Spectral Analysis of EEG Signals for Automatic Imagined Speech Recognition

Ashwin Kamble, *Member, IEEE*, Pradnya H. Ghare, *Senior Member, IEEE*, Vinay Kumar, *Senior Member, IEEE*, Ashwin Kothari, *Senior Member, IEEE*, and Avinash G. Keskar, *Senior Member, IEEE*

***Abstract*— Brain–computer interface (BCI) systems are intended to provide a means of communication for both the healthy and those suffering from neurological disorders. Imagined speech conveys users intentions. This article investigates the feasibility of spectral characteristics of the electroencephalogram (EEG) signals involved in imagined speech recognition. Eleven subjects were recruited to perform the speech imagination task. This article analyses the spectral features for binary and multiclass classification of imagined words in six different frequency bands (FBs). The 1-D EEG signals were converted into time–frequency representation (TFR) plots using smoothed pseudo-Wigner–Ville distribution (SPWVD) and classified using a convolutional neural network (CNN). In addition, the analysis was performed for subject-dependent, subject-independent, and leave-one-subject-out (LOSO) approaches along with the all-data approach. The proposed method achieved promising results in the gamma band with a binary classification accuracy of 82.04% ± 2.45%, 81.66% ± 4.93%, 78.97% ± 3.12%, and 81.04% ± 3.08% in all-data, subject-dependent, subject-independent, and LOSO approaches, respectively, and a multiclass classification accuracy of 51.44% ± 3.55%, 50.20% ± 1.35%, 49.93% ± 1.72%, and 50.42% ± 2.18% in all-data, subject-dependent, subject-independent, and LOSO approaches, respectively. Finally, the multiclass scalability in decoding the imagined words is investigated by increasing the number of classes from 2 to 15. The study's findings demonstrate that the EEG-based imagined speech recognition using spectral analysis has the potential to be an effective tool for speech recognition in practical BCI applications. The contribution of this article lies in developing an EEG-based automatic imagined speech recognition (AISR) system that offers high accuracy and reliability while also providing a noninvasive method for speech recognition.**

***Index Terms*— Automatic imagined speech recognition (AISR), convolutional neural network (CNN), data collection, electroencephalogram (EEG), smoothed pseudo-Wigner–Ville Distribution (SPWVD), spectral analysis, time–frequency representation (TFR).**

Manuscript received 30 May 2023; accepted 10 July 2023. Date of publication 1 August 2023; date of current version 11 August 2023. The Associate Editor coordinating the review process was Dr. Lihui Peng. *(Corresponding author: Ashwin Kamble.)*

This work involved human subjects in its research. Approval of all ethical and experimental procedures and protocols was granted by the Visvesvaraya National Institute of Technology (VNIT), Nagpur, India.

Ashwin Kamble, Pradnya H. Ghare, Ashwin Kothari, and Avinash G. Keskar are with the Electronics and Communication Engineering Department, Visvesvaraya National Institute of Technology, Nagpur 440010, India (e-mail: kashwin94@gmail.com; phghare@ece.vnit.ac.in; ashwinkothari@ece.vnit.ac.in; agkeskar@ece.vnit.ac.in).

Vinay Kumar was with the Electronics and Communication Engineering Department, Visvesvaraya National Institute of Technology, Nagpur 440010, India. He is now with the Electronics and Communication Engineering Department, Motilal Nehru National Institute of Technology Allahabad, Prayagraj 211004, India (e-mail: vinay.k@mnnit.ac.in).

Digital Object Identifier 10.1109/TIM.2023.3300473

I. Introduction

BRAIN–COMPUTER interfaces (BCIs) are being developed to help people with locked-in syndrome or paralysis to directly communicate with their external surroundings [1], [2]. Researchers are developing BCI systems, so that both patients and healthy individuals can communicate through their imagination. The conventional paradigms, such as event-related potentials and motor imagery, offer enhancement in communication but are restricted by the required stimuli and the number of classes for practical communication [3], [4]. Furthermore, these paradigms have been shown to be inefficient in terms of BCI (i.e., users who are unable to assert control when using a specific BCI framework), implying the need for a simpler paradigm [5]. For BCI communication to be effective, the user must be able to interact with the devices readily and intuitively. As a result, an intuitive framework that is simple to use and communicates user intent immediately is required [6].

Imagined speech is referred to as the imagination of words or things that a person wants to communicate without any articulation. The imagined speech is captured from the electroencephalogram (EEG) signals, which have evolved as the most common choice of researchers because of their noninvasive nature. Using EEG signals to recognize imagined speech directly from the brain could be helpful for direct BCI control with less training time [3]. In recent years, several attempts have been made to classify the EEG signals generated for imagined speech. These attempts were mainly focused on the classification of imagined vowels, imagined syllables, or a few imagined words [6], [7], [8], [9], [10], [11], [12], [13], [14]. The choice of the words in imagined speech application may increase the decoding performance.

For a human being, natural communication takes place through speech. The imagined speech application has the advantage of having an unlimited number of words, and each word represents a class [10], [13]. Unlike the other ways of communication, such as motor imagery, where the number of classes is restricted by the number of motor movements, imagined speech has constant discrimination between the words despite the number of words increasing in the classification problem [2]. These properties of imagined speech demonstrate its usefulness for BCI applications with more words.

The previous work demonstrates the reliability of the EEG signals in imagined speech applications but fails to attain actual communication along with the restrictions on a number of classes [15]. In recent years, several studies have

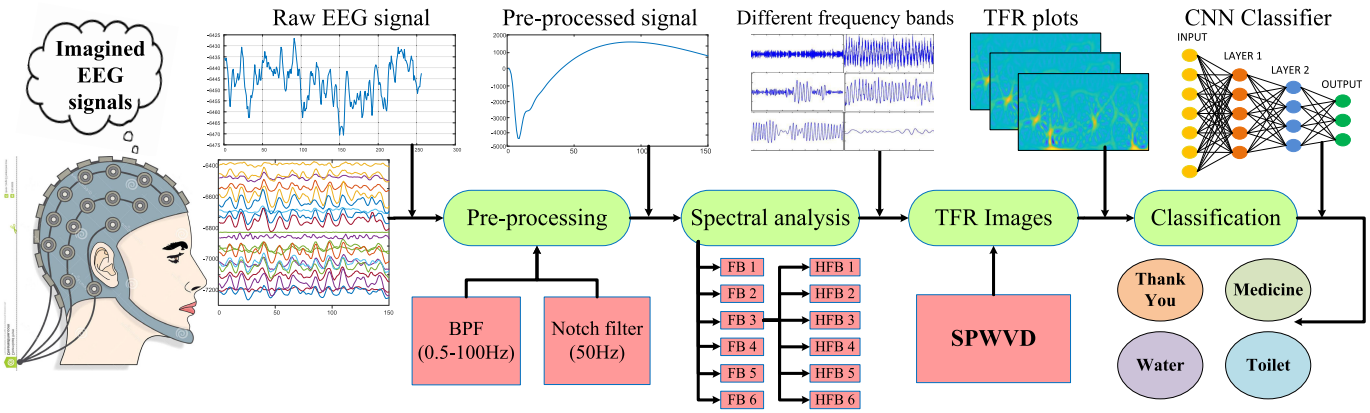

Fig. 1. Global architecture of the proposed model. The proposed model is divided into five parts: 1) collection of the dataset; 2) preprocessing of the EEG dataset; 3) dividing EEG signals into different FBs; 4) generating TFR images from 1-D EEG signal; and 5) classification of EEG signals using CNN.

investigated the use of spectral analysis techniques to gain insights into different aspects of brain activity and cognitive processes, such as Chikhi et al. [16] in assessing mental workload, Knyazev [17] in understanding neural dynamics and cognitive processes, and Zhang [18] in the event-related synchronization. These studies demonstrate the importance of spectral analysis in unraveling the complexities of EEG signals. Along with the spectral analysis, an extensive investigation of the time–frequency content of EEG signals captured for imagined speech applications is required.

Hence, this article presents the analysis of EEG signals for imagined speech applications for various frequency bands (FBs), such as delta, theta, alpha, mu, beta, and gamma. The objective of the study is to analyze the spectral features of EEG signals to develop automatic imagined speech recognition (AISR). The spectral features provide effective information about EEG signals that can help classify EEG signals into distinct categories. We hypothesize that the proposed method with spectral features can facilitate an improved imagined speech recognition than the conventional approaches available in the literature.

The global architecture of the proposed system is shown in Fig. 1. It shows the proposed model is divided into five parts: 1) collection of the dataset with a large number of words; 2) preprocessing of the EEG dataset; 3) dividing EEG signals into different FBs; 4) generating time–frequency representation (TFR) images from 1-D EEG signal; and 5) classification of EEG signals into binary and multiclass classification using convolutional neural network (CNN).

The objective of this article is to create a dataset with large number of words and develop an AISR system for recognition of imagined words. The TFR images will be generated from EEG signals and given input to deep-learning (DL)-based CNN algorithm. We hypothesize that the amalgamation of the DL-based CNN algorithm with the TFR-based methods can facilitate improved imagined speech recognition than the conventional machine learning-based approaches available in the literature. The major contributions of this article are as follows.

1) *Fresh Data Collection:* Fresh EEG data are collected that provides a valuable resource for investigating the feasibility and performance of automatic speech recognition.
2) *Spectral Analysis:* A detailed spectral analysis of the EEG signals is conducted to explore the frequency characteristics associated with imagined speech.
3) *TFR:* The TFR technique is employed to capture both temporal and spectral information from the EEG signals and investigate the dynamic changes in neural activity across different FBs during imagined speech.
4) *CNN-Based Automatic Feature Extraction:* The CNN is utilized for automatic feature extraction from the EEG signals and captures relevant spatial and temporal patterns in the EEG data.

The remainder of this article is arranged as follows. Section II describes the dataset acquisition, smoothed pseudo-Wigner–Ville distribution (SPWVD) for TFR, and CNN for classification. Section III presents the experimental results, and Section IV provides the discussion and conclusion of this article.

II. Materials and Methods

This section presents detailed information about the procedure used to collect the dataset for imagined speech recognition, the methods used for TFR, and classification.

A. Dataset Collection

Fifteen healthy subjects ($S1$–$S15$, all males, and mean age 28) took part in the imagined speech experiment. The study involving human participants was reviewed and approved by the committee formed at the institutional level at the Visvesvaraya National Institute of Technology (VNIT), Nagpur, India. All participants provided their written informed consent before the experiment. None of the subjects had a history of any neurological disease or language disorder. During the experiments, the subjects were instructed to pronounce these words internally in their minds and avoid any overt vocalization or muscle movements. The subjects were receiving instructions about the desired word based on visual cues from a computer monitor. The overall experimental setup used for each subject is shown in Fig. 2. All of our recordings took

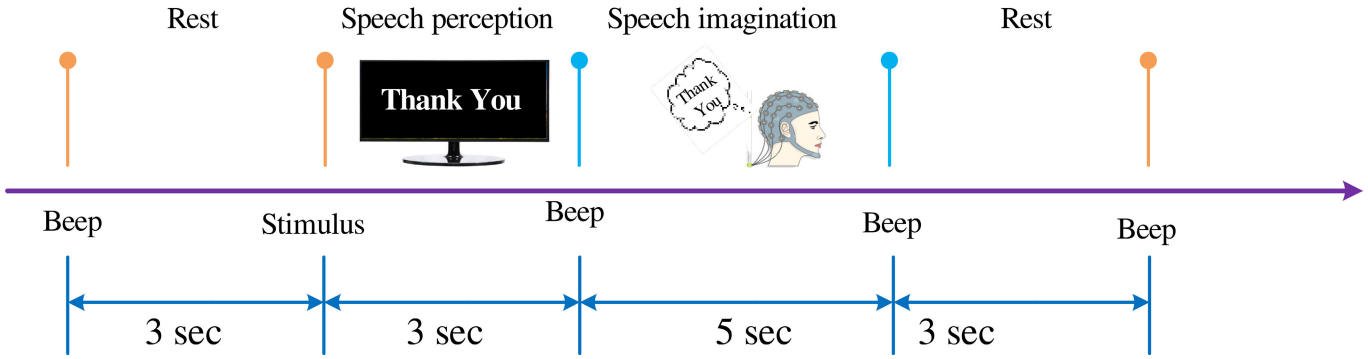

Fig. 2. Graphical representation of the imagined speech experimental paradigm. In each event, EEG signals for imagination of speech were recorded; 3 s after the first beep sound, the required word is presented on the screen for 3 s. The subject imagines the same word, which was shown on the LCD screen for 2 s.

Fig. 3. Experimental setup. The subjects involved in data collection process is wearing the EEG electrodes and looks at an LCD monitor a few inches away. A consent was obtained from the subject to show their picture.

place during the afternoon hours, in relatively peaceful settings in a room located at the New Academic Building, VNIT.

B. Experimental Setup

The EEG signals were acquired using a Medicaid amplifier system. The electrodes were placed according to the 10/20 international system. The data were recorded at 256 Hz. The subject is wearing the EEG electrodes and looks at an liquid crystal display (LCD) monitor a few inches away, as shown in Fig. 3. The monitor shows the word that the subject must imagine.

In this work, subjects performed the imagined speech of "*HELP*," "*LIGHT*," "*PAIN*," "*STOP*," "*YES*," "*NO*," "*RIGHT*," "*LEFT*," "*THANK YOU*," "*BACKWARD*," "*DOWN*," "*TOILET*," "*TELEVISION*," "*WATER*," and "*MEDICINE*." The words are selected keeping the requirement of the bedridden patients suffering from paralysis, speech disorder, and aphasia for patient-friendly communication. Each subject performed around 15 sessions, and each session corresponds to the imagination of the abovementioned 15 words. This indicates that each participant has imagined each word approximately 15 times. The words were presented in a random order in each session, so that the subject would not remember the sequence and, thus, avoid time-correlated artifacts.

The session started with a relaxation period of 3 s, followed by a screen time of 3 s, a relaxation period of 3 s, and an imagination period of 5 s. Screen time refers to the time the word was shown on the screen. Every interval was separated by a beep. At the beginning of the trial, the subject was also prompted with a visual cue indicating the desired word to be imagined. The cue lasted for 3 s. The subject was instructed to perform speech imagery with closed eyes and to keep going until the beep sound was heard. Finally, the trial ended with a rest period of approximately 3 s where no cue and no sounds were present.

C. Preprocessing

EEG data were preprocessed using the EEGLAB toolbox [19] in MATLAB 2022 (The MathWorks Inc. (2022), institute license version). A fifth-order Butterworth filter with a passband of 0.5–100 Hz is used to suppress the noise and artifacts [2]. In addition, a 60-Hz notch filter is used to remove the power line noise. Afterward, common average referencing was performed for rereferencing [3], [20].

D. Spectral Analysis

The preprocessed signal between 0.5- and 100-Hz frequency range consists of several FBs, each having significant information in each band. These FBs are FB1 (delta, 0.5–4 Hz), FB2 (theta, 4–8 Hz), FB3 (mu, 8–12 Hz), FB4 (alpha, 8–13 Hz), FB5 (beta, 13–30 Hz), and FB6 (gamma, 30–100 Hz), as shown in Table I. A fifth-order Butterworth bandpass filter was used to filter the EEG data into six FBs. This is performed to evaluate the effect of the imagination of speech in different FBs. Binary and multiclass classifications were performed for each FB.

E. Smoothed Pseudo-Wigner–Ville Distribution

The SPWVD is a widely used time–frequency analysis tool in signal processing and engineering applications [21], [22].

TABLE I
FBs FOR SPECTRAL ANALYSIS

Groups	Description
Frequency band (FB)	FB1 (0.5-4 Hz, Delta), FB2 (4-8 Hz, Theta), FB3 (8-12 Hz, Mu), FB4 (8-13 Hz, Alpha), FB5 (13-30 Hz, Beta), FB6 (30-100 Hz, Gamma)
High-Frequency band (HFB)	HFB1 (30-50 Hz), HFB2 (30-80 Hz), HFB3 (30-100 Hz), HFB4 (50-80 Hz), HFB5 (50-100 Hz), HFB6 (80-100 Hz)

It is a member of Cohen's class of bilinear time–frequency distributions, which aim to capture the time–frequency properties of nonstationary signals. The SPWVD is a modification of the WVD, which is a well-known bilinear time–frequency distribution. The WVD provides a high time–frequency resolution but suffers from cross terms. The SPWVD overcomes this limitation by smoothing the WVD with a Gaussian kernel, which suppresses the cross terms and enhances the energy concentration in the signal's time–frequency components.

The SPWVD can be defined mathematically as follows:

$$\mathrm{SPWVD}_x(t,\omega) = \int_{-\infty}^{+\infty} x\left(t+\frac{\tau}{2}\right)x^*\left(t-\frac{\tau}{2}\right)g(\tau)e^{-j\omega\tau}d\tau \tag{1}$$

where $x(t)$ is the signal under analysis, $g(\tau)$ is a Gaussian smoothing kernel, and x^* denotes the complex conjugate of x. The SPWVD estimates the energy density of the signal in the time–frequency domain and provides a smoothed representation of the WVD. The SPWVD has several advantages over other time–frequency analysis techniques. It provides a high time–frequency resolution and is insensitive to noise and nonstationarity. In addition, it has low computational complexity and can be easily implemented in digital signal processing systems.

F. Convolutional Neural Network

CNN is a type of DL algorithm that is widely used in computer vision applications for image classification and object detection [23], [24]. Proposed by LeCun et al. [25], CNN is comprised of self-optimized neurons. CNN is made up of layers upon layers of neurons that have been carefully trained for feature extraction and classification. CNN automatically learns to identify and categorize characteristics. They are inspired by the structure and function of the human visual system and are designed to learn hierarchical representations of visual information. With this, CNN has replaced conventional feature extraction and classification methods.

CNN consists of an input layer, multiple intermediate layers known as hidden layers, and an output layer. A hidden layer consists of convolutional layer, pooling layer, and fully connected (FC) layer. The convolutional and pooling layers consists of kernels of fixed size that extract different features from the images, and the classification task is carried out at FC layer.

The convolution operation of an image can be written as follows:

$$(M*N)(m,n) = \sum_{i,j} M(i,j)N(m+i,n+j). \tag{2}$$

TABLE II
HYPERPARAMETERS AND THEIR RANGE OF OPTIMIZED CNN MODEL FOR AISR SYSTEM USING SPWVD AND CNN

Sr No.	Hyperparameter	Range
1	conv_1_ kernels	96
2	conv_1_ kernel_size	3
3	conv_2_ kernels	128
4	conv_2_ kernel_size	7
5	conv_3_ kernels	216
6	conv_3_ kernel_size	7
7	conv_4_ kernels	256
8	conv_4_ kernel_size	9
9	dropout_1	0.30
10	dense_1_units	56
11	Learning rate	0.0001
12	Activation	Relu
13	Output layer	Sigmoid (binary), Softmax (multiclass)
14	Pool	Max (2, 2)
15	Batch size	64
16	Stride	1

The output of a convolution is a feature map that has a smaller dimension than that of the input image. For an input image of width W, height H, and number of channels H, with K filters of size r, the output of convolution can be calculated as follows:

$$W_{\text{new}} = \frac{W-K+2P}{S}+1 \tag{3}$$

$$H_{\text{new}} = \frac{H-K+2P}{S}+1 \tag{4}$$

where W_{new} and H_{new} are the width and height of the feature map, S is the stride, K is the kernel size, and P is the number of pixels used on each side for padding.

The next layer, pooling layer, is the subsampling layer that follows convolution layer. In pooling layer, creating downsampled feature maps is the primary goal. It lowers the number of factors and measurements. Overfitting can also be controlled by the pooling layer using max or mean functions. Pooling layer is followed by an FC layer. The FC layer takes a 2-D feature map and makes it 1-D. The result is transformed into probabilities in the softmax layer, and then, an item is classified using an algorithm.

III. RESULTS

The raw EEG data collected from 15 subjects were preprocessed to remove artifacts and noise. The preprocessed EEG data were then analyzed to identify the most powerful FBs containing significant information. TFR images were generated using SPWVD by utilizing EEG signals from all FBs. These TFR images were used as input for a CNN classifier. Hyperparameters of the CNN model are tuned using a recently developed Keras-Tuner library in Python (Python Software Foundation, version 3.9) to obtain a good combination of hyperparameters that highly improves the performance of the model. Table II presents the specific hyperparameters used for binary and multiclass classification of imagined words. The architecture of the optimized CNN is shown in Fig. 4.

A. Spectral Analysis in FBs

Since the dataset consists 15 words, classification can be performed for maximum of 15 classes. Consequently,

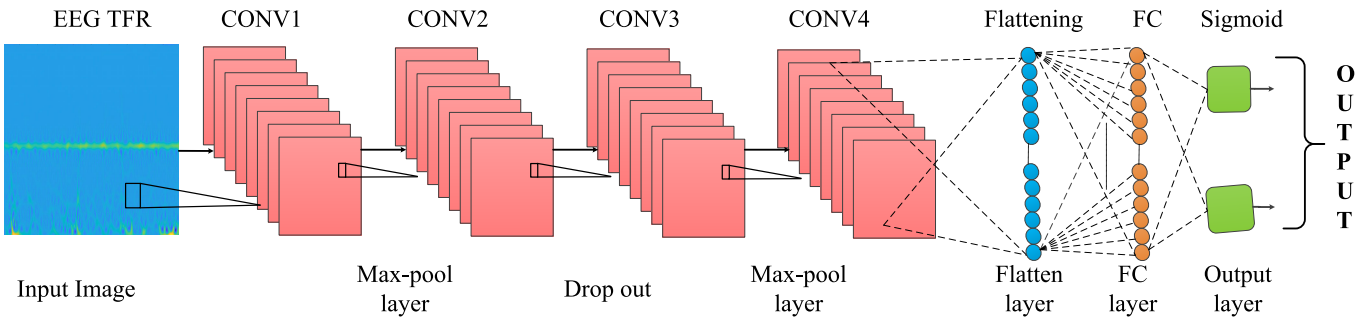

Fig. 4. CNN architecture used for binary classification. It has four convolutional layers (CONV: convolutional layer and FC: fully connected layer).

TABLE III
BINARY AND MULTICLASS ACCURACIES (%) IN DIFFERENT FBS

Frequency bands	Category	
	Binary	Fifteen class
FB1	79.82±3.19	49.98±2.21
FB2	82.01±3.03	48.48±1.89
FB3	80.98±2.45	49.98±1.67
FB4	81.86±2.12	48.91±3.30
FB5	80.68±2.04	51.44±3.55
FB6	**82.04±2.45**	**51.44±3.55**

we conduct binary classifications for various word combinations and multiclass classification (with 15 classes) on each of the six FBs, as shown in Table III. The optimized CNN model is employed to test binary and multiclass classification across the six FBs. The performance of the proposed system is evaluated using several performance metrics, including accuracy (ACC), $F1$ score, precision (PREC), recall (REC), Cohen's kappa (κ), and area under the curve (AUC).

1) Binary Classification Results: Binary classification was performed for all possible combinations of words, resulting in a total of 105 pairwise comparisons for each FB. The average binary classification accuracy, which represents the mean accuracy across all combinations, is presented in Table III. The table also highlights statistically significant differences in binary classification accuracy among the six FBs, with bold values indicating the highest scores. Among the FBs, FB6 achieved the highest binary classification accuracy of 82.04% ± 2.45%, while FB1 exhibited the lowest accuracy of 79.82% ± 3.19%. The binary classification experiment was repeated four times for each FB. Fig. 5 illustrates the training and validation curves for the binary classification of the word pairs "*THANK YOU*" and "*WATER*."

Fig. 6 provides classification performance for all 105 binary combinations involving the 15 words. The classification performance varied across different combinations of words. Among the 105 combinations, the pairing of "*LEFT*" and "*PAIN*" proved to be the most challenging, achieving an accuracy of 68%. The overall classification rates ranged from 68% to 86%. The combination of "*THANK YOU*" and "*WATER*" exhibited the highest accuracy of 86%, indicating the best discrimination between these two words. Notably, the words "*THANK YOU*," "*WATER*," "*TOILET*," and "*MEDICINE*" demonstrated improved discrimination when combined in other words. This observation suggests that subjects may have maintained better concentration during the imagination of these particular words, which is relevant in the application of the system for individuals with speech disabilities.

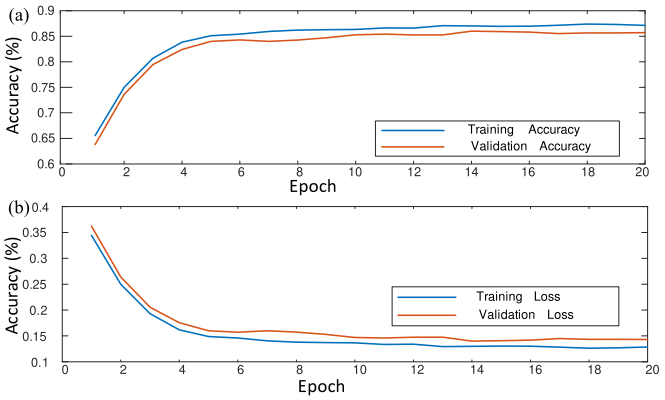

Fig. 5. Training and validation curve for binary classification of pair "*THANK YOU*" and "*WATER*." (a) Training and validation accuracy. (b) Training and validation loss.

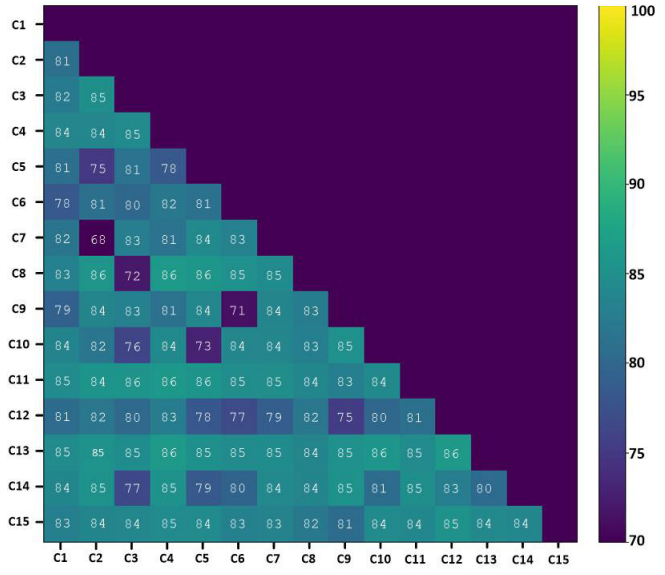

Fig. 6. Classification performance of 105 binary combinations of 15 classes. Binary classification was performed using the data collected from all the 11 subjects (C = class, $C1$: "*HELP*," $C2$: "*PAIN*," $C3$: "*YES*," $C4$: "*NO*," $C5$: "*STOP*," $C6$: "*LIGHT*," $C7$: "*LEFT*," $C8$: "*RIGHT*," $C9$: "*DOWN*," $C10$: "*BACKWARD*," $C11$: "*TOILET*," $C12$: "*THANK YOU*," $C13$: "*WATER*," $C14$: "*TELEVISION*," and $C15$: "*MEDICINE*").

2) Multiclass Classification Results: Multiclass classification accuracies were computed for each FB using data from all subjects. The mean and standard deviations of these accuracies were determined. Among the six FBs, FB6 achieved the highest mean accuracy of 51.44% ± 3.55% in the 15-class classification task, while FB2 exhibited the lowest accuracy of 48.48% ± 1.89%. The classification experiment was repeated four times for each FB to ensure robustness. Table III presents the statistically significant differences in classification accuracy among the six FBs.

3) Statistical Analysis: Differences in binary accuracy between six FBs (FB1–FB6) are investigated using t-test values and the corresponding confidence intervals. FB6 exhibited a significantly higher accuracy compared with FB1, with a mean difference of 2.39 ($t = 10.47$ and $p < 0.05$) and a 95% confidence interval of [1.947, 2.833]. However, the differences in accuracy between FB6 and the other bands (FB2–FB5) were not statistically significant, with the mean differences of 0.16 ($t = 0.76$ and $p > 0.05$; CI: [−0.256, 0.576]), 1.25 ($t = 5.59$ and $p < 0.05$; CI: [0.911, 1.589]), 0.51 ($t = 3.27$ and $p < 0.05$; CI: [0.204, 0.816]), and 1.40 ($t = 9.50$ and $p < 0.05$; CI: [1.110, 1.690]), respectively. These results indicate that the accuracy of FB6 is significantly higher than FB1, while the differences between FB6 and the other bands are not statistically significant. These results show that FB6 is a promising candidate for accurate signal recognition.

B. Spectral Analysis in HFBs

It was observed that the performance of the gamma band was best compared with that of other bands. Thus, gamma band is again divided into six different bands for the analysis, and we called these bands high-frequency bands (HFBs). A fifth-order Butterworth bandpass filter was used to filter the EEG data into six HFBs. These HFBs are HFB1 (30–50 Hz), HFB2 (30–80 Hz), HFB3 (30–100 Hz), HFB4 (50–80 Hz), HFB5 (50–100 Hz), and HFB6 (80–100 Hz), as shown in Table I.

1) Binary Classification: Binary classification accuracies with mean and standard deviations were calculated for each HFB using the data collected from all of the participants. The binary classification was performed for all the possible 105 binary combinations of 15 words available, and the average values are kept for each HFB. Table IV shows the statistically significant difference in binary classification accuracy for each of the six HFBs. The classification experiment was repeated four times for each of the FB. Among all the six HFBs, the highest binary classification accuracy of 82.04% ± 2.45% was obtained with HFB3, which is the same as FB6. The HFB3 is followed by HFB5, which obtained the binary classification accuracy of 80.79% ± 1.83%.

2) Multiclass Classification: Multiclass classification accuracies with mean values and standard deviations were calculated for each FB for the data collected from all the subjects, as shown in Table IV. The classification experiment was repeated four times for each of the HFB. Among all the six HFBs, the highest 15-class classification accuracy of 51.44% ± 3.55% was obtained with HFB3, which is the same as FB6. The HFB3 is followed by HFB5, which obtained the binary classification accuracy of 49.83% ± 2.38%.

TABLE IV
BINARY AND MULTICLASS ACCURACIES (%) IN DIFFERENT HFBS

Frequency bands	Category	
	Binary	Fifteen class
HFB1	79.84±2.79	48.57±2.01
HFB2	80.29±3.42	48.63±2.30
HFB3	**82.04±2.45**	**51.44±3.55**
HFB4	80.47±3.34	48.60±1.99
HFB5	**80.79±1.83**	**49.83±2.38**
HFB6	79.68±2.34	48.15±1.20

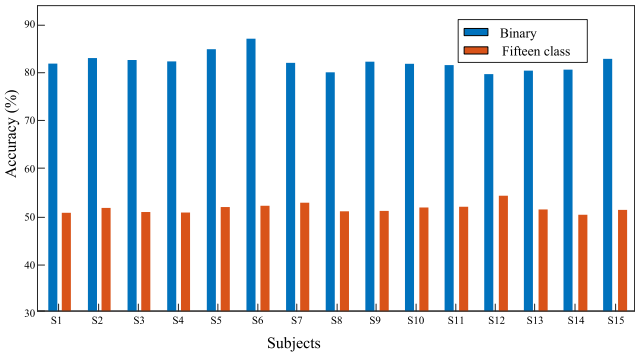

Fig. 7. Subject-dependent accuracies for binary and 15-class classification.

C. Performance Evaluation

The accuracies reported in this study were calculated using the "all-data" approach, which involves combining the data collected from all subjects and performing a tenfold cross validation to split the data into training and validation sets. While the subject-dependent approach is commonly used in BCI applications, the subject-independent approach is considered more practical. Therefore, the performance of the proposed method was assessed using both the subject-independent approach and the leave-one-subject-out (LOSO) approach, in addition to the subject-dependent approach.

In the subject-dependent approach, the classification is performed individually for each subject using a tenfold cross-validation technique. In the subject-independent approach, the performance of the proposed model is assessed by training the model with data collected from 12 randomly selected subjects, while testing it with data from the remaining three subjects. This process is performed once, ensuring that all subjects have an opportunity to serve as part of the testing set. In the LOSO approach, the model is trained using data from all subjects except one and then tested using the data from the excluded subject. This process is repeated for each subject individually, with each subject's data being withheld once for testing. The accuracies obtained from each iteration are averaged to derive the final accuracy.

Fig. 7 shows the accuracies received using subject-dependent approach for binary and 15-class classification. For subject-dependent approach, the highest binary accuracy achieved by $S6$ was 81.66% ± 4.93%, and the highest multiclass classification accuracy achieved by $S12$ was 50.20% ± 1.35%. For subject-independent approach, the highest binary

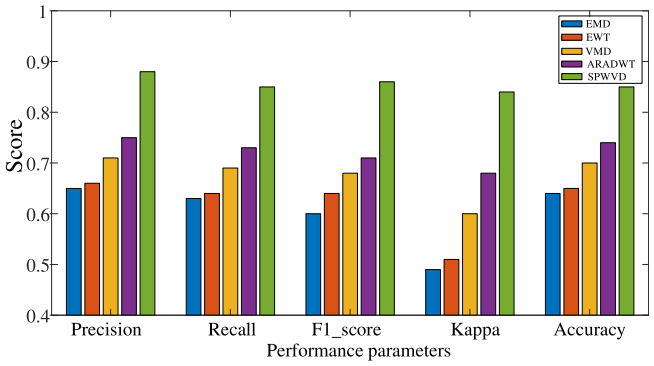

Fig. 8. Performance validation using the previously developed methods EMD, EWT, VMD, ARADWT, and SPWVD.

and multiclass classification accuracy achieved was 78.97% ± 3.12% and 49.93% ± 1.72%, respectively. Similarly, for the LOSO approach, the highest binary and multiclass classification accuracy achieved was 81.04% ± 3.08% and 50.42% ± 2.18%, respectively. Hence, it is clear that the proposed method performs well for all three approaches.

D. Comparative Analysis

The performance of the proposed model is compared against the previously developed models [1], [26], as shown in Fig. 8. All the models used for comparative analysis, such as empirical mode decomposition (EMD), variational mode decomposition (VMD), empirical wavelet transform (EWT), and adaptive rational dilation wavelet transform (ARADWT), employed all-data approach with tenfold cross validation. The performance is compared using performance evaluation metrics, such as $F1$ score, REC, PREC, κ, and AUC. Fig. 8 shows that the proposed method demonstrates superior performance compared with previously developed methods across multiple evaluation metrics, including $F1$ score, REC, PREC, κ, and AUC. The results consistently show higher values for these metrics, indicating the improved accuracy, completeness, and balance of our approach in capturing the desired performance aspects. These findings suggest that our method presents a significant advancement in achieving robust and reliable results in the context of the studied task.

Fig. 9 shows the region of convergence (ROC) plots for the binary classification. The optimized CNN reported the highest AUC of 0.86 followed by ARADWT (AUC = 0.82), VMD (AUC = 0.77), EWT (AUC = 0.72), and EMD (AUC = 0.70). Fig. 9 shows that CNN can be used for an imagined speech recognition application.

E. Expanding Number of Classes

As we have 15 words of data available to us, the classification can be performed for a maximum of 15 classes. Fig. 10 shows the confusion matrix for the 15-class classification. The true positive rate of "*HELP ME*" was the highest at 0.84 followed by "*WATER*," 0.77; "*PAIN*," 0.74; "*MEDICINE*," 0.71; and "*DOWN*," 0.65. The "*NO*" had the lowest true positive rate of 0.51.

IV. Discussion and Conclusion

The EEG-based BCI for imagined speech recognition has applications for both healthy and disabled people, such as those with speech production and speaking disabilities, neuromuscular illness, and other diseases [14], [15], [27]. In addition, this will enhance the quality of rehabilitation and neurology. This article focuses on the significant characteristics of EEG signals for imagined speech applications, such as spectral analysis, and the scalability of multiclass classification. In this article, the term "automatic" refers to the ability of the proposed model to recognize imagined speech directly from EEG signals without manual feature extraction. The experiments conducted in this study were performed offline, focusing on evaluating the model's performance using prerecorded EEG data.

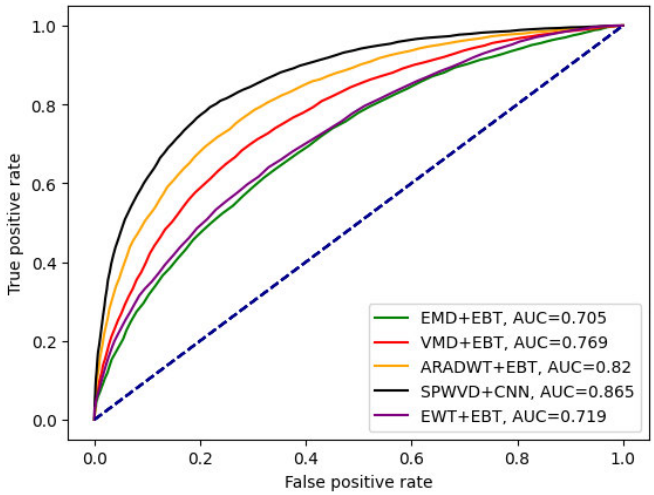

Fig. 9. ROC curve for binary classifications using EMD, EBT, EWT, VMD, ARADWT, and SPWVD.

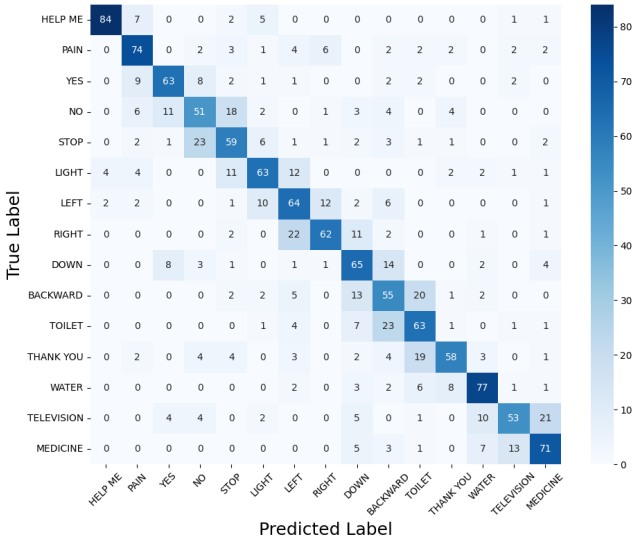

Fig. 10. Confusion matrix for 15-class classification using EEG signals from gamma band.

A. Superior Performance in the FB

The EEG signals of the imagined speech are divided into six FBs, and the decoding performance was observed. It has been observed that the gamma band has shown a remarkable enhancement in decoding performance. It is known that the

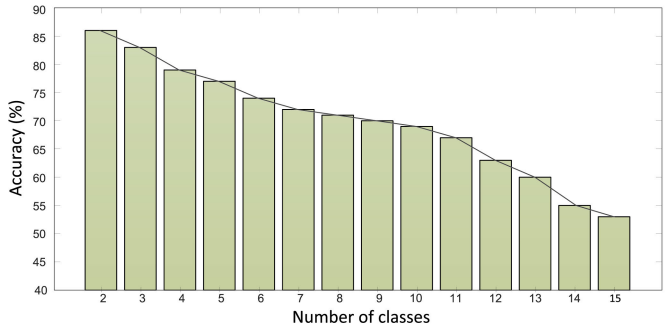

Fig. 11. Drop in multiclass classification accuracy with an increase in the number of classes.

HF activity shown by the gamma band can help with natural language processing [28]. When compared with any other spectral characteristic throughout the whole brain, the gamma band has shown considerable improvement in the classification of imagined speech. The gamma band was again divided into six HFBs, and the performance was evaluated for six different HFBs. It was observed that HFB3 had shown the best performance among the other HFBs.

Furthermore, the HFB is evident during cross-modal memory and sensory processing of objects and sounds [29]. The mental arousal of the pregiven stimulus is involved in the imagined speech, and the memory recall process may be represented in the HF band [6]. Although the 0.5–100-Hz band encompasses the HF range, selecting only this range resulted in substantially better performance [28].

B. Robustness in Expanding the Number of Classes

Classifications were performed from binary to 15-class classifications for different word combinations. The chance-level accuracy, denoted by the letter C, is defined as follows for a classification problem with N classes:

$$C(\%) = \frac{1}{N} * 100. \tag{5}$$

The drop in the classification accuracy with the increase in the number of classes is shown in Fig. 11. The average binary classification accuracy was 86%, while the 15-class classification accuracy was 53.5%. The results show that, though the accuracy drops with an increase in the number of classes, the accuracy is better compared with that of chance level [30], [31].

C. Conclusion

This article presents the effectiveness of spectral features for imagined speech recognition. This analysis allowed us to gain insights into the neural correlates of different speech-related processes, contributing to our understanding of imagined speech recognition. To explore the possibility of increasing number of classes, a dataset with a large number of words is created. The choice of words plays a crucial role in the imagined speech recognition system. The vocabulary is restricted to patients suffering from aphasia, paralysis, and speech disorder and their frequency of using these words for patient-friendly communication.

Although the proposed model shows better results, we believe that a dataset with more subjects and more words is required for the investigation. The presented dataset is biased toward the male subjects and age restricted. Furthermore, length of different words may contribute to understanding the imagined speech, and hence, special care needs to be taken while selecting words with different lengths. In future, we intend to add more words to the dataset, more trials will be recorded from each subject, and the effect of different brain regions on imagined speech recognition will be analyzed.

ACKNOWLEDGMENT

The authors acknowledge the support received from the Center of Excellence on Commbedded Systems, Department of Electronics and Communication Engineering, Visvesvaraya National Institute of Technology (VNIT), Nagpur, Maharashtra, India. The project grant provided the EEG data acquisition system used in this study that enabled us to carry out this research. All participants provided their written informed consent before the experiment.

REFERENCES

[1] A. Kamble, P. Ghare, and V. Kumar, "Machine-learning-enabled adaptive signal decomposition for a brain–computer interface using EEG," *Biomed. Signal Process. Control*, vol. 74, Apr. 2022, Art. no. 103526.

[2] 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, Feb. 2018, Art. no. 016002.

[3] X. Pei, D. L. Barbour, E. C. Leuthardt, and G. Schalk, "Decoding vowels and consonants in spoken and imagined words using electrocorticographic signals in humans," *J. Neural Eng.*, vol. 8, no. 4, Aug. 2011, Art. no. 046028.

[4] Y. Chen et al., "A high-security EEG-based login system with RSVP stimuli and dry electrodes," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 12, pp. 2635–2647, Dec. 2016.

[5] M.-H. Lee et al., "EEG dataset and OpenBMI toolbox for three BCI paradigms: An investigation into BCI illiteracy," *GigaScience*, vol. 8, no. 5, May 2019, Art. no. giz002.

[6] S.-H. Lee, M. Lee, J.-H. Jeong, and S.-W. Lee, "Towards an EEG-based intuitive BCI communication system using imagined speech and visual imagery," in *Proc. IEEE Int. Conf. Syst., Man Cybern. (SMC)*, Oct. 2019, pp. 4409–4414.

[7] B. Min, J. Kim, H.-J. Park, and B. Lee, "Vowel imagery decoding toward silent speech BCI using extreme learning machine with electroencephalogram," *BioMed Res. Int.*, vol. 2016, pp. 1–11, Dec. 2016.

[8] B. M. Idrees and O. Farooq, "Vowel classification using wavelet decomposition during speech imagery," in *Proc. 3rd Int. Conf. Signal Process. Integr. Netw. (SPIN)*, Feb. 2016, pp. 636–640.

[9] S. Deng, R. Srinivasan, T. Lappas, and M. D'Zmura, "EEG classification of imagined syllable rhythm using Hilbert spectrum methods," *J. Neural Eng.*, vol. 7, no. 4, Aug. 2010, Art. no. 046006.

[10] M. N. I. Qureshi, B. Min, H.-J. Park, D. Cho, W. Choi, and B. Lee, "Multiclass classification of word imagination speech with hybrid connectivity features," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 10, pp. 2168–2177, Oct. 2018.

[11] N. Hashim, A. Ali, and W.-N. Mohd-Isa, "Word-based classification of imagined speech using EEG," in *Proc. Int. Conf. Comput. Sci. Technol.* Cham, Switzerland: Springer, 2017, pp. 195–204.x

[12] E. F. González-Castañeda, A. A. Torres-García, C. A. Reyes-García, and L. Villaseñor-Pineda, "Sonification and textification: Proposing methods for classifying unspoken words from EEG signals," *Biomed. Signal Process. Control*, vol. 37, pp. 82–91, Aug. 2017.

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

[14] S. Martin et al., "Word pair classification during imagined speech using direct brain recordings," *Sci. Rep.*, vol. 6, no. 1, pp. 1–12, May 2016.
[15] P. Saha, M. Abdul-Mageed, and S. Fels, "SPEAK YOUR MIND! Towards imagined speech recognition with hierarchical deep learning," 2019, *arXiv:1904.05746*.
[16] S. Chikhi, N. Matton, and S. Blanchet, "EEG power spectral measures of cognitive workload: A meta-analysis," *Psychophysiology*, vol. 59, no. 6, p. e14009, Jun. 2022.
[17] G. G. Knyazev, "Cross-frequency coupling of brain oscillations: An impact of state anxiety," *Int. J. Psychophysiology*, vol. 80, no. 3, pp. 236–245, Jun. 2011.
[18] Z. Zhang, "Spectral and time-frequency analysis," in *EEG Signal Processing and Feature Extraction*. London, U.K.: Springer, 2019, pp. 89–116.
[19] A. Delorme and S. Makeig, "EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *J. Neurosci. Methods*, vol. 134, no. 1, pp. 9–21, Mar. 2004.
[20] M. Alhaddad, "Common average reference (CAR) improves P300 speller," *Int. J. Eng. Sci. Technol.*, vol. 2, no. 3, p. 21, Mar. 2012.
[21] E. P. de Souza Neto, M.-A. Custaud, J. Frutoso, L. Somody, C. Gharib, and J.-O. Fortrat, "Smoothed pseudo Wigner–Ville distribution as an alternative to Fourier transform in rats," *Autonomic Neurosci.*, vol. 87, nos. 2–3, pp. 258–267, Mar. 2001.
[22] S. Chaudhary, S. Taran, V. Bajaj, and A. Sengur, "Convolutional neural network based approach towards motor imagery tasks EEG signals classification," *IEEE Sensors J.*, vol. 19, no. 12, pp. 4494–4500, Jun. 2019.
[23] S. K. Khare, V. Bajaj, and U. R. Acharya, "SPWVD-CNN for automated detection of schizophrenia patients using EEG signals," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–9, 2021.
[24] S. K. Khare and V. Bajaj, "Time–frequency representation and convolutional neural network-based emotion recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 7, pp. 2901–2909, Jul. 2021.
[25] Y. LeCun et al., "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, Dec. 1989.
[26] A. Kamble, P. H. Ghare, and V. Kumar, "Optimized rational dilation wavelet transform for automatic imagined speech recognition," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–10, 2023.
[27] M. Matsumoto and J. Hori, "Classification of silent speech using support vector machine and relevance vector machine," *Appl. Soft Comput.*, vol. 20, pp. 95–102, Jul. 2014.
[28] M. Lee et al., "Network properties in transitions of consciousness during propofol-induced sedation," *Sci. Rep.*, vol. 7, no. 1, p. 16791, Dec. 2017.
[29] M. A. Kisley and Z. M. Cornwell, "Gamma and beta neural activity evoked during a sensory gating paradigm: Effects of auditory, somatosensory and cross-modal stimulation," *Clin. Neurophysiol.*, vol. 117, no. 11, pp. 2549–2563, Nov. 2006.
[30] Y.-J. Liu, M. Yu, G. Zhao, J. Song, Y. Ge, and Y. Shi, "Real-time movie-induced discrete emotion recognition from EEG signals," *IEEE Trans. Affect. Comput.*, vol. 9, no. 4, pp. 550–562, Oct. 2018.
[31] A. Sun, B. Fan, and C. Jia, "Motor imagery EEG-based online control system for upper artificial limb," in *Proc. Int. Conf. Transp., Mech., Electr. Eng. (TMEE)*, Dec. 2011, pp. 1646–1649.

Ashwin Kamble (Member, IEEE) received the B.E. degree from the MET's Institute of Engineering, Bhujbal Knowledge City, Nashik, India, in 2011, and the M.Tech. degree from the Veermata Jijabai Technological Institute (VJTI), Mumbai, India, in 2015. He is currently pursuing the Ph.D. degree with the Electronics and Communication Engineering Department, Visvesvaraya National Institute of Technology (VNIT), Nagpur, India.

His research interests include biomedical signal processing, machine learning, and deep neural networks.

Pradnya H. Ghare (Senior Member, IEEE) received the M.Tech. and Ph.D. degrees from the Electronics and Communication Engineering Department, Visvesvaraya National Institute of Technology (VNIT), Nagpur, India, in 2012 and 2017, respectively.

She is currently an Assistant Professor with the Department of Electronics and Communication Engineering, VNIT. She has developed some Internet-of-Things (IoT)-based products. Her research interests include wireless sensor networks (WSNs), body area networks, and the IoT.

Vinay Kumar (Senior Member, IEEE) received the bachelor's degree in electronics and communications engineering from Dr. A. P. J. Abdul Kalam Technical University, Lucknow, India, in 2006, and the M.Tech. and Ph.D. degrees from the Electronics and Communication Engineering Department, Motilal Nehru National Institute of Technology (MNNIT) Allahabad, Prayagraj, India, in 2010 and 2015, respectively.

He is currently an Assistant Professor with the Electronics and Communication Engineering Department, MNNIT Allahabad. His research interests include underwater, underground, and terrestrial wireless sensor network (WSN).

Ashwin Kothari (Senior Member, IEEE) received the B.E. and M.Tech. degrees in electronics engineering and the Ph.D. degree from the Visvesvaraya National Institute of Technology (VNIT), Nagpur, India, in 1994, 2005, and 2010, respectively.

He is currently working as a Professor with the VNIT, where he is one of the Coordinators for the Center of Excellence on Commbedded Systems: Hybridization of Communications and Embedded Systems under the World Bank-Assisted Project "Technical Education Quality Improvement Programme (TEQIP)." He has 26 years of experience in teaching. His research interests include image processing, antennas and wave propagation, electromagnetic, communication, and rough sets.

Avinash G. Keskar (Senior Member, IEEE) was born in Nagpur, India, in 1959. He received the B.E. degree (Hons.) from the Visvesvaraya National Institute of Technology (VNIT), Nagpur, in 1979, the M.E. degree (Hons.) from the Indian Institute of Science (IISc), Bengaluru, India, in 1983, and the Ph.D. degree from Nagpur University, Nagpur, in 1997.

He has 30 years of teaching and seven years of industrial experience. He is currently working as a Professor and the Head of the Department of Electronics and Communication Engineering, VNIT. His current research interests include computer vision, soft computing, embedded systems, and fuzzy logic.