Hindawi
Advances in Human-Computer Interaction
Volume 2022, Article ID 1374880, 10 pages
https://doi.org/10.1155/2022/1374880

Research Article

Multiclass Classification of Imagined Speech Vowels and Words of Electroencephalography Signals Using Deep Learning

Nrushingh Charan Mahapatra [1,2] **and Prachet Bhuyan** [2]

[1]*Intel Technology India Pvt Ltd, Bengaluru 560103, India*
[2]*School of Computer Engineering, Kalinga Institute of Industrial Technology, Bhubaneswar 751024, India*

Correspondence should be addressed to Nrushingh Charan Mahapatra; 1981030@kiit.ac.in

Received 4 April 2022; Revised 23 June 2022; Accepted 2 July 2022; Published 20 July 2022

Academic Editor: Christos Troussas

The paper's emphasis is on the imagined speech decoding of electroencephalography (EEG) neural signals of individuals in accordance with the expansion of the brain-computer interface to encompass individuals with speech problems encountering communication challenges. Decoding an individual's imagined speech from nonstationary and nonlinear EEG neural signals is a complex task. Related research work in the field of imagined speech has revealed that imagined speech decoding performance and accuracy require attention to further improve. The evolution of deep learning technology increases the likelihood of decoding imagined speech from EEG signals with enhanced performance. We proposed a novel supervised deep learning model that combined the temporal convolutional networks and the convolutional neural networks with the intent of retrieving information from the EEG signals. The experiment was carried out using an open-access dataset of fifteen subjects' imagined speech multichannel signals of vowels and words. The raw multichannel EEG signals of multiple subjects were processed using discrete wavelet transformation technique. The model was trained and evaluated using the preprocessed signals, and the model hyperparameters were adjusted to achieve higher accuracy in the classification of imagined speech. The experiment results demonstrated that the multiclass imagined speech classification of the proposed model exhibited a higher overall accuracy of 0.9649 and a classification error rate of 0.0350. The results of the study indicate that individuals with speech difficulties might well be able to leverage a noninvasive EEG-based imagined speech brain-computer interface system as one of the long-term alternative artificial verbal communication mediums.

1. Introduction

Speech is an essential communication channel for people to connect with society. There are difficulties with external speech stimulation due to the medical conditions of some individuals, such as speech delay, autism, brain stroke, old age, Down syndrome, and other neurological diseases. Advanced human-computer interface (HCI) technology based on neural signals, also known as brain-computer (machine) interface (BCI/BMI), attempts to connect individuals with speech difficulties to society by decoding messages from the brain's neural activity through mental imagery rather than depending on natural speech mechanisms [1,2]. Speech imagery, or imagined speech, is defined as the neural representation of speech in the absence of natural speech, which occurs when a person imagines or thinks about syllables or words but does not produce natural sounds [3–5]. The availability of noninvasive EEG devices for measuring speech neural activity in the brain and advanced deep learning techniques has contributed to the development of the imagined speech-based BCI, which is expected to be the imminent verbal communication alternative for speech-disordered individuals. The imagined speech EEG-based BCI system decodes or translates the subject's imaginary speech signals from the brain into messages for communication with others or machine recognition instructions for machine control [6]. Decoding imagined speech from brain signals to benefit humanity is

one of the most appealing research areas. In the absence of any traceable auditory output that is synced to the imagery speech of subjects' brain activity, decoding imagined speech is challenging.

We reviewed previous scientific work in the discipline of imagined speech decoding from neural EEG signals. Dasalla et al. [3] classified the English vowels using the feature extraction common spatial patterns (CSP) and a machine learning classifier, support vector machine (SVM). Wang et al. [7] experimented with two Chinese characters, extracting signal feature information with CSP and classifying them with SVM. Kim et al. [8] demonstrated effective vowel classification by utilizing a linear discriminant analysis (LDA) classifier and feature extraction approaches such as CSP and empirical mode decomposition (EMD). Min et al. [9] used the extreme learning machine (ELM) and classifier SVM to decode the imagined speech. Yoshimura et al. [10] decoded Japanese vowels two-class using sparse logistic regression (SLR). Sun et al. [11] classified the ten phonemes using feature extraction techniques such as the restricted Boltzmann machine (RBM) and neural network- (NN-) based classifiers. Nguyen et al. [12] improved the accuracy of multiclass imagined speech decoding by combining the Riemannian manifold feature extraction approach with classifier relevance vector machines (RVM). Saha et al. [13] demonstrated the classification of eleven speech sounds using temporal and spatial CNN with a deep autoencoder (DAE). Cooney et al. [14] investigated the imagined speech vowel classification using a deep CNN transfer learning (TL) model and observed that TL produced relatively better accuracy. Panachakel et al. [15] showed comparable higher accuracy than previous research when decoding imagery speech of two words using the feature extraction technique CSP and DWT, a deep neural network (DNN) classifier model. Cooney et al. [16] evaluated the impact of hyperparameter optimization of imagined speech EEG signals on deep learning models with CNN. Tamm et al. [17] used the classifiers CNN and TL to decode imagined vowels. Pawar et al. [18] demonstrated the covert speech multiclass classification of four distinct words using a kernel-based extreme learning machine (kernel ELM). Li et al. [19] classified the imagined speech of eight words using a hybrid convolution network. Sarmiento et al. [20] used the CNN-based model to classify five English vowels. Panachakel and Ganesan [21] used sliding window data augmentation and TL on the underlying ResNet50 model to classify imagined spoken words and vowels.

Even though multiple studies on this topic have been conducted over the last decade, we examined the relevant publications and found that the accuracy of decoding imagined speech necessitates more attention and investigation. The primary purpose of the work is to employ advanced deep learning (DL) approaches to decode or classify multiclass imagined speech from multichannel EEG neural signals of multiple subjects with better accuracy.

Mathematical Representation. The hypothesis is defined as a supervised multiclass classifier model. Assuming that the model's input includes labelled EEG signals, each labelled EEG signal is represented as (E_i, s_i) for the record (i), the true imagined speech is $s_i \in \{Speech\ Labels\}$, and the input $E_i \in R^{n\times m}$ is a preprocessed signal of a two-dimensional vector, where n is the number of EEG channels and m is the number of signal sample points for each channel. The model output classified or decoded imagined speech is $\widehat{s_i} \in \{Speech\ Labels\}$ and $s_i = \widehat{s_i} + Error_i$, where classification error is $Error_i$. The classifier model is mathematically defined as in equation (1).

$$f(E_i) = \widehat{s_i} + \text{Error}_i. \tag{1}$$

In this paper, we presented a novel supervised deep learning model that integrated temporal convolutional networks (TCN) with CNN for the multiclass imagined speech decoding of EEG signals. The preprocessing method DWT was used for feature extraction and artifact removal. The input network layer and multiple hidden network layers were used to extract and transform the underlying signal features. The output network layer was to classify the imagined speech vowels and words. To evaluate the classification outcome performance of the model, validation indicators such as recall, precision, f1-score, Cohen's kappa score, and confusion matrix were used. The following is an outline of the contribution of the paper. The wavelet-based analysis, together with the deep learning model consolidated by the TCN and CNN, demonstrated effectiveness in decoding the multiclass eleven imagined speech EEG signals into vowels and words.

2. Materials and Methods

The architecture of the EEG-based imagined speech BCI is depicted in Figure 1. The system architecture consists of modules for acquiring neural signals, signal preprocessing, and the classification algorithm for converting the signals into decoded imagined speech. The DL technique was used to eradicate manual feature extraction since traditional machine learning algorithms have a challenging learning process that requires the manual extraction of the feature from the signals.

Multichannel imagined speech neural signals are collected using EEG electrodes placed on the scalp. Signal processing techniques such as resampling, band-pass filtering, notch filtering, artifact removal, and feature extraction are used to preprocess the raw signals. The deep learning model was trained and evaluated for the classification performance of the preprocess signal.

2.1. EEG Data. The data used for the experiment in this study was an open-access dataset of EEG signals from fifteen subjects [22]. According to the ten-twenty international system of electrode scalp locations [23], the collected signals have six channels such as C3, F3, P3, C4, F4, and P4. The data were collected from multiple subjects, and each subject recorded multiple trials. The signals measured during the subject's imagined speech contained all five English vowels and six Spanish words. The signal sampling rate was 1024 Hz. Figure 2 illustrates the EEG signal acquisition with speech classes.

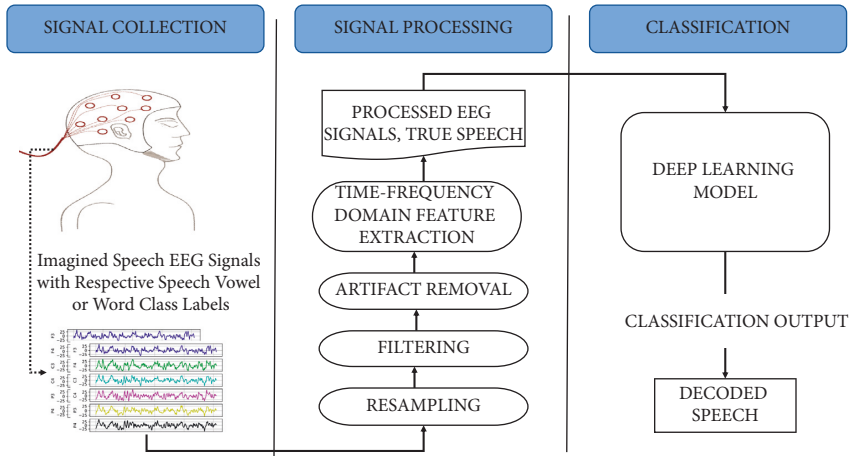

Figure 1: The architecture of an EEG-based imagined speech BCI system.

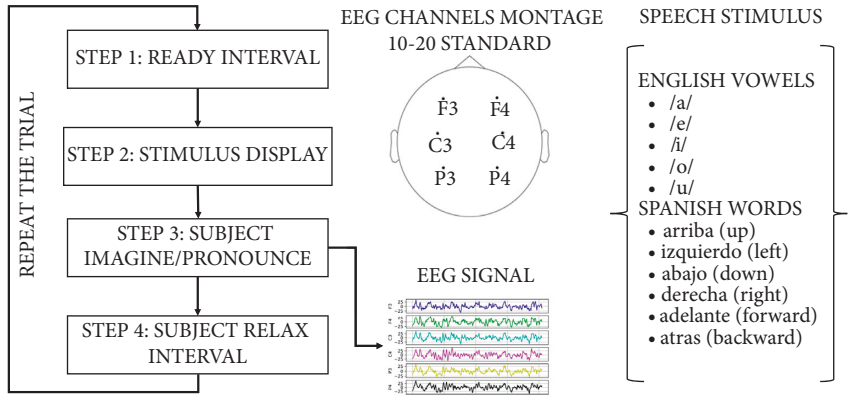

Figure 2: The diagram shows the procedure of acquiring an EEG signal.

Step 1: contained a two-second preparedness phase to notify the subject that the trial was about to begin

Step 2: involved presenting the stimulus word or vowel for two seconds

Step 3: entailed capturing the subject's brain wave EEG for a four-second period of the subject's imagined or spoken utterance of the word or vowel

Step 4: involved the subject relaxing for four seconds and preparing for the next stimulus

2.2. EEG Signal Preprocessing. The raw EEG signals were preprocessed before being used in the deep learning model. The EEG signal sampling rate was recorded at 1024 Hz, and the signal was recorded for four seconds. Each signal record was represented as a two-dimensional vector of (n, m), where $n = 6$ was the number of EEG channels and $m = 4096$ was the signal sampling points of each channel. The sampling rate of the signal was rescaled from 1024 Hz to 512 Hz. The resultant resampled signal was a two-dimensional ($n = 6$, $m = 2048$) vector. The frequencies of the signals below 0.01 Hz and above 100 Hz were filtered out to keep all the frequency bands for data interpretation. The signals were notch filtered to reduce the 50 Hz noise created by the electrical environment around the collected EEG data.

2.2.1. Independent Component Analysis (ICA). The EEG signal that caught the artifacts was generated by biological signals other than the brain, such as eye movements, heart rhythms, and muscle activity. The ICA blind source separation approach separates statistically independent signal components and can automatically eliminate artifacts from EEG signals [24]. The signals were processed by the FastICA algorithm to remove artifacts [25].

2.2.2. Discrete Wavelet Transformation (DWT). The EEG signal is nonstationary and nonlinear in composition [26, 27]. The wavelet time-frequency analysis gives the best performance results on the nonstationary input signals [28]. The EEG signal that caught the artifacts was generated by biological signals other than the brain, such as eye movements, heart rhythms, and muscle activity. The DWT was used to denoise the signal and extract features such as temporal information and local spectral information from signals. A wavelet is defined as a time-restricted wavelike oscillation. The orthogonal property of the mother wavelet db10 allows for smooth signal reconstruction. The mother wavelet db10 is notably effective for feature extraction in EEG wavelet based in several applications [29–31]. Although different mother wavelets were attempted in the study to

increase classification performance, at the end of the analysis, wavelets db10 offered higher classification performance. With the mother wavelets db10, the raw EEG signal is decomposed into 4 levels. The frequencies of the resulting subbands are in the ranges of 0–12.5 Hz, 12.5–25 Hz, 25–50 Hz, and 50–100 Hz. The signal's energy is almost entirely contained in these decomposed subbands. Because the mother wavelet closely resembles the signal, higher coefficients corresponding to eye movements and lower coefficients relating to noise are generated. We also noticed that, after the 4-level decomposition, there was no improvement in classification performance. As a result, the signals were preprocessed using the mother wavelet, Daubechies (db10) with level 4. The DWT computes the wavelet present in the signal given the scale and position on the discrete grid. The signal's breakdown into several time series of wavelet coefficients depicts the signal's temporal evolution in the associated frequency band. To denoise the signals, the threshold value was applied to their decomposed wavelet coefficients to produce the estimated value of the wavelet coefficients. The threshold value for the decomposed wavelet is computed using the formula universal threshold defined in equation (2). The inverse wavelet transform reconstructs the signals using the estimated wavelet coefficient values:

$$\text{threshold} = \text{sigma}^{*}\sqrt{(2^{*}\log(\text{signal length}))},$$
$$\text{where sigma} = \frac{(\text{mean absolute deviation (wavelet coefficient)})}{0.6745}. \tag{2}$$

2.3. Proposed Deep Learning Model. The model architecture was designed with the objective of learning the feature representations from the EEG multichannel or multidimensional signals of imagined speech of the subjects. Figure 3 details the architectural diagram of the proposed model. The TCN stream was responsible for learning temporal features, whereas the CNN stream was responsible for learning spatial features from the preprocessed EEG signals. As a result, the model was able to learn both the temporal and spatial characteristics of the signals. The concatenation of the learned information streams was input into the fully connected layers for feature transformation and multiclass imagined speech classification.

In Figure 3, the input layer, TCN layer, CNN layer, and output layer are presented. A dilated casual convolution operation with the dilation factor and a kernel of size three is used to demonstrate the TCN residual block. The classification layer and the decoded speech label are displayed.

2.3.1. TCN Branch. The TCN consists of a stack of residual blocks of one-dimensional causal dilated convolution, with the network layer input and output sequence lengths always being the same [32]. The TCN was used for the extraction of temporal-recurrent unique electrical signatures of each speech class from the imagined speech signal. The residual block consists of one-dimensional dilated causal convolution, batch normalization, nonlinear activation function, one-dimensional spatial dropout, and residual connection. One-dimensional causal convolution is divided into two segments such as causal convolution and one-dimensional dilated convolution.

Causal Convolution. Causal convolutions are a type of convolution used for temporal signals that preserve and will not compromise the order of the information. The causal convolution of output o_t at time step t for a given channel's input signal sampling points $e_0, e_1, \ldots, e_{m-1}$ and output $o_0, o_1, \ldots, o_{m-1}$ is only performed on $e_0, e_1, \ldots, e_t$ earlier observed sampling points, not the future observed sampling points $e_{t+1}, \ldots, e_{m-1}$ [33]. The convolution operation can be represented as a mathematical prediction function $\Pr(o_t|e_0, e_1, \ldots, e_t)$ for time step $t = 0, 1, \ldots, m-1$.

One-Dimensional Dilated Convolution. This is a convolutional variation in which the kernel is expanded by inserting gaps between the kernel elements. A convolution type one-dimensional (1D) computation creates a 1D output signal by applying a 1D filter to a 1D input signal. In the dilated convolution, the convolution filter size of k was applied to the sampling points of an input signal of a length equal to or greater than the filter size k by skipping the signal sampling points with the dilation step, which means that as the depth increased, the dilated convolution generated an expanding receptive field [34]. As the depth of the layer is increased, the dilation step d is increased by a factor of two, resulting in $d = 2^{h-1}$ at the layer $h = 1, 2, 3, \ldots$ of the block. Figure 4 shows examples of one-dimensional dilated convolution. A dilated convolution with a dilation value of one produced the conventional convolution.

In this case, the input signal is two-dimensional (number of channels and number of sampling points). A one-dimensional kernel filter has a size of 3, has 1 filter, and has 2 dilation steps.

For the input signal sampling point sequence $e \in E$, the dilation step d, and the filter size of $k \in R$, the one-dimensional dilated causal convolution function dilcasualconv1d at the sampling time t is defined as in equation (3), where $t - d.j$ is the index of the past sampling time and filter(j) is the convolutional filter function:

$$\text{dilcasual conv1d}(e, t) = \sum_{j=0}^{k-1} e_{t-d.j} \odot \text{filter}(j). \tag{3}$$

Residual Connection. The residual block structure comprises multiple layers for increased receptivity. To allow a multilayer in the model while avoiding gradient explosion or vanishing, a residual connection of the input was added to the output residual transformation function after applying one-dimensional convolution to the input and allowing information to flow across the layers. The residual block output for the input e is defined by the following equation:

$$\text{Residual}_{\text{output}} = \text{Conv1D}(e) + \text{Residual}_{\text{transform}}(e). \tag{4}$$

The nature of neural signals has a distribution of features across the dataset due to the EEG data being collected from

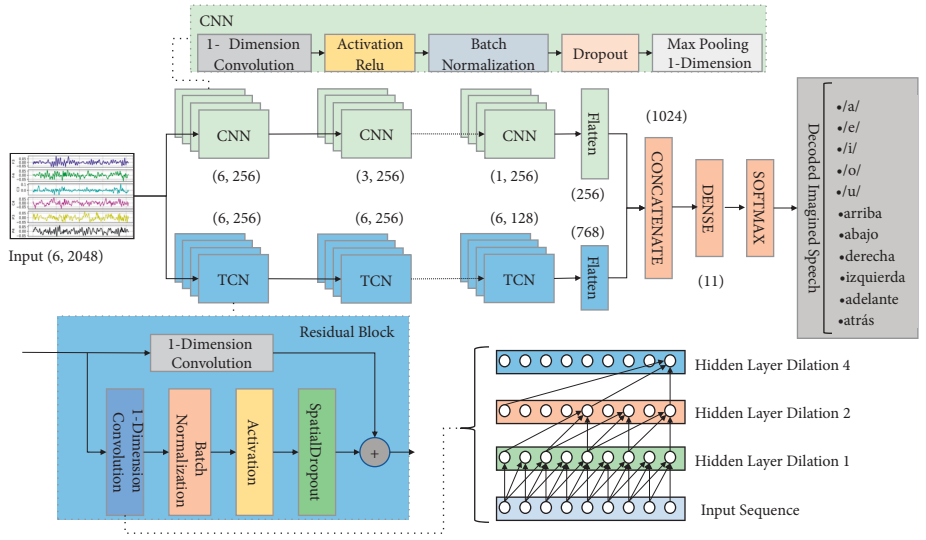

Figure 3: The proposed deep learning model architecture is depicted in the diagram.

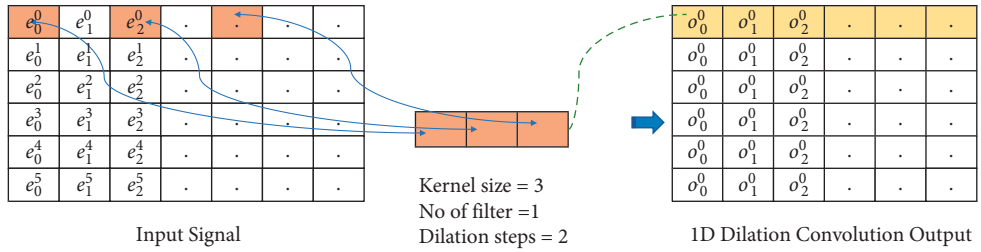

Figure 4: Example of one-dimensional dilated convolution.

multiple subjects having independent medical conditions. The model experienced an internal covariate shift from layer to layer, and the model was unable to learn the features. The batch normalization was added before the activation layer to prevent the internal covariate shift in the model [35]. To avoid model overfitting in the residual block, regularization such as one-dimensional spatial dropout was used. The one-dimensional spatial dropout with probability removed the complete feature learning of the convoluted filter channels, not simply a few neurons from each channel [36]. The nonlinear activation function gated activation unit was used in the residual block [37], and the respective activation function formula is defined as in the following equation:

$$o_i = \tanh(z_i) \odot \text{sigmoid}(z_i),$$

$$\tanh(z_i) = \frac{\left(e^{z_i} - e^{-z_i}\right)}{\left(e^{z_i} + e^{-z_i}\right)}, \tag{5}$$

$$\text{sigmoid}(z_i) = \frac{1}{\left(1 + e^{-z_i}\right)},$$

where z_i is the input neuron and $\odot$ is the elementwise dot product.

2.3.2. CNN Branch. The CNN stream consists of a series of convolution blocks. Every convolution block has a sequence of components such as one-dimensional convolution, activation unit, batch normalization, regularization dropout, and one-dimensional max-pooling. The nonlinear activation function of the rectified linear unit known as the ReLU is used. The output of ReLU is nearly linear, defined as ReLU $(z_i) = \max(0, z_i)$, where s is the neuron input [38]. The regularization dropout technique was used to improve model generalization and avoid overfitting by randomly dropping the units from the model training [39]. The signals were downsampled using one-dimensional max-pooling, which takes the maximum value over a pool-sized sliding window by applying it to the signals, retaining only the high-resolution feature information.

2.3.3. Transformation and Classification Layer. The model performed the feature transformation and multiclass classification on the CNN and TCN stream outputs. The CNN stream output was flattened and created a one-dimensional feature of the learning vector. The TCN stream output was flattened, resulting in a one-dimensional feature learning vector. The one-dimensional feature learning vectors of TCN and CNN were then concatenated. The activation function exponential linear unit (ELU) defined in equation (6) was applied to the combined feature vector [40]. The combined feature vector was used in the dense or fully connected layers for the feature information transformation

or extraction, and the extracted information was used in the multiclass speech classification.

$$\mathrm{ELU}(z_i) = \begin{cases} z_i, & \text{if } z_i > 0, \\ \gamma(e^{z_i} - 1), & \text{if } z_i < 0, \end{cases} \quad (6)$$

where the hyperparameter γ>0 controls the value to saturate for the z_i<0 negative neuron input.

The softmax activation function was applied in the model output layer to achieve multiclass speech classification. The softmax formula, which calculates the probability value of each speech class, is defined as in equation (7). The model classification result was the decoded speech class with the highest softmax value.

$$\mathrm{softmax}(s_i^j) = \frac{e^{s_i^j}}{\sum_{k=1}^{l} e^{s_i^k}}, \quad (7)$$

where the input vector $s_i^j \in \{s_i^1, s_i^2, \ldots, s_i^l\}$, where the expected speech label of record (i) and l represents the number of imagined speech classes. The total softmax value of all input vectors is equal to 1, $\sum_{j=1}^{l} \mathrm{softm}\mathbf{a}\mathrm{x}(s_i^j) = 1$.

2.4. Model Training and Experiments. The proposed deep learning supervised multiclass model was developed, trained, and validated using the scikit-learn, Keras, and TensorFlow frameworks [41–43]. The optimization algorithm Adam was used in the model training [44]. The categorical cross-entropy cost function, which computes the loss between the true speech labels and the model's decoded imagined speech outputs, is defined in equation (8) and was used in model training. The model architecture parameters were frozen after the model attained a certain level of performance and continued with hyperparameter optimization. Table 1 shows the detailed summary of model architectural parameters.

$$\mathrm{loss} = -\frac{1}{T}\sum_{i=1}^{T}\sum_{k=1}^{l} s_i^k \log \widehat{s}_i^k, \quad (8)$$

where s_i denoted as $[s_i^1, s_i^2, s_i^3, \ldots, s_i^l]^T$ is the categorical representation of actual speech label of record (i), $\widehat{s}_i$ denoted as $[\widehat{s}_i^1, \widehat{s}_i^2, \widehat{s}_i^3, \ldots, \widehat{s}_i^l]^T$ is the decoded categorical representation (softmax) of imagined speech for record (i), l represents the number of imagined speech classes, and T stands for the total number of records.

All the subjects' EEG signal data were aggregated once the signal preprocessing was completed. In the preprocessed signal, the distribution of speech classes was almost balanced. The fivefold cross-validation technique was applied for the proposed model training and evaluation.

3. Results and Discussion

3.1. Classification Results of Different Models. We carried out a few experiments involving signal processing (DWT; ICA), and the DL model architecture as the ICA (FastICA algorithm) blind source separation approach separates statistically independent signal components and can automatically eliminate artifacts from EEG signals. Table 2 displays the different experiments, mean cross-validation classification accuracy results, and comparisons. According to the outcomes, the wavelet-based signal processing and the DL model united TCN with the CNN and improved the performance.

Table 1: Summary of the model architectural parameters.

	Parameter	Values
Input layer	Signal dimension	(6, 2048)
TCN layer	No. of blocks	5
	No. of filters	256
	conv1d	Causal
	Dilation factor	1,2,4
	Kernel	3
	Activation	Gated activation unit
	Spatialdropout1d	0.2
CNN layer	No. of blocks	5
	Convolution	conv1d
	No. of filters	256
	Activation	ReLU
	Dropout	0.1
	Pooling	maxpool1d (2)
Transformation layer	Fully connected	1024
	Activation	ELU
Classification layer	Fully connected	11
	Activation	Softmax
	Loss function	Cross entropy
Training	Optimizer	Adam
	Learning rate	0.00001
	Epochs	300
	Batch size	256

In the case of experiment 5, both the ICA and DWT were used in the preprocessing of the EEG signals. As both the ICA and DWT are used in signal preprocessing, most artifacts are eliminated. In the case of experiment 6, only DWT was used in the preprocessing of the EEG signals, and the DWT removed some artifacts by thresholding the decomposed wavelet, but the resultant signals still have some artifacts [45]. As in the case of experiment 6, where signals contain some artifacts, the model's generalization capability increases, adding variation, and the proposed model performs better than experiment 5, which combined ICA and DWT signal preprocessing.

3.2. Outcomes of Proposed Methods. The experimental outcomes of the proposed model (DWT; TCN + CNN) for multiclass decoding of imagined speech achieved an overall accuracy of 0.9649. The overall multiclass decoding error rate was 0.0350. The model's statistical metric precision, which measures the ability to decode each speech label, was in the range of 0.92–0.99. The model's statistical metric recall, which measures the ability to decode all relevant speech labels, was in the range of 0.95–0.99. The f1-score harmonizes the mean value between recall and precision, and the model f1-score ranged from 0.94 to 0.98. Table 3 summarizes the details of the precision, recall, and f1-score of each imagined speech vowel and word.

The model's confusion matrix demonstrates how the model correctly decodes each imagined speech class and

Table 2: Classification accuracy of distinct experiments.

Experiment number	Signal processing	DL model	Overall accuracy (%)
1	ICA	TCN	72.29
2	DWT	TCN	81.48
3	ICA + DWT	TCN	82.37
4	ICA	TCN + CNN	94.35
5	ICA + DWT	TCN + CNN	95.63
6	DWT	TCN + CNN	96.49

Table 3: Model evaluation reports.

Imagined speech	Precision	Recall	F1-score
$\|a\|$	0.97	0.95	0.96
$\|e\|$	0.92	0.96	0.94
$\|i\|$	0.98	0.96	0.97
$\|o\|$	0.96	0.98	0.97
$\|u\|$	0.99	0.97	0.98
Up	0.98	0.99	0.98
Down	0.93	0.98	0.95
Forward	0.98	0.95	0.96
Back	0.98	0.98	0.98
Right	0.96	0.95	0.96
Left	0.98	0.96	0.97

becomes wrongly classified or confused when trying to decode each imagined speech signal. Figure 5 depicts the entire experiment report confusion matrix.

A statistical function such as Cohen's kappa [46] was used to measure the degree of agreement between model predictions and true labels to examine the randomness of the proposed model and is defined in equation (9). Cohen's kappa value of the proposed model was 0.9614.

$$k_{\text{score}} = (p_o - p_e)/(1 - p_e), \tag{9}$$

where p_o represents the actual model overall accuracy and p_e represents the expected random prediction accuracy.

Figure 6 depicts the overall classification accuracy of imagined speech for all the fifteen subjects. Compared to the other subjects, subjects S05, S07, S13, and S15 had the highest performance accuracy, while subjects S03 and S14 had the lowest performance accuracy. Compared to other subjects, the proposed model's overall accuracy for subjects S03 and S14 is slightly lower. The significant aspect is the individual differences in multidimensional neural EEG signals [47], which limit the model's generalizability, which might be one of the reasons.

The confusion matrix in Figure 5 exhibits considerable values across the diagonal from left top to right bottom and lower values of the diagonal, emphasizing the significance of the proposed model's uniform imagined speech decoding of all vowels and words. The proposed model's precision, recall, and f1-score reports revealed that the model is effective for multiclass decoding of imagined speech. The Cohen-Kappa result shows that the model has a strong level of agreement on the decoding of multiclass imagined speech, and the model adequately learns the underlying features from the input signals.

3.3. Comparison of State of the Art with Proposed Model. The proposed model result was compared to the present state of the art for classification or decoding imagined speech. The proposed model performs slightly better in comparison to the present state of the art in EEG-based imagined speech decoding in terms of multiclass decoding or classification accuracy. Table 4 shows the comparison findings. The proposed model was compared with the existing literature on CNN-based models, and it was observed that it has an efficacious influence on performance accuracy, although comparability is difficult as the differences in EEG signals occur because of variables in the collecting environment, such as different participants, instruments, and tasks.

3.4. Wavelet, Temporal-Spatial Information, and Convolution Network Significance. The proposed TCN-CNN integrated model incorporates time-frequency resolution signals as input, permitting the network to learn more about the temporal and spatial properties of the signals using both the convolution path and enhanced classification accuracy. Considering that EEG signals are nonstationary, trying to find the right mother wavelet for the time-frequency analysis technique is instrumental for strengthening the decoding model's performance, and wavelet db10 has the capacity to illustrate the signal information with a significant time-frequency resolution, thus employed in the signal processing.

The EEG signal contained high temporal and low spatial information, and both types of information were important for improving the performance of decoding and analysing EEG signals. Additionally, we observe that the model's learning potential is reduced, and its accuracy falls when only TCN filters are used to extract the signal's temporal information. Earlier research [19] on imagined speech recognition from EEG signals indicated that both these features included in the input EEG signals could be utilized in the decoding, and the convolution-based network could become compelling for information extraction. Earlier research also illustrates the potential for decoding performance in other types of BCI activities, such as motor imagery, with the help of both these features of the EEG signals using the convolution-based deep learning model. In the study [48], one-dimensional convolutions were employed to extract both these features from the EEG signal, and the accuracy was significantly improved. In another study [49], temporal one-dimensional convolution was used first, followed by spatial one-dimensional convolution, to extract both of these features, allowing discriminative feature learning to classify the signals. Another study [50] found that both features can help with recognition in a BCI system based on steady-state visual evoked potential, where temporal filters were combined with spatial filters to improve event detection accuracy in noisy signals.

The methodological consideration of mother wavelet db10 for the discrete time-frequency analysis and the deep learning framework learn the temporal features with one-

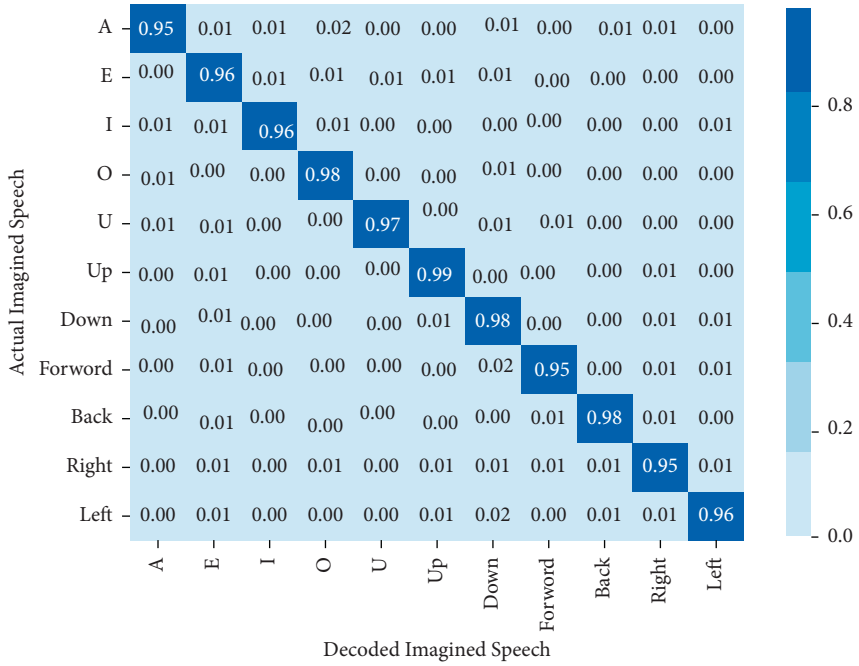

Figure 5: The model evaluation confusion matrix. The column represents the ground truth of the actual speech, and the row represents the predicted speech by the model. The diagonal from top-left to bottom-right represents the true positive for each speech label.

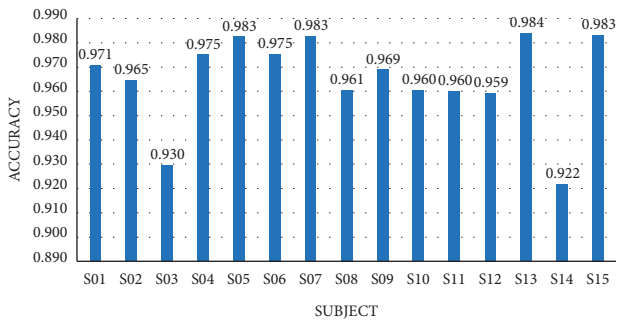

Figure 6: The histogram depicts the proposed model evaluation accuracy for the imagined speech classification of EEG signals across all subjects.

Table 4: Comparison of the proposed model with the state of the art.

Sl. no.	Authors	Model	Accuracy (%)
1	Saha et al. [13]	CNN + DAE	83.42
2	Cooney et al. [14]	CNN and TL	35.68
3	Panachakel et al. [15]	DNN	71.8
4	Panachakel et al. [21]	TL ResNet50	86.34
5	Li et al. [19]	Dilated CNN	54.31
6	Sarmiento et al. [20]	CNN	85.66
7	Proposed method	TCN + CNN	96.49

*In the absence of average classification accuracy in the paper, the overall accuracy was estimated as the mean of all subjects' accuracy.

dimension diluted casual convolution and spatial features with one-dimension convolution simultaneously, as well as consolidations of these features into classifications, which accelerates accuracy.

3.5. Limitations. Although the proposed method enhances multiclass classification performance accuracy, it has limitations due to the model's training adopting an offline supervised learning method. To generalize the approach to larger imagined speech decoding tasks, the proposed method requires a large corpus of labelled signals for model learning. However, gathering such a corpus of neural signals is extremely challenging. A reinforcement learning technique, like a biological agent's learning process, is the most likely answer, in which an artificial model learns by interacting with the environment through feedback-based processes. The artificial model decodes the neural imagined speech signals and interacts with the environment to determine the correctness of the decoding. The model learns from its environment, which provides feedback in the form of rewards for both accurate and erroneous decoding.

4. Conclusions

In this study, we demonstrated that using a deep learning integrated model built with CNN and TCN to extract the spatial and temporal activity of EEG data for multiclass decoding of imagined speech was one of the most effective ways. Using an open-access EEG dataset, we exhibited an improvement in model performance accuracy in the decoding of imagined speech vowels and words. We preprocessed the raw EEG signals with filtering and extracted features, as well as removed artifacts. The proposed deep learning model was trained and validated to decode the imagined speech by using preprocessed signals. The model achieved an overall accuracy of 96.49% for multiclass decoding of imagined speech. Although the proposed model attained higher overall accuracy in the EEG-based decoding

of imagined speech, the overall decoding error rate is a bit higher. Therefore, the decoding error rate needs to be further improved.

4.1. Future Direction for Optimization. The method used in the experiment was aligned with the specific imagined speech tasks and the specific set of subjects' neural EEG signals. This learning can be utilized in the generalization of other imagined speech tasks and neural EEG signals from other sets of subjects. This further research could be the convolutional neural network-based cross-task learning of imagined speech decoding.

Data Availability

The dataset used in this study is open access and available at https://doi.org/10.1117/12.2255697 from the earlier study and is cited at relevant places within the text as reference.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain–computer interfaces for communication and control," *Clinical Neurophysiology*, vol. 113, no. 6, pp. 767–791, 2002.

[2] E. W. Sellers, D. B. Ryan, and C. K. Hauser, "Noninvasive brain-computer interface enables communication after brainstem stroke," *Science Translational Medicine*, vol. 6, no. 257, 2014.

[3] C. S. DaSalla, H. Kambara, M. Sato, and Y. Koike, "Single-trial classification of vowel speech imagery using common spatial patterns," *Neural Networks*, vol. 22, no. 9, pp. 1334–1339, 2009.

[4] S. Deng, R. Srinivasan, T. Lappas, and M. D'Zmura, "EEG classification of imagined syllable rhythm using Hilbert spectrum methods," *Journal of Neural Engineering*, vol. 7, no. 4, Article ID 046006, 2010.

[5] M. D'Zmura, S. Deng, T. Lappas, S. Thorpe, and R. Srinivasan, "Toward EEG sensing of imagined speech," in *Proccedings of the Human-Computer Interaction. New Trends*, Berlin, Heidelberg, 2009.

[6] K. Mohanchandra, S. Saha, and G. M. Lingaraju, "EEG based brain computer interface for speech communication: principles and applications," *Intelligent Systems Reference Library*, vol. 74, pp. 273–293, 2015.

[7] L. Wang, X. Zhang, X. Zhong, and Y. Zhang, "Analysis and classification of speech imagery EEG for BCI," *Biomedical Signal Processing and Control*, vol. 8, no. 6, pp. 901–908, 2013.

[8] J. Kim, S.-K. Lee, and B. Lee, "EEG classification in a single-trial basis for vowel speech perception using multivariate empirical mode decomposition," *Journal of Neural Engineering*, vol. 11, no. 3, Article ID 036010, 2014.

[9] B. Min, J. Kim, H. j. Park, and B. Lee, "Vowel imagery decoding toward silent speech BCI using extreme learning machine with electroencephalogram," *BioMed Research International*, vol. 2016, Article ID 2618265, 11 pages, 2016.

[10] N. Yoshimura, A. Nishimoto, A. N. Belkacem et al., "Decoding of covert vowel articulation using electroencephalography cortical currents," *Frontiers in Neuroscience*, vol. 10, 2016.

[11] P. Sun and J. Qin, *Neural Networks Based EEG-Speech Models*, arXiv, Ithaca, NY, USA, 2017.

[12] C. H. Nguyen, G. K. Karavas, and P. Artemiadis, "Inferring imagined speech using EEG signals: a new approach using Riemannian manifold features," *Journal of Neural Engineering*, vol. 15, no. 1, Article ID 016002, 2018.

[13] P. Saha, M. Abdul-Mageed, and S. Fels, *Speak Your Mind! Towards Imagined Speech Recognition with Hierarchical Deep Learning*, arXiv, Ithaca, NY, USA, 2019.

[14] C. Cooney, R. Folli, and D. Coyle, "Optimizing layers improves CNN generalization and transfer learning for imagined speech decoding from EEG," in *Proceedings of the EEE International Conference on Systems, Man and Cybernetics (SMC)*, Bari, Italy, 2019.

[15] J. T. Panachakel, A. G. Ramakrishnan, and T. V. Ananthapadmanabha, *A Novel Deep Learning Architecture for Decoding Imagined Speech from EEG*, arXiv, Ithaca, NY, USA, 2020.

[16] C. Cooney, A. Korik, R. Folli, and D. Coyle, "Evaluation of hyperparameter optimization in machine and deep learning methods for decoding imagined speech EEG," *Sensors*, vol. 20, no. 16, p. 4629, 2020.

[17] M.-O. Tamm, Y. Muhammad, and N. Muhammad, "Classification of vowels from imagined speech with convolutional neural networks," *Computers*, vol. 9, no. 2, 2020.

[18] D. Pawar and S. Dhage, "Multiclass covert speech classification using extreme learning machine," *Biomedical Engineering Letters*, vol. 10, no. 2, pp. 217–226, 2020.

[19] F. Li, W. Chao, Y. Li et al., "Decoding imagined speech from EEG signals using hybrid-scale spatial-temporal dilated convolution network," *Journal of Neural Engineering*, vol. 18, no. 4, Article ID 0460c4, 2021.

[20] L. C. Sarmiento, S. Villamizar, O. López, A. C. Collazos, J. Sarmiento, and J. B. Rodríguez, "Recognition of EEG signals from imagined vowels using deep learning methods," *Sensors*, vol. 21, no. 19, 2021.

[21] J. T. Panachakel and R. A. Ganesan, "Decoding imagined speech from EEG using transfer learning," *IEEE Access*, vol. 9, pp. 135371–135383, 2021.

[22] G. A. Pressel Coretto, I. E. Gareis, and H. L. Rufiner, "Open access database of EEG signals recorded during imagined speech," in *12th International Symposium on Medical Information Processing and Analysis*, Tandil, Argentina, 2017.

[23] G. H. Klem, H. O. Lüders, H. H. Jasper, and C. Elger, "The ten-twenty electrode system of the international federation. The international federation of clinical neurophysiology," *Electroencephalography & Clinical Neurophysiology-Supplement*, vol. 52, pp. 3–6, 1999.

[24] J. F. Gao, Y. Yang, P. Lin, P. Wang, and C. X. Zheng, "Automatic removal of eye-movement and blink artifacts from EEG signals," *Brain Topography*, vol. 23, no. 1, pp. 105–114, 2010.

[25] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Networks*, vol. 13, no. 4–5, pp. 411–430, 2000.

[26] A. Ya. Kaplan, A. A. Fingelkurts, A. A. Fingelkurts, S. V. Borisov, and B. S. Darkhovsky, "Nonstationary nature of the brain activity as revealed by EEG/MEG: methodological, practical and conceptual challenges," *Signal Processing*, vol. 85, no. 11, pp. 2190–2212, 2005.

[27] C. J. Stam, "Nonlinear dynamical analysis of EEG and MEG: review of an emerging field," *Clinical Neurophysiology*, vol. 116, no. 10, pp. 2266–2301, 2005.

[28] D. Chen, S. Wan, J. Xiang, and F. S. Bao, "A high-performance seizure detection algorithm based on Discrete Wavelet Transform (DWT) and EEG," *PLoS One*, vol. 12, no. 3, Article ID e0173138, 2017.

[29] M. Mumtaz, M. Afzal, and A. Mushtaq, *Sensorimotor Cortex EEG Signal Classification Using Hidden Markov Models and Wavelet Decomposition*, IEEE, Piscataway, NJ, USA, 2018.

[30] H. Anila Glory, C. Vigneswaran, and V. S. Shankar Sriram, "Identification of suitable basis wavelet function for epileptic seizure detection using EEG signals," in *Proceedings of the First International Conference on Sustainable Technologies for Computational Intelligence*Jaipur, Rajasthan, India, 2020.

[31] H. Ines, Y. Slim, and E. Noureddine, "EEG Classification Using Support Vector Machine," in *Proceedings of the 10th International Multi-Conferences on Systems, Signals & Devices*, Hammamet, Tunisia, 2013.

[32] S. Bai, J. Z. Kolter, and V. Koltun, *An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling*, arXiv, Ithaca, NY, USA, 2018.

[33] A. van den Oord, S. Dieleman, H. Zen et al., *WaveNet: A Generative Model for Raw Audio*, arXiv, Ithaca, NY, USA, 2016.

[34] F. Yu and V. Koltun, *Multi-Scale Context Aggregation by Dilated Convolutions*, arXiv, Ithaca, NY, USA, 2016.

[35] S. Ioffe and C. Szegedy, *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*, arXiv, Ithaca, NY, USA, 2015.

[36] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using Convolutional Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Piscataway, NJ, USA, 2015.

[37] A. van den Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, and K. Kavukcuoglu, *Conditional Image Generation with PixelCNN Decoders*, arXiv, Ithaca, NY, USA, 2016.

[38] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proceedings of the International Conference on Machine Learning*, San Diego, CA, USA, 2010.

[39] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[40] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, *Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)*, arXiv, Ithaca, NY, USA, 2016.

[41] F. Pedregosa, G. Varoquaux, A. Gramfort et al., "Scikit-learn: machine learning in Python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011.

[42] F. Chollet, M. Ganger, E. Duryea, and W. Hu, *Keras*, vol. 4, 2015.

[43] M. Abadi, A. Agarwal, P. Barham et al., "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems," *Preliminary White Paper*, vol. 19, 2015.

[44] D. P. Kingma and J. Ba, *Adam: a Method for Stochastic Optimization*, arXiv, Ithaca, NY, USA, 2017.

[45] B. Azzerboni, M. Carpentieri, F. La Foresta, and F. C. Morabito, "Neural-ICA and wavelet transform for artifacts removal in surface EMG," in *Proceedings of the IEEE International Joint Conference on Neural Networks*, Budapest, Hungary, 2004.

[46] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[47] Z. Wan, R. Yang, M. Huang, N. Zeng, and X. Liu, "A review on transfer learning in EEG signal analysis," *Neurocomputing*, vol. 421, pp. 1–14, 2021.

[48] J. Chen, Z. Yu, Z. Gu, and Y. Li, "Deep temporal-spatial feature learning for motor imagery-based brain–computer interfaces," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 11, pp. 2356–2366, 2020.

[49] L. Yang, Y. Song, K. Ma, and L. Xie, "Motor imagery EEG decoding method based on a discriminative feature learning strategy," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 368–379, 2021.

[50] J. Jin, Z. Wang, R. Xu, C. Liu, X. Wang, and A. Cichocki, "Robust similarity measurement based on a novel time filter for SSVEPs detection," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.