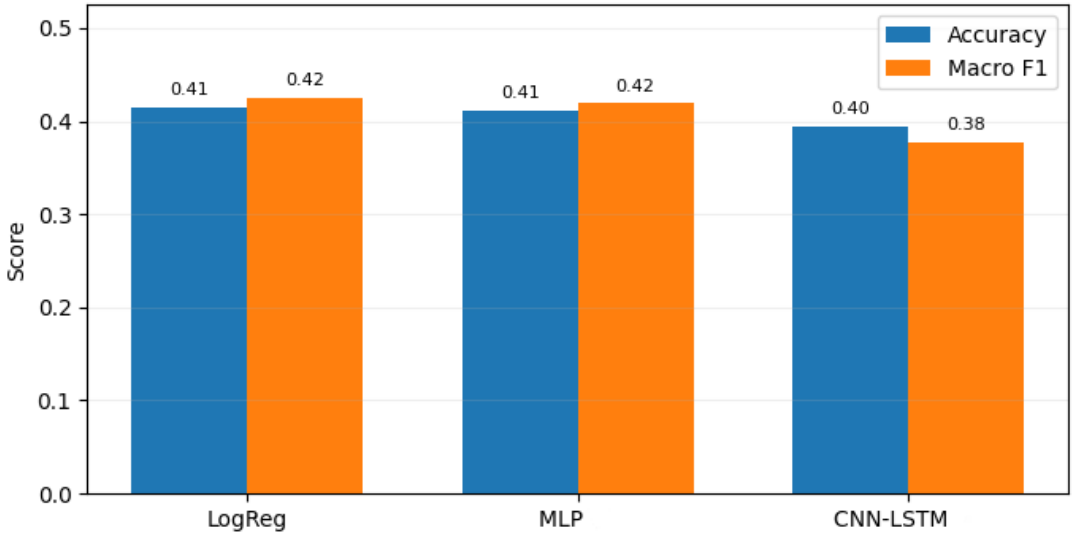

Fig. 4. Comparison of test accuracy and macro-F1 for logistic regression, MLP, and CNN–BiLSTM–FCNN on the reduced ASU features.

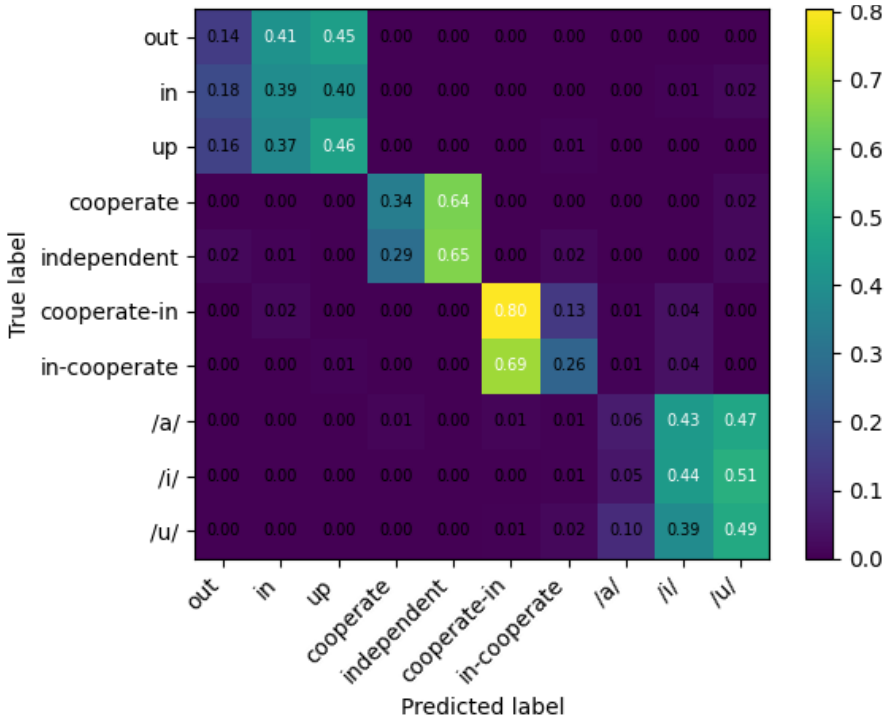

Fig. 5. Normalised confusion matrix for the CNN–BiLSTM–FCNN classifier on the ASU test set, showing systematic confusions among directional commands and vowels.

B. *Block-wise Contribution of Engineered Feature Families*

To understand how much each feature family contributes to the final performance, we conducted two complementary ablation studies on the logistic regression classifier: 1) single-block runs where the model receives only one family of features, and 2) leave-one-block-out (LOBO) runs where a given block is removed from the full feature set. The four blocks considered here are CSP spatial filters, autocorrelation statistics, band-power features from the power spectral density (PSD), and Hjorth parameters capturing activity, mobility, and complexity. These blocks have a long history in EEG decoding and are known to capture partially complementary aspects of oscillatory dynamics and temporal structure.

Table III summarises the macro-F1 obtained in the single-block and LOBO conditions, which correspond to the bar plots in Fig. 6.

TABLE III. BLOCK-WISE MACRO-F1 SCORES WITH ONE FEATURE FAMILY AND CHANGE IN MACRO-F1 WHEN EACH BLOCK IS REMOVED (LEAVE-ONE-BLOCK-OUT ANALYSIS).

Feature block	Macro-F1 with only this block	Δ macro-F1 when removed from the full set*
CSP	0.38	+0.015 (performance drops when removed)
Autocorrelation	0.42	−0.001
PSD	0.41	−0.003
Hjorth	0.40	−0.005

*Δ macro-F1 is defined as $F_1^{\text{all}} - F_1^{(-b)}$, where F_1^{all} is the macro-F1 with all blocks and $F_1^{(-b)}$ is the score with block b removed.

Several patterns emerge. First, when used in isolation, all four blocks produce macro-F1 scores in the range 0.38–0.42. Autocorrelation features are the strongest single source of discriminative information, slightly outperforming PSD and Hjorth statistics, which aligns with the intuition that imagined speech relies on intricate temporal dependencies rather than purely stationary band-power shifts. Second, the LOBO analysis reveals that CSP is the only block whose removal consistently harms the combined model: excluding CSP reduces macro-F1 by about 0.015, whereas excluding other blocks produces negligible or slightly positive changes at the third decimal place. In other words, CSP contributes a unique spatial structure that is not fully recoverable from the remaining blocks, even after dimensionality reduction, see Fig. 6.

These results suggest a division of labour across blocks. The temporal and spectral features (autocorrelation, PSD, and Hjorth) provide broadly redundant evidence about the underlying dynamics, which helps stabilise training but does not dramatically change the decision boundary when one block is removed. CSP, by contrast, contributes a smaller number of highly informative dimensions that appear to encode class-specific spatial patterns, consistent with its established role in motor imagery BCIs and more recent imagined-speech studies. From a design perspective, this analysis justifies keeping all four families in the reduced 3440-dimensional representation, but it also highlights that any future attempts at aggressive feature pruning should preserve CSP components and at least one of the temporal blocks to maintain robustness.

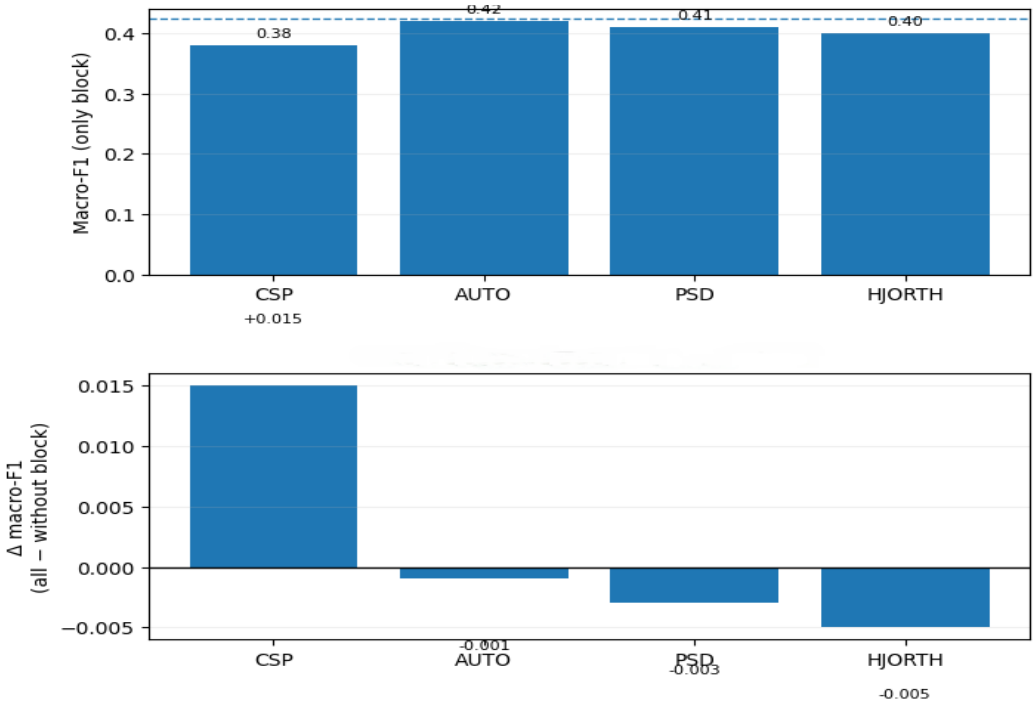

Fig. 6. Top: macro-F1 when each feature block (CSP, autocorrelation, PSD, Hjorth) is used in isolation. Bottom: change in macro-F1 when each block is removed from the full feature set.

C. Geometry of the Reduced Feature Space

To better understand why all three classifiers converge to similar performance levels, the geometry of the 3440-dimensional reduced feature space was examined. Fig. 7 visualises the test set using a two-dimensional t-SNE embedding that preserves local neighbourhood structure. Several loose clusters emerge, yet almost all of them are multi coloured. The three directional commands share overlapping territories, which reflects the symmetric confusions seen in the confusion matrix. The longer phrases occupy broader regions with slightly more homogeneous colour, consistent with their somewhat higher F1 scores, but there is no clear linear margin separating them from the short commands. The vowel trials disperse across multiple t-SNE islands with substantial intermixing, which again matches their poor separability at the classifier level.

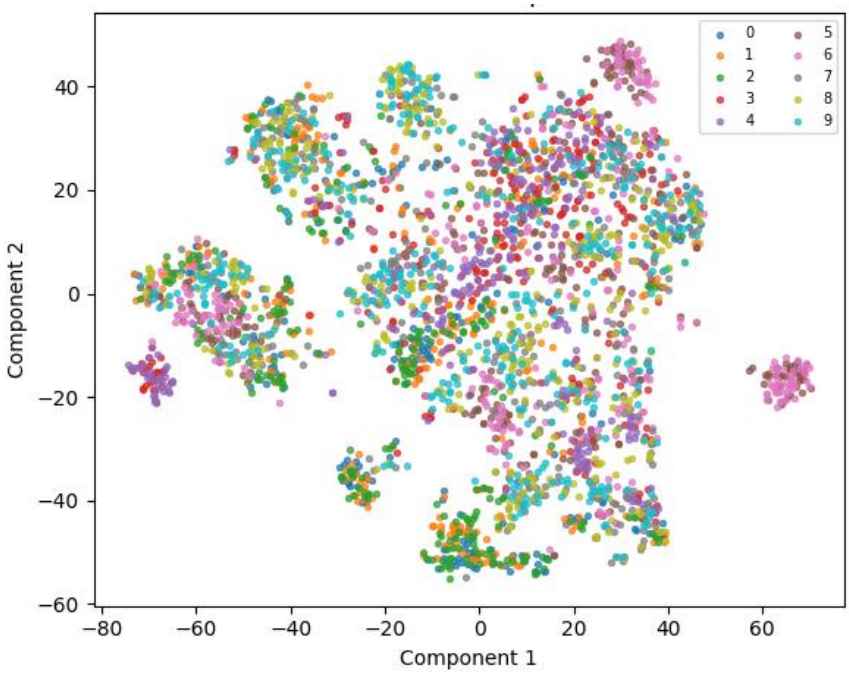

Fig. 7. t-SNE embedding of the reduced ASU feature space, coloured by class label, showing loosely clustered but strongly overlapping class manifolds.

The PCA projection in Fig. 8 provides a complementary, more linear view. The first principal component appears to capture a generic energy or amplitude factor: classes with higher overall band power, such as the composite phrases, are shifted toward the positive side of the axis, whereas short commands and vowels populate the centre and negative side. Along the second component the classes fan out in a wedge shaped pattern, but the degree of overlap remains high. This suggests that, although the block selection and dimensionality reduction pipeline has removed redundant and noisy dimensions, the remaining axes still encode a continuum of task difficulty rather than clearly separated class specific manifolds

Taken together, the t-SNE and PCA plots explain why the linear and non-linear models end up so close in performance. The reduced feature space is only weakly structured at the class level: there are pockets where neighbourhoods align with labels, particularly for some long phrases and composite commands, but large swathes of the space remain mixed. Similar overlapping manifolds have been reported in other imagined speech datasets, even when more aggressive representation learning or Riemannian geometry based features are used [36]. In that context, achieving around 0.41 macro F1 on a ten class, cross trial setting appears consistent with the intrinsic geometry of the data rather than a failure of the classifier family. This observation motivates our later shift toward transfer learning and column wise progressive architectures, where the goal is not only to learn a decision boundary in a fixed space but also to reshape that space through shared representations drawn from larger, more diverse corpora.

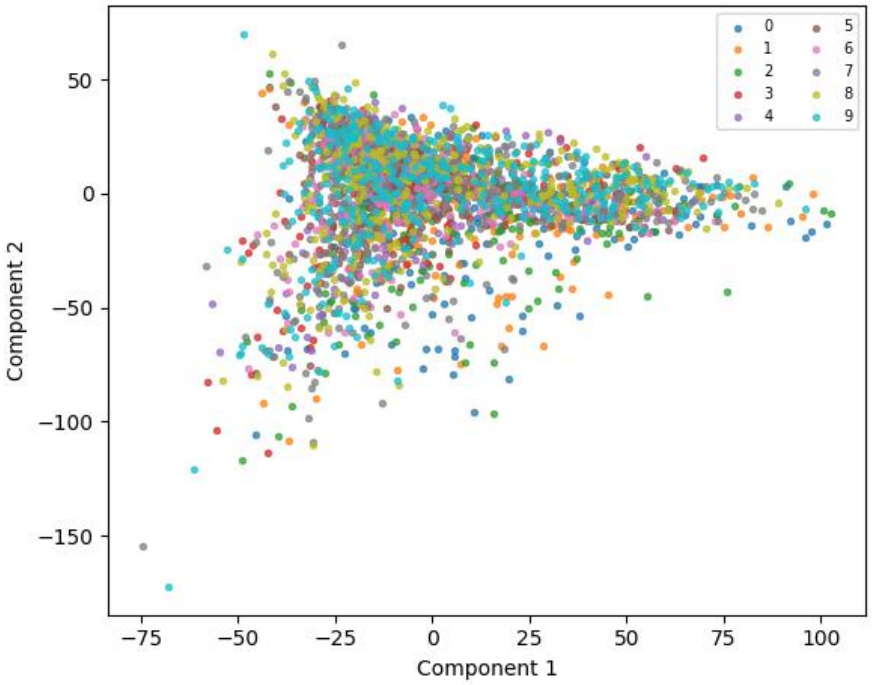

Fig. 8. First two principal components of the reduced ASU feature space, indicating a dominant energy dimension and persistent overla p between commands, phrases, and vowels.

D. Discussion

Taken together, these results confirm that, given the engineered feature space, model choice has limited impact: all three classifiers remain in a narrow band around 0.40–0.42 macro-F1. The ablations in Fig. 5 show that CSP contributes the only clearly non-redundant block, while temporal statistics are largely interchangeable, which explains why deeper models cannot extract much extra structure as described in Section IV-B. Within this context, the confusion matrix in Fig. 4 can be read as a stress test of the feature space. Sentences that differ in high-level semantics but share similar acoustic imagery, such as “in”, “out”, and “up”, are frequently confused, while commands that combine distinctive prosody and articulation, such as “cooperate-in”, are recognised more reliably. The vowels /a/, /i/, and /u/ form another tightly coupled group, with systematic pairwise confusions that align with their proximity in both articulatory and formant space. These patterns reinforce the message from the feature-block analysis: the current representation is sensitive to broad articulatory structure but struggles to separate fine-grained variations that are subtle even at the overt-speech level.

It is important to interpret the approximately 0.41 test accuracy and 0.42 macro-F1 in the light of how the ASU corpus has been used in earlier work. Most published studies restrict attention to a much smaller label space, typically binary contrasts such as “in” versus “cooperate” or three-class vowel sets, and often operate in subject-dependent mode with carefully selected trials. Under these conditions, reported accuracies frequently reach 70–90 per cent, but the task differs fundamentally from the ten-class, subject-wise split considered here. A cross-corpus studies, for example, demonstrates strong performance on ASU and related datasets, yet focuses on a reduced vocabulary and task structure that is closer to command-selection than to continuous imagined speech decoding [37]-[40].

Against this backdrop, the present results sit in the middle of what recent reviews describe as the "moderate-accuracy regime" for non-invasive imagined speech: performance is clearly above chance, robust across folds, and consistent between accuracy and macro-F1, yet still far from the reliability required for high-bandwidth communication [41]-[45] . Our contribution is less about pushing the absolute numbers and more about unpacking where the remaining errors come from. By combining block-wise ablations, classifier comparisons, learning curves, and low-dimensional visualisations, the figures in this section provide a multi-angle diagnosis that is often absent from earlier studies, where only a single classifier and headline accuracy are reported. This richer analysis shows that, for ASU, the bottleneck lies in the front-end representation and label design rather than in a lack of deep learning capacity [46]-[52].

Many imagined-speech systems that report high recognition rates do so within individual subjects, implicitly allowing models to specialise to idiosyncratic neural patterns. In contrast, the overlapping class manifolds observed here indicate that, once trials from multiple talkers and sessions are pooled, the neural signatures of individual words are less separable. This observation aligns with recent calls in the BCI literature to move away from over-optimistic subject-specific benchmarks and toward more realistic cross-subject evaluation protocols.

E. Limitations and Implications for Future Work

Several limitations follow directly from the figures and diagnostics presented above. First, all analyses are restricted to a single corpus. Although the ASU dataset is attractive because it combines ten imagined commands with controlled acquisition, it does not capture cross-linguistic variation, free-running internal speech, or very large vocabularies. The absence of cross-dataset validation means that we cannot yet claim that the identified feature blocks, such as autocorrelation and CSP, will remain optimal when transferred to other corpora like KARA ONE or BCI Competition datasets. Prior cross-corpus work suggests that substantial recalibration is usually required when switching between datasets with different channel layouts or cueing schemes.

Second, the present framework treats each trial as an isolated five-second segment and ignores finer temporal structure within the cue window. While this is appropriate for the feature blocks used here, it limits what the recurrent and convolutional layers can learn. Architectures that operate on lower-level time–frequency maps or directly on sensor-space waveforms might exploit temporal dynamics more fully, at the cost of more complex regularisation. Finally, the t-SNE and PCA plots are inherently descriptive. They suggest that class manifolds overlap and that some clusters are more compact than others, but they do not provide formal guarantees about separability or sample complexity.

Despite these constraints, the analysis points to several concrete directions. The redundancy revealed in Fig. 5 argues for more principled feature selection or sparsity-inducing penalties at the feature level, rather than only at the classifier weights. The confusion patterns in Fig. 4 suggest that future vocabulary design for imagined-speech BCIs should prioritise articulatorily distinct commands and avoid near-minimal pairs until more powerful representations are available. Finally, the modest yet stable performance of the logistic baseline indicates that well-regularised linear models remain a strong reference for imagined-speech EEG and should accompany any claims of deep-learning gains in future studies.

V. Conclusion

This work examined how far carefully engineered EEG features can support imagined speech recognition on the ASU corpus under a ten-class, cross-trial setting. Starting from 6,520 trials, a 3,440-dimensional representation was constructed that combines CSP spatial filters, temporal autocorrelation, PSD band power, and Hjorth parameters, and three classifiers were then evaluated on this representation: logistic regression, a three-layer MLP, and a CNN–BiLSTM–FCNN hybrid. All three models converged to a narrow performance band around 0.41 test accuracy and 0.42 macro-F1, suggesting that, within the evaluated protocol, the main bottleneck lies in the front-end representation and label structure rather than in classifier capacity. The block-wise ablations and geometric analyses clarify this constraint. CSP emerged as the only block whose removal clearly degraded performance on the full set, while temporal statistics were largely interchangeable, and both t-SNE and PCA revealed strongly overlapping class manifolds, especially among short commands and vowels. These patterns offer a plausible account of why deeper models failed to deliver systematic gains and place the present results within the "moderate-accuracy" regime reported for non-invasive imagined speech under realistic evaluation protocols. Several limitations temper the scope of these conclusions. First, the evaluation is confined to a single corpus (ASU); whether the block-level ranking, and in particular the primacy of CSP, holds under different channel layouts, cueing schemes and subject populations cannot be determined from this study alone. Second, all reported numbers are single-seed point estimates; confidence intervals from multi-seed and bootstrap resampling are not yet provided. Third, the deeper classifier treats a heterogeneous feature vector as a 1-D sequence, which is a conservative rather than an architecturally ideal choice for tabular inputs. Against these constraints, the study still provides a transparent and reproducible baseline, showing what can be achieved with well-understood EEG descriptors and conceptually simple classifiers, and pinpointing where progress is most needed. Several concrete next steps follow directly from the analysis. A first priority is cross-corpus validation, replicating the block-wise ablation on Kara One, FEIS and the BCI Competition imagined-speech track to test whether CSP remains the single non-redundant block under different montages and paradigms. A second priority is statistical robustness, repeating every model with at least five seeds and reporting bootstrap confidence intervals on accuracy and macro-F1. A third priority is the classifier side: gradient-boosted ensembles, regularised linear models with sparsity, and attention-based tabular architectures are a more principled match to the heterogeneous feature vector than the CNN–BiLSTM hybrid used here, and quantifying whether any of them escape the 0.42 macro-F1 ceiling on the same representation would be informative. A fourth priority is representation learning and transfer: progressive neural networks and similar architectures that reshape the feature space using larger and more diverse corpora remain the most plausible route to breaking out of the geometry described in Section IV D.

Finally, once a stable offline baseline is established across corpora, a real-time feasibility study – with streaming feature extraction and subject-adaptive calibration – would bridge the offline analysis presented here to the practical BCI setting. In this sense, the present analysis is intended as a diagnostic front end for subsequent stages of a broader programme toward reliable imagined-speech interfaces.

AUTHORS' CONTRIBUTION

The authors' contributions are as follows: Masrullizam Mat Ibrahim and Jamil Abedalrahim Jamil Alsayaydeh handled conceptualization; Safarudin Gazali Herawan was responsible for methodology; Jamil Abedalrahim Jamil Alsayaydeh and Hatem T M Duhair managed software development; Masrullizam Mat Ibrahim and Mazen Farid; validated and conducted formal analysis; Jamil Abedalrahim Jamil Alsayaydeh and Mazen Farid carried out the investigation; resources were provided by Hatem T M Duhair; the original draft was written by Jamil Abedalrahim Jamil Alsayaydeh and Hatem T M Duhair; Hatem T M Duhair reviewed and edited the writing; and Safarudin Gazali Herawan secured funding.

ACKNOWLEDGMENT

The authors express their gratitude to the Centre for Research and Innovation Management (CRIM) at Universiti Teknikal Malaysia Melaka (UTeM) for their valuable support in this research.

DATA AVAILABILITY STATEMENT

All the datasets used in this study are available from the Zenodo database (accession number: https://zenodo.org/records/17816093).

REFERENCES

[1] N. Fitriah, H. Zakaria, and T. L. E. Rajab, "EEG-Based Silent Speech Interface and its Challenges: A Survey," International Journal of Advanced Computer Science and Applications, vol. 13, no. 11, doi: 10.14569/IJACSA.2022.0131173.

[2] X.-Y. Liu et al., "Recent applications of EEG-based brain-computer-interface in the medical field," Military Medical Research, vol. 12, no. 1, p. 14, Mar. 2025, doi: 10.1186/s40779-025-00598-z.

[3] J. T. Panachakel and A. G. Ramakrishnan, "Decoding Covert Speech From EEG-A Comprehensive Review," Front. Neurosci., vol. 15, Apr. 2021, doi: 10.3389/fnins.2021.642251.

[4] K. Su and L. Tian, "Systematic review: progress in EEG-based speech imagery brain-computer interface decoding and encoding research," PeerJ Comput. Sci., vol. 11, p. e2938, Jun. 2025, doi: 10.7717/peerj-cs.2938.

[5] C. H. Nguyen, G. K. Karavas, and P. Artemiadis, "Inferring imagined speech using EEG signals: a new approach using Riemannian manifold features," J Neural Eng, vol. 15, no. 1, p. 016002, Feb. 2018, doi: 10.1088/1741-2552/aa8235.

[6] J. T. Panachakel, A. G. Ramakrishnan, and T. V. Ananthapadmanabha, "Decoding imagined speech using wavelet features and deep neural networks," in 2019 IEEE 16th India Council International Conference (INDICON), IEEE, 2019, pp. 1–4. Accessed: Dec. 05, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9028925/?casa_token=-IZJ8H3BhbUAAAAA:15xGxiarXLBBMqYwlHi-yAIccIACdvz8e_PX7-TqwwpQG293tZoBtqDPXCgKRGYRKlc5ySNgAg

[7] N. C. Mahapatra and P. Bhuyan, "Multiclass Classification of Imagined Speech Vowels and Words of Electroencephalography Signals Using Deep Learning," Advances in Human-Computer Interaction, vol. 2022, 2022, doi: 10.1155/2022/1374880.

[8] S. Alzahrani, H. Banjar, and R. Mirza, "Systematic Review of EEG-Based Imagined Speech Classification Methods," Sensors, vol. 24, no. 24, Art. no. 24, Jan. 2024, doi: 10.3390/s24248168.

[9] D. -Y. Lee, M. Lee, and S. -W. Lee, "Decoding Imagined Speech Based on Deep Metric Learning for Intuitive BCI Communication," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 29, pp. 1363–1374, 2021, doi: 10.1109/TNSRE.2021.3096874.

[10] M. M. Abdulghani, W. L. Walters, and K. H. Abed, "Imagined Speech Classification Using EEG and Deep Learning," Bioengineering (Basel), vol. 10, no. 6, p. 649, May 2023, doi: 10.3390/bioengineering10060649.

[11] Y. F. Alharbi and Y. A. Alotaibi, "Decoding Imagined Speech from EEG Data: A Hybrid Deep Learning Approach to Capturing Spatial and Temporal Features," Life, vol. 14, no. 11, Art. no. 11, Nov. 2024, doi: 10.3390/life14111501.

[12] M. Bisla and R. S. Anand, "Optimized CNN-Bi-LSTM–Based BCI System for Imagined Speech Recognition Using FOA-DWT," Advances in Human-Computer Interaction, vol. 2024, no. 1, p. 8742261, 2024, doi: 10.1155/2024/8742261.

[13] J. T. Panachakel and R. A. Ganesan, "Decoding Imagined Speech From EEG Using Transfer Learning," IEEE Access, vol. 9, pp. 135371–135383, 2021, doi: 10.1109/ACCESS.2021.3116196.

[14] A. Einizade, M. Mozafari, S. Jalilpour, S. Bagheri, and S. Hajipour Sardouie, "Neural decoding of imagined speech from EEG signals using the fusion of graph signal processing and graph learning techniques," Neuroscience Informatics, vol. 2, no. 3, p. 100091, Sept. 2022, doi: 10.1016/j.neuri.2022.100091.

[15] Z. Zhang et al., "Chisco: An EEG-based BCI dataset for decoding of imagined speech," Sci Data, vol. 11, no. 1, p. 1265, Nov. 2024, doi: 10.1038/s41597-024-04114-1.

[16] P. He, G. Wilson, and C. Russell, "Removal of ocular artifacts from electro-encephalogram by adaptive filtering," Med. Biol. Eng. Comput., vol. 42, no. 3, pp. 407–412, May 2004, doi: 10.1007/BF02344717.

[17] T. Memmott et al., "BciPy: brain–computer interface software in Python," Brain-Computer Interfaces, vol. 8, no. 4, pp. 137–153, Oct. 2021, doi: 10.1080/2326263X.2021.1878727.

[18] A. Akbarinia, Optimising EEG decoding with refined sampling and multimodal feature integration. 2024. doi: 10.48550/arXiv.2409.20086.

[19] G. Zhang, D. R. Garrett, and S. J. Luck, "Optimal filters for ERP research I: A general approach for selecting filter settings," Psychophysiology, vol. 61, no. 6, p. e14531, Jun. 2024, doi: 10.1111/psyp.14531

[20] M. Gyurkovics, G. M. Clements, K. A. Low, M. Fabiani, and G. Gratton, "The impact of 1/f activity and baseline correction on the results and interpretation of time-frequency analyses of EEG/MEG data: A cautionary tale," NeuroImage, vol. 237, p. 118192, Aug. 2021, doi: 10.1016/j.neuroimage.2021.118192.

[21] F. Taleb, M. Vasco, N. Rajabi, M. Björkman, and D. Kragic, "Challenging Deep Learning Methods for EEG Signal Denoising under Data Corruption," in 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), July 2024, pp. 1–4. doi: 10.1109/EMBC53108.2024.10782132.

[22] J. Yin, A. Liu, L. Wang, R. Qian, and X. Chen, "Integrating spatial and temporal features for enhanced artifact removal in multi-channel EEG recordings," J. Neural Eng., vol. 21, no. 5, p. 056018, Sept. 2024, doi: 10.1088/1741-2552/ad788d.

[23] B. Pester and C. Ligges, "Does independent component analysis influence EEG connectivity analyses?" in Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC), Jul. 2018, pp. 1007–1010, doi: 10.1109/EMBC.2018.8512425.

[24] A. Echtioui, W. Zouch, M. Ghorbel, M. B. Slima, A. B. Hamida, and C. Mhiri, "Automated EEG Artifact Detection Using Independent Component Analysis," 2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), pp. 1–5, Sept. 2020, doi: 10.1109/ATSIP49331.2020.9231574.

[25] X. Yin, M. Meng, Q. She, Y. Gao, and Z. Luo, "Optimal channel-based sparse time-frequency blocks common spatial pattern feature extraction method for motor imagery classification," Math. Biosci. Eng., vol. 18, no. 4, pp. 4247–4263, 2021, doi: 10.3934/mbe.2021213.

[26] B. Wang et al., "Common Spatial Pattern Reformulated for Regularizations in Brain-Computer Interfaces," IEEE Trans. Cybern., vol. 51, no. 10, pp. 5008–5020, 2021, doi: 10.1109/TCYB.2020.2982901.

[27] D. Pawar and S. Dhage, "Imagined Speech Classification using EEG based Brain-Computer Interface," in 2022 IEEE 11th International Conference on Communication Systems and Network Technologies (CSNT), IEEE, 2022, pp. 662–666. Accessed: Dec. 05, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9787644/?casa_token=q5YZSt-70HgAAAAA:3jTrKlxnQOJjSbaPu6dC9Ilr1B1hhSbia7GrKko6EdG9lSa-blnvzPpn7EBaWnD0JjEkF1fd7A

[28] A. Moura, S. Lopez, I. Obeid, and J. Picone, "A comparison of feature extraction methods for EEG signals," in 2015 IEEE Signal Processing in Medicine and Biology Symposium (SPMB), Dec. 2015, pp. 1–2. doi: 10.1109/SPMB.2015.7405430.

[29] I. Carrara and T. Papadopoulo, "Classification of BCI-EEG based on augmented covariance matrix," 2023, doi: 10.48550/ARXIV.2302.04508.

[30] S.-H. Lee, M. Lee, and S.-W. Lee, "Neural decoding of imagined speech and visual imagery as intuitive paradigms for BCI communication," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 28, no. 12, pp. 2647–2659, 2020.

[31] W. H. Alawee, A. Basem, and L. A. Al-Haddad, "Advancing biomedical engineering: Leveraging Hjorth features for electroencephalography signal analysis," J Electr Bioimpedance, vol. 14, no. 1, pp. 66–72, doi: 10.2478/joeb-2023-0009.

[32] O. George et al., "State-of-the-Art Versus Deep Learning: A Comparative Study of Motor Imagery Decoding Techniques," IEEE Access, vol. 10, pp. 45605–45619, 2022, doi: 10.1109/ACCESS.2022.3165197.

[33] F. Qi, W. Wu, K. Liu, T. Yu, and Y. Cao, "A Logistic Regression Based Framework for Spatio-Temporal Feature Representation and Classification of Single-Trial EEG," May 2021. doi: 10.1007/978-981-16-2336-3_36.

[34] D. Alonso-Vázquez, O. Mendoza-Montoya, R. Caraza, H. R. Martinez, and J. M. Antelis, "From pronounced to imagined: improving speech decoding with multi-condition EEG data," Front. Neuroinform., vol. 19, June 2025, doi: 10.3389/fninf.2025.1583428.

[35] P. Saha, S. Fels, and M. Abdul-Mageed, "Deep Learning the EEG Manifold for Phonological Categorization from Active Thoughts," in ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2019, pp. 2762–2766. doi: 10.1109/ICASSP.2019.8682330.

[36] M. A. Bakhshali, M. Khademi, A. Ebrahimi-Moghadam, and S. Moghimi, "EEG signal classification of imagined speech based on Riemannian distance of correntropy spectral density," Biomedical Signal Processing and Control, vol. 59, 2020, doi: 10.1016/j.bspc.2020.101899.

[37] J. T. Panachakel, A. G. Ramakrishnan, and T. V. Ananthapadmanabha, "A novel deep learning architecture for decoding imagined speech from EEG," arXiv:2003.09374, Mar. 2020.

[38] H. T. M. Duhair, M. bin mat Ibrahim, J. A. J. Alsayaydeh, R. Bacarra, and A. H. Ahmed, "Progressive transfer learning Unifies Multi-Corpus EEG for robust and scalable Imagined-Speech decoding," Biomedical Signal Processing and Control, vol. 116, p. 109475, May 2026, doi: 10.1016/j.bspc.2026.109475.

[39] J. A. J. Alsayaydeh, W. A. Y. Khang, W. A. Indra, V. Shkarupylo and J. Jayasundar. 2019. Development of smart dustbin by using apps. ARPN Journal of Engineering and Applied Sciences. 14(21): 37033711.

[40] A. A. AlZubi, "Bunch graph based dimensionality reduction using auto-encoder for character recognition," Multimedia Tools and Applications, 2022.

[41] F. Lotte et al., "A review of classification algorithms for EEG-based brain-computer interfaces: a 10 year update," J Neural Eng, vol. 15, no. 3, p. 031005, June 2018, doi: 10.1088/1741-2552/aab2f2.

[42] J. A. J. Alsayaydeh, W. A. Indra, A. W. Y. Khang, V. Shkarupylo, and D. A. P. P. Jkatisan, "Development of vehicle ignition using fingerprint," ARPN Journal of Engineering and Applied Sciences, vol. 14, no. 23, pp. 4045–4053, 2019. [Online]. Available: http://www.arpnjournals.org/jeas/research_papers/rp_2019/jeas_1219_8024.pdf

[43] L. Kovalchuk, D. Kaidalov, A. Nastenko, M. Rodinko, O. Shevtsov, and R. Oliynykov, "Decreasing security threshold against double spend attack in networks with slow synchronization," Computer Communications, vol. 154, pp. 75–81, 2020, doi: 10.1016/j.comcom.2020.01.079.

[44] J. A. J. Alsayaydeh, W. A. Indra, A. W. Y. Khang, A. K. M. Z. Hossain, V. Shkarupylo, and J. Pusppanathan, "The experimental studies of the automatic control methods of magnetic separators performance by magnetic product," ARPN Journal of Engineering and Applied Sciences, vol. 15, no. 7, pp. 922-927, 2020.

[45] N. F. B. A. Rahim, A. W. Y. Khang, A. Hassan, S. J. Elias, J. A. M. Gani, J. Jasmis, and J. A. J. Alsayaydeh, "Channel Congestion Control in VANET for Safety and Non-Safety Communication: A Review," in 2021 6th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE), 2021, pp. 1-6, doi: 10.1109/ICRAIE52900.2021.9704017.

[46] J. A. J. Alsayaydeh et al., "Development of vehicle door security using smart tag and fingerprint system," International Journal of Engineering and Advanced Technology, vol. 9, no. 1, pp. 3108–3114, Oct. 2019, doi: 10.35940/ijeat.E7468.109119

[47] A. Mehmood, A. Abugabah, A. A. AlZubi, and L. Sanzogni, "Early diagnosis of Alzheimer's disease based on convolutional neural networks," Computer Systems Science and Engineering, vol. 43, no. 1, pp. 305–315, 2022.

[48] J. A. Alsayaydeh, M. Nj, S. N. Syed, A. W. Yoon, W. A. Indra, V. Shkarupylo and C. Pellipus, "Homes appliances control using bluetooth," ARPN Journal of Engineering and Applied Sciences, vol. 14 (19), pp. 3344-3357, 2019.

[49] A. Al Smadi, S. Yang, A. Abugabah, A. A. AlZubi, and L. Sanzogni, "A pansharpening based on the non-subsampled contourlet transform and convolutional autoencoder: Application to QuickBird imagery," IEEE Access, vol. 10, pp. 44778–44788, 2022.

[50] V. Shkarupylo, I. Blinov, A. Chemeris, V. Dusheba, J. A. J. Alsayaydeh, and A. Oliinyk, "Iterative approach to TLC model checker application," in 2021 IEEE 2nd KhPI Week on Advanced Technology (KhPIWeek), Kharkiv, Ukraine, 2021, pp. 283–287, doi: 10.1109/KhPIWeek53812.2021.9570055.

[51] H. T. M. Duhair, M. Bin Mat Ibrahim, M. Farid, J. A. J. Alsayaydeh, and S. G. Herawan, "EEG-Based Imagined-Speech Decoding: A Review," Intl. J. Adv. Comput. Sci. Appl., vol. 17, no. 1, pp. 380–390, 2026, doi: 10.14569/IJACSA.2026.0170136.

[52] A. Karatkevich and I. Grobelna, "Deadlock detection in Petri nets: One trace for one deadlock?," 2014 7th International Conference on Human System Interactions (HSI), Costa da Caparica, Portugal, 2014, pp. 227-231, doi: 10.1109/HSI.2014.6860480.