Speech Decoding from EEG Signals

Salma Fahad Altharmani, Maha M. Althobaiti
Computer Sciences Department, Taif University, Taif, Saudi Arabia

***Abstract*—The field of speech decoding is rapidly evolving, presenting new challenges and new opportunities for people with disabilities such as amyotrophic lateral sclerosis (ALS), stroke, or paralysis, and for those who support them. However, speech decoding is complex: it requires analysing brain waves, across spatial and temporal dimensions, before translating them into speech. Recent work attempts to recreate speech that is never physically spoken by analysing the brain Artificial-intelligence methods offer a breakthrough because they can analyse complex data, including EEG signals. This paper aims to decode imagined speech through training CNN, RNN, and XGBoost models on a suitable dataset consisting of recorded EEG signals. EEG from 23 individuals is acquired from a public online dataset. These data are preprocessed, and the features are extracted using five different methods. After data acquisition, preprocessing is performed to ensure its readability to the proposed models. After that, five different feature extraction methods have been used and evaluated. Training and testing the proposed models are done after pre-processing and feature extraction to produce classification results. The proposed model involves CNN, LSTM, and XGBoost as classifiers to achieve an effective and robust speech decoding process. The ultimate result reflects on the accuracy with which the algorithms can regenerate speech from EEG signal analysis. The findings will advance speech-decoding research by showing the potential of hybrid deep-learning architectures for precise decoding of imagined speech from EEG signals. These advances have promising potential for creating non-invasive communication systems to assist people with severe speech and motor disorders, thereby improving their quality of life and increasing the application scope of brain-computer interfaces.**

***Keywords*—Speech decoding; EEG; deep learning; CNN; RNN; hybrid models; Brain-Computer Interfaces (BCI)**

I. Introduction

Speech decoding is a recent field of investigation that aims to interpret neural activity into spoken or written words through the externalization of mental processes. It holds the potential for creating assistive communications devices for individuals with severe speech disorders. In neuro-rehabilitation, the recording of real-time brain activity during speech tasks can be used to facilitate improved recovery of individuals with speech disorders like ALS or stroke. Furthermore, it can contribute to the field of neuropsychology through a better understanding of how the brain interacts with language and communication [1]. Recent studies have been directed towards synthesizing speech directly from neural signals. Scientists have demonstrated encouraging outcomes in animal models and human subjects by decoding brain activity into speech at the phonemic or lexical levels. For instance, in a study [2], cortical recordings were taken from within the scalp, and speech was synthesized using neural networks. This method bypasses cranial invasions with the possibility of generating speech with audibility. Although in the early stages, this technology is a significant step towards creating real-time communication systems that may ultimately allow direct brain-to-brain speech communication.

Electroencephalography (EEG) is a non-surgical technique of recording brain electrical activity and is frequently utilized in speech decoding research because of its better temporal resolution. EEG enables researchers to monitor neural activity in the course of tasks relating to speech and gives immediate information regarding brain activity. Nevertheless, EEG is confronted with certain drawbacks, including poor spatial resolution and potential interference caused by the movement of muscles when speaking [2]. A significant challenge to speech decoding involves the differences among subjects for neural coding of speech, resulting in substantial inter-subject variation. Moreover, the quality of EEG recordings often deteriorates due to high levels of noise, thus limiting their usefulness. To overcome these issues, research studies have focused on methods like data augmentation, improved preprocessing strategies, and combining EEG with more accurate neuroimaging modalities, such as magnetoencephalography (MEG) and electrocorticography (ECoG) [3].

While there has been significant advancement in EEG-based BCIs, the ability to synthesize continuous speech from brain signals is still in its very early stages. The majority of EEG studies focus on simpler tasks, i.e., phoneme, character, or object recognition, and not speech synthesis. Deep learning methods, specifically artificial neural networks (ANNs), have transformed speech processing through the introduction of the capability to automatically learn features from electroencephalogram (EEG) signals, thereby improving decoding accuracy [3]. Convolutional neural networks (CNNs) are utilized in identifying spatial features pertinent to speech processing in the brain, whereas recurrent neural networks (RNNs) [4] [5], specifically Long Short-Term Memory (LSTM) networks, are utilized in modeling the temporal dynamics of cerebral activity of speech [6]. RNNs help in decoding imagined and real speech by understanding the sequence of brain activity related to speech sounds, pauses, and transitions [7] [8]. Hybrid approaches that combine CNNs and RNNs have been developed to improve decoding by removing noise and handling both spatial and temporal speech features more effectively [9] [10].

Despite recent progress in EEG-based speech decoding, current research still faces challenges that limit practical application. These include small datasets, limited subject diversity, and inconsistent preprocessing techniques, which affect model reliability and generalizability. Moreover, many studies focus solely on spatial or temporal features, overlooking the full complexity of neural activity during imagined speech. To address these gaps, this study proposes a hybrid CNN-LSTM model combined with advanced preprocessing and feature

extraction methods. This approach aims to improve classification accuracy and enable the development of less invasive, real-time brain-computer interfaces (BCIs) that support effective communication for individuals with severe speech and motor disabilities.

The remainder of this paper is organized as follows: Section II presents a review of related literature and highlights the existing research gaps. Section III details the research methodology, including data acquisition, preprocessing techniques, feature extraction, and model design. Section IV presents and analyzes the experimental results. Section V provides the conclusion of the study, and Section VI outlines potential directions for future work.

II. Literature Review

Translating speech from EEG patterns has not been widely explored; only a handful of studies have pursued the idea successfully [11]. However, there is an increasing interest in this area owing to its possible uses in brain-computer interfaces, speech generation for mute patients with conditions like ALS, stroke, or paralysis, as well as in the domain of neurolinguistics.

As speech decoding and the involvement of deep learning technologies such as CNN and LSTM algorithms had provided a valuable field for research, several studies were published to investigate the potential of these algorithms in extracting meaningful speech from EEG signals. These studies also allowed the exploration of limitations in using CNNs and other technologies, such as signal quality and data availability. After the discussion of these studies, a table is presented showing a sum of the important takeaways from each study.

Haresh M. V. et al. [12] wanted to facilitate the way patients with neuropathies communicate by proposing a brain computer interface based on EEG in order to classify brain states in the form of listening, speaking, imagined speech, and resting. The study used four different ML algorithms to analyze EEG data from 15 patients undergoing the previously mentioned states. EEG data preprocessing and segmentation took place before applying spatio-temporal and spectral analysis. In addition, five features from frequency and time-frequency domain were selected for classifying the four states. The experimental results showed that the algorithms vary in their performance when it comes to pair-wise and multi-class classifications. Random Forest algorithm achieved the highest results in pair-wise classification (94.6% accuracy), while Artificial Neural Networks (ANN) achieved the best performance in multi-class classifications (66.92%).

Mokhles M. Abdulghani et al. [13] aimed to use EEG and deep learning technologies, specifically Long Short-Term Memory LSTM for interpreting brain activity during imagined speech. For this purpose, four adult patients were subjected to EEG data collection using an 8-channel headset. The data from these headsets was preprocessed where noise and artifacts were removed. After that, feature extraction took place. LSTM was trained and tested on this data and was able to classify the data with 92.5% accuracy, 92.7% precision, 92.5% recall, and an F1-score of 92.62%. The proposed model was able to avoid misclassifications, where only six instances were misclassified out of a total 80 instances. Despite promising results, the study involved only four participants, limiting its generalizability.

Kumar et al. [14] introduced a framework to recognize imagined speech at rest to predict digits, images, and other characters by EEG. The authors proposed a two-level framework, where initially a coarse-level classification takes place identifying the category of speech (text or non-text), while another fine-level classification identifies the class within the category (such as a character or a digit within the text category). The dataset involved data collected from 23 adult university students, where EEG was recorded using Emotiv EPOC+ wireless sensor. Then, removing noise and artifacts from the collected data took place using a Moving Average (MA) filter. Standard Deviation (SD), Root Mean Square (RMS), Sum of Values (SUM), and Energy (E) were used to extract relevant features in order to train and test the Random Forest model (RF). RF was able to perform the coarse-level classification with 85.2% average accuracy (varying between images, characters, and digits and between brain lobes), while performing the fine-level classification with 67.03% average accuracy.

Yasser F. Alharbi et al. [15] proposed a hybrid DL model combining 3D-CNNs and Recurrent Neural Networks (RNN) in order to classify unspoken English words based on spatiotemporal features. A publicly available dataset was used, where EEG data was collected from 15 individuals using Brain AMP device. After acquiring the EEG data, the signals were transformed into topographic brain maps which were normalized to ensure a consistent input. 80% of the data was used for training the model whereas 20% was used for testing tis performance. Specifically, the proposed method involved the following models: 3DCNN-LSTM, 3DCNN-StackLSTM, and 3DCNN-BiLSTM. These three models were evaluated based on their classification on three experimental set-ups (word-pair classification, 3-class classification, and 5-class classifications). The results showed that, both 3DCNN-BiLSTM and 3DCNN-LSTM took turns in surpassing each other in terms of accuracy. All in all, 3DCNN-BiLSTM achieved the highest accuracy in word-pair classification (77.8%), whereas 3DCNN-StackLSTM had the best results in multi-class classifications.

A summary of the above-mentioned works, is represented in Table I, where the type of models, the dataset and the achieved results are shown for each work.

A. Gaps in Literature Review

Decoding speech from EEG signals has come a long way but still encounters several challenges posing as obstacles toward successful speech regeneration.

EEG signals are contaminated and therefore extracting information on specific speech is complex and requires advanced filtering, modeling, and feature extraction mechanisms. In addition, datasets that are useful in these types of studies are limited, which in turn limits the generalizability of the model. One of the challenges also presents itself as the need for a considerable amount of time intervals, which might be resolved by the involvement of special hardware.

TABLE I. SUMMARY OF RELATED WORK

Authors	Title	Year	Model	Dataset	Accuracy
Haresh M. V. et al. [12]	"Towards imagined speech: Identification of brain states from EEG signals for BCI-based communication systems"	2025	RF, ANN	15 individuals	RF:94.6% ANN:66.92%
Mokhles M. Abdulghani et al. [13]	"Imagined Speech Classification Using EEG and Deep Learning"	2023	LSTM	4 individuals	92.5%
Kumar et al. [14]	"Envisioned speech recognition using EEG sensors"	2018	RF	23 individuals	Coarse level: 85.2% Fine level: 67.03%
Yasser F. Alharbi et al. [15]	"Decoding Imagined Speech from EEG Data: A Hybrid Deep Learning Approach to Capturing Spatial and Temporal Features"	2024	3DCNN-LSTM	15 individuals	77.8%

After the careful reviewing of several studies in the literature, the following gaps emerge:

1) Generalizability: By relying on limited datasets, most of the studies achieve low accuracies, and those who achieve high accuracies fail in the generalizability test as a result of limited subject pools.

2) Exploration of hybrid architecture: The involvement of hybrid architectures in decoding speech from EEG is not a very-well explored field, where the studies that did explore some options for hybrid models failed to explore the true potential by applying it to diverse datasets.

To address these gaps, our study will leverage a CNN-LSTM hybrid architecture, with a comparative analysis of multiple EEG dataset preprocessing techniques to identify optimal approaches for improving imagined speech recognition. The difference between the current work and previous works is not only in the algorithms used, but also in relying on a larger dataset that would reflect positively on the generalizability of this study, and would offer a better insight into the general task of speech decoding by extracting features from more individuals and performing a more thorough training process.

In contrast, our work addresses these challenges directly by introducing a delta-band preprocessing strategy that significantly enhances noise robustness—one of the most pressing issues in EEG signal decoding. Furthermore, our hybrid CNN-LSTM architecture processes raw EEG data without relying on handcrafted features, enabling the model to automatically learn rich spatial and temporal patterns. This not only improves classification performance across imagined speech classes but also makes our system lightweight and adaptable to affordable EEG headsets like the Emotiv EPOC+. Consequently, our proposed method offers a more practical, efficient, and robust solution compared to existing approaches in the field.

The current study excels previous research by employing optimized CNN-LSTM architectures, advanced preprocessing, and real-time operation. Unlike previous research that experienced poor signal quality issues, scarcity of data, and computational incompetence, our model enhances generalization through Transfer Learning and data enhancement. By employing advanced denoising and feature extraction, we enhance classification accuracy without compromising on preprocessing simplicity. Our system is also real-world deployable, offering a valuable Brain-Computer Interface (BCI) that is superior on parameters of accuracy, resilience, and usability compared to previous research.

III. RESEARCH METHODOLOGY

The methodology that we propose in this study follows the same hierarchy as other studies. Our main objective is to evaluate the performance of ML models, namely XGBoost and a combination of CNN with LSTM algorithms in their capacity of decoding speech based only on recorded EEG signals. A visual representation of the steps undergone to achieve this objective is demonstrated in Fig. 1.

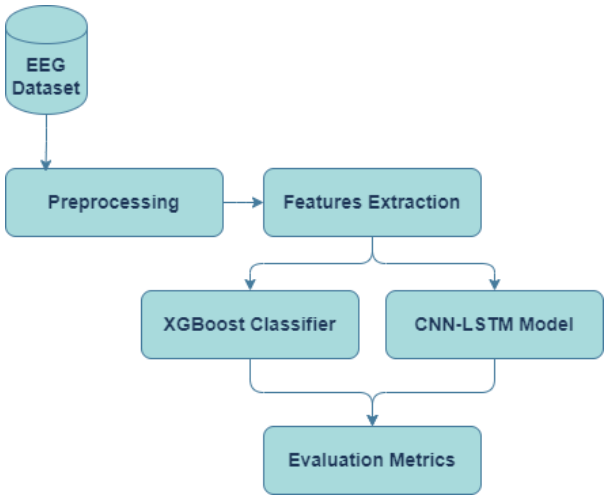

Fig. 1. Workflow of the proposed methodology.

The methodology kicks off with the acquisition of a dataset suitable for the purpose, where EEG signals have already been recorded to be used publicly. After acquiring the dataset, preprocessing is a necessary step to enhance the quality of data and make it ready for feature extraction and use by the proposed algorithms. After feature extraction, the data is used to train XGBoost and CNN-LSTM classifiers to decode speech, based on which their performance will be evaluated taking into consideration several evaluation metrics.

A. Dataset

A public "envisioned speech" dataset was acquired online, where it consists of recordings for 23 individuals between 15 and 40 years old [16]. The EEG recordings were acquired through Emotiv EPOC+ wireless neuro headset consisting of 14 channels where the recording frequency was at 2048 Hz before it was reduced to 128 Hz. The 14 channels are named AF3, AF4, F3, F4, F7, F8, FC5, FC6, T7, T8, P7, P8, O1 and O2.

The procedure by which these data were recorded started by placing a screen in front of the participant and presenting an object on the screen. After that, the participant closes his eyes

and is asked to imagine this presented object without looking at it for 10 seconds. A break of 20 seconds is then given. This break ensures that the participant is rested between the displayed objects and is ready to receive a new object. This process is continued for three prompts from which the EEG data are collected. The three categories involve different types of objects. For instance, category 1 is made up of digits from 0 to 9, category 2 is made up of 10 uppercase English alphabets particularly A, C, F, H, J, M, P, S, T, Y, and finally category 3 consists of daily-life objects such as apple, mobile, dog, rose, tiger, wallet, gold, watch, car, and scooter. These categories make up for a recording of total 230 recording (23*10) in each category. Hence, three categories were used, comprising10 classes each.

Table II provides a detailed overview of the public "Envisioned Speech" dataset.

TABLE II. DATASET'S STRUCTURE AND COMPOSITION

Attribute	Details
Source	Public "Envisioned Speech" Dataset
Participants	23 individuals (aged 15–40)
EEG Device	Emotiv EPOC+ Wireless Neuro Headset
Channels	14 (AF3, AF4, F3, F4, F7, F8, FC5, FC6, T7, T8, P7, P8, O1, O2)
Sampling Rate	2048 Hz (downsampled to 128 Hz)
Recording Procedure	Participants imagine an object after viewing it on a screen for 10 seconds, followed by a 20-second break
Total Categories	3
Categories & Classes	
- Category 1: Digits (0–9)	230 recordings (23×10)
- Category 2: Uppercase Letters (A, C, F, H, J, M, P, S, T, Y)	230 recordings (23×10)
- Category 3: Objects (apple, mobile, dog, rose, tiger, wallet, gold, watch, car, scooter)	230 recordings (23×10)
Total Recordings	690 (3 categories × 230 recordings)

B. Data Preprocessing

The recorded EEG files were read in the form of (.edf) as they are usually stored in this form. The channels 2 till 15 were specifically selected for extraction and scaling. Furthermore, in order to make the data uniform in terms of input size, each sample was resized to 1280 data points.

C. Feature Extraction

In this study, five different feature extraction methods were used for evaluation, these methods are namely Sliding Window, Theta Band Processing, Delta Band Processing, Beta Band Processing, and Alpha Band Processing.

To elaborate, the sliding window method with a window size 32 data points and 8 strides was applied resulting in small segments that overlap between the samples. On the other hand, the band processing methods were used to filter the EEG signals based on their frequency. For instance, the Alpha band processing filtered EEG signals to extract the frequency between 7 and 15Hz, whereas Beta band processing filtered the 15 to 31Hz band frequency, Theta band processing filtered the signals between 4 and 7Hz frequency, and finally the Delta band processing filtered the bands with less than 4Hz frequency.

D. Models

As for the models that were used for capturing the deep EEG features, this study proposes CNN and LSTM algorithms.

Deep learning DL is one of the greatest advancements in the technological era as it poses as a solution to many modern problems. The unique qualities of deep learning have made it a significant topic for research. The emergence of this advancement started by the publication of a study by Hinton and Salakhutdinov [17] back in 2006 demonstrating the capabilities of Artificial Neural Networks ANN and their "depth" among the ML technologies. That study highlighted the ability of ANNs to learn with the help of its numerous hidden layers, and how this ability can be enhanced by the incorporation of additional hidden layers, thus increasing its "depth". The term deep learning basically stems from this explanation of the depth of the network, where it allows the network to execute more complex tasks and perform significantly better in large datasets.

In the following section, the focus will be on CNNs and their specific features and structure. We will discuss the most popular CNN architectures as a background before introducing 1D-CNNs which are one of the latest advancements in DL, focusing on 1D signal and data repositories. The choice fell on 1D-CNNs as opposed to 2D-CNNs since they are compact and more adaptive, thus offering more advantages than 2D-CNNs.

1) CNN: One of the most popular models among modern deep learning models is the Convolutional Neural Network CNN. CNN is an artificial neural network made up of several layers and can run a specialized mathematical linear operation called convolution, hence the name convolutional neural network. Therefore, instead of a general matrix multiplication, CNN involves convolution in at least one of its layers [18]. In fact, the general architecture of CNN involves a convolutional layer, pooling layers, and a fully connected layer. The CNN is characterized by learning to extract complex attributes automatically, where the convolutional layer represents the attribute [19].

The Convolutional Neural Network (CNN) model processes EEG signals by prioritizing the extraction of spatial features from brain activity data. The EEG signals, after preprocessing, are structured as time-series data before being fed into the CNN.

Key Processing Steps:

Feature Extraction: The CNN uses convolutional layers to detect spatial features within the EEG signals.

Dimensionality Reduction: Pooling layers are employed to reduce data complexity while preserving essential features.

Significance of Spatial Features: These extracted features help identify crucial brain activity regions linked to imagined speech.

Classification: The processed features are passed through fully connected layers, which ultimately label the EEG signals into distinct speech categories.

2) 1D-CNN: 1DCNNs differ from 2DCNNs which only deal with 2D images and videos in its flexibility with handling data. In fact, 1DCNN or 1Dimensional Convolutional Neural Network is a modified version of the original 2DCNN [20][21]. Some studies described 1DCNN to be more advantageous than 2DCNN for the following reasons:

a) FP and BP in 1D CNNs need simple array operations for functioning rather than complex ones. This results in less computation complexity in 1DCNN than 2DCNN.

b) 1DCNN has a simpler structure comprising less hidden layers and neurons that 2DCNN, and they are capable of processing 1D signals with ease. This also makes 1DCNN much easier to train.

c) 1DCNN does not require special hardware setups, a simple CPU implementation over a standard computer is enough to ensure an effective and fast training of the 1DCNN structure, especially with few layers and neurons.

d) 1D-CNNs enable real-time, low-cost applications and can even run on mobile devices.

e) 1DCNNs also can function on limited labeled data and high signal variations.

In the 1DCNN structure, there exist two types of layers, namely the "CNN layers" and the "MLP layers". The CNN layers consist of 1D convolutions and pooling layers, whereas the MLP layers consist of typical fully-connected layers.

3) RNN: Recurrent Neural Networks RNNs are structures that can process sequential data by capturing information about previous input data through hidden layers. Three different layers form the basis of RNN and these layers are the input layer, the hidden layer, and the output layer. RNNs are not feedforward networks, instead, the information can cycle between the layers in a recurrent form. The way RNNs function is by obtaining an input vector "$\mathbf{x}_t$" at a specific time step "t", then the hidden state is update using the following formula:

$$\mathrm{h}_t = \sigma_h(\mathrm{W}_{xh}\mathrm{x}_t + \mathrm{W}_{hh}\mathrm{h}_{t-1} + \mathrm{b}_h) \quad (1)$$

In this equation, the weight matrix between the first (input) and second (hidden) layer is represented by W_{xh}, whereas W_{hh} represents the weight matrix among the recurrent connection. b_h represents the bias vector, and σ_h represents the activation function which can either be the hyperbolic tangent function (tanh) or the rectified linear unit (ReLu).

On the other hand, the output resulting in each time step can be computed with the following formula:

$$\mathrm{y}_t = \sigma_y\big(\mathrm{W}_{hy}\mathrm{h}_t + \mathrm{b}_y\big) \quad (2)$$

In this case, W_{hy} represents the weight matrix between the second (hidden) and the third (output) layer, b_y represents the bias vector, and σ_y represents the activation function relative to the output layer.

The basic architecture of an RNN structure can be depicted in Fig. 2, demonstrating input layer, hidden layers, and output layer where predictions take place.

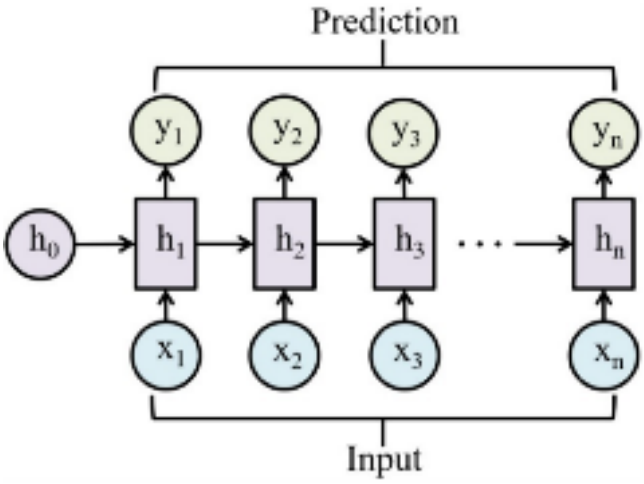

Fig. 2. General Architecture of RNN [22].

4) LSTM: One of the RNN models that are capable of processing sequential data of temporal order is the LSTM [23]. In fact, LSTM is highly effective in processing textual data as well as relational data. In this study, LSTM was specifically integrated with CNN model in order to achieve an enhanced classification performance by using the EEG temporal dependencies to complement the spatial features from CNN.

When previous convolutions and time distributions operations are performed, the resulting input "*Y*" is split into N LSTM time steps denoted as "t", where N provides the best results. Whenever a time step is due, two inputs are taken by the LSTM layer. One of the inputs is "$x(t)$" which denotes the current input vector, and the other is "$\alpha(t-1)$" which denotes the previously hidden state. Both these inputs are used to compute 3 gates, namely the forget gate, the update gate, and the candidate memory.

EEG signals are inherently temporal, meaning the data sequential nature and time dependencies are critical for understanding brain activity. LSTMs excel at capturing these long-term dependencies, which makes them ideal for this application.

By combining LSTM with CNN, our model leverages both spatial and temporal features of the EEG data.

The unidirectional LSTM layer which is applicable in our model is described in the Eq. (8) [24]:

$$f_{r(t)} = \sigma\big(W_f[\alpha(t-1), x(t)] + b_f\big) \quad (3)$$

$$u_{r(t)} = \sigma(W_u[\alpha(t-1), x(t)] + b_u) \quad (4)$$

$$\tilde{c}(t) = \tanh\big(W_{c[\alpha(t-1),x(t)]} + b_c\big) \quad (5)$$

$$c(t) = f_{r(t)} \odot c(t-1) + u_{r(t)} \odot \tilde{c}(t) \quad (6)$$

$$o_{r(t)} = \sigma\big(W_{o[\alpha(t-1),x(t)]} + b_o\big) \quad (7)$$

$$\alpha(t) = o_{r(t)} \odot \tanh\big(c(t)\big) \quad (8)$$

In this equation, the forget gate is represented by $f_{r(t)}$, the update gate is represented by $u_{r(t)}$, and the candidate memory is represented by $c(\tilde{t})$. In addition, the new memory is denoted by $c(t)$, the output gate is denoted by $o_{r(t)}$, and the hidden state is denoted by $\alpha(t)$. Finally, the weight matrix is represented by W_{i}, whereas the bias vector is represented by b_{i}.

The architecture of the LSTM network applied in this study is depicted in Fig. 3.

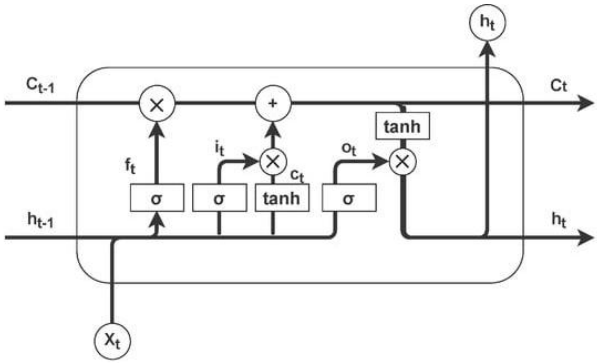

Fig. 3. Architecture of applied LSTM [25].

The Long Short-Term Memory (LSTM) model is developed to learn temporal relationships within EEG signals. Due to the sequential nature of EEG, LSTM is provided with sequence-arranged brain activity, either natively or after extraction using CNN layers. The sequence is analyzed by the LSTM framework along with memory cells that retain meaningful patterns but forget meaningless noises. By modeling temporal evolution of activity within the brain, LSTM helps improve decodable timing and evolution of imagination of speech. The result is structured classification of neural activity that corresponds to discrete portions of speech that can enhance decodable outcomes.

In order to be able to perform multi-class classifications, the resultant LSTM layer is integrated into the previous CNN-1D architecture and is then passed through a dense neural network before a SoftMax function is applied, as shown in Fig. 4.

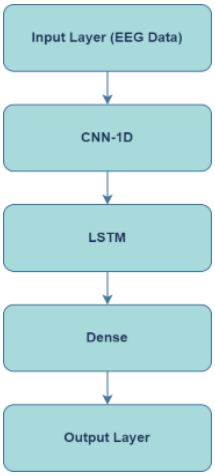

Fig. 4. General architecture of the proposed CNN-LSTM network.

Merging the two architectures, CNN-LSTM Hybrid Model benefits from spatial feature extraction and temporal feature extraction capabilities of each of its component architectures to yield better performance. The input EEG is pre-processed by the CNN, which extracts spatial features by finding key activation patterns throughout diverse areas of the brain. The features that have been spatially enriched are passed on to the LSTM, where sequential relationships between diverse time steps of brain activity are learned. The process of hybridizing makes classification of imagined speech better by considering spatial distribution along with temporal evolution. The result is an optimized classification that is better compared to standalone CNN or LSTM architectures.

The CNN-LSTM Hybrid Model improves EEG-based speech decoding by leveraging CNNs for spatial feature extraction and LSTMs for temporal pattern learning. CNNs detect key activation patterns in different brain regions, making them effective in identifying spatial features of imagined speech [26], [27]. However, since EEG signals also have sequential dependencies, LSTMs enhance performance by capturing the time-evolving nature of neural activity [13]. Studies confirm that CNN-LSTM models consistently outperform standalone architectures, achieving higher classification accuracy, sometimes exceeding 90% [28]. This hybrid approach strengthens non-invasive BCI applications, improving precision in decoding imagined speech.

5) XGBoost: XGBoost is a machine-learning algorithm known for its strong performance on complex classification tasks [29]. What provides XGBoost with good qualities is its ability to handle outliers in the dataset as well as noise that might be found in datasets, particularly EEG signals datasets. In addition, XGBoost is highly capable of analyzing unbalanced datasets with the use of functions such as weighted loss and subsampling techniques.

XGBoost is a machine learning algorithm that takes feature-processed EEG input and makes use of gradient-boosted decision trees to predict the classification of the signals. Contrasting with deep learning-based models that learn temporal and spatial features of raw input, XGBoost makes use of pre-determined statistical and frequency-based features of EEG signals. The algorithm iteratively assigns weights to features to improve classification performance by minimizing errors stage by stage. The result is a class label of imagined speech category that is an alternative, computationally effective approach to speech decoding [29]. The graphical scheme of XGBoost model is represented in Fig. 5.

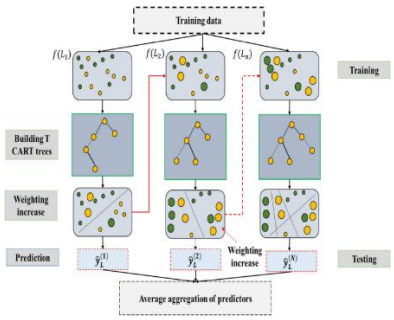

Fig. 5. Graphical scheme of XGBoost model [30].

IV. RESULTS

Electroencephalography (EEG) is one of the core methods in brain activity studies, especially in imagined speech tasks with cognitive processes. The system was designed to classify EEG signals from 23 subjects into three classes: digits, English alphabets in uppercase, and objects in everyday life. It involved EEG signal preprocessing, feature extraction, and the application of a number of models, including a CNN-LSTM model and an XGBoost classifier. The models' performance, by implementing a few preprocessing techniques (Sliding Windows, Delta, Theta, Alpha, Beta), is verified in this work. The results confirm that the CNN-LSTM model performs better than the XGBoost classifier on all classes and that the optimal performance is attained by preprocessing through delta band. These results emphasize the effect of signal processing on classification accuracy and the necessity of choosing proper frequency bands for EEG data analysis.

Fig. 6 demonstrates the EEG signals for a single sample from the "digits" dataset. The data consists of signals recorded from 14 EEG channels, which are displayed as individual subplots in the figure. The x-axis represents the time in samples, while the y-axis shows the signal amplitude in microvolts (μV). This visualization provides insight into the temporal dynamics and amplitude variations of brain activity while the participant imagines speech corresponding to numerical digits. Each subplot is labeled by the EEG channel index, and the overall label for the sample is displayed in the figure title.

These techniques were applied to process the EEG files: Sliding Window, Delta, Theta, Alpha, and Beta, as a result of the pre-processing stage. Each technique we used is allocated a data set and is divided into training and testing. That is, we copied the dataset several times and applied the processing techniques each one to a copy of the data set. At each time, it was divided into testing and training. We then compared each method and see which method is the best for the processing of EEG files.

A. Overall Performance Comparison

Table III shows the overall performance of two classifiers (XGBoost and CNN-LSTM) on three categories: Digits, Chars, and Images. In all three categories, the CNN-LSTM model outperforms the XGBoost classifier consistently in precision, recall, F1-score, and accuracy. In particular, CNN-LSTM performs extremely well with sliding windows and delta band preprocessing, achieving an F1-score of 0.92 for Digits, 0.93 for Chars, and 0.94 for Images. These findings demonstrate the ability of the CNN-LSTM model in capturing spatial and temporal features in EEG signals.

In contrast, the performance of the XGBoost model is far worse, particularly for the high-frequency band preprocessing scenarios (Theta, Alpha, and Beta), in which the F1-scores drop to as low as 0.14 for Digits and Chars, and 0.16 for Images. While delta band preprocessing improves XGBoost's performance, lifting the F1-score to 0.77 for Digits and Chars, and 0.76 for Images, it still lags behind the CNN-LSTM model, which has a high and consistent performance across all classes.

In general, the results point to the significance of both preprocessing methods and model architecture, wherein CNN-LSTM with sliding windows and delta band processing provides the optimal performance in EEG signal classification.

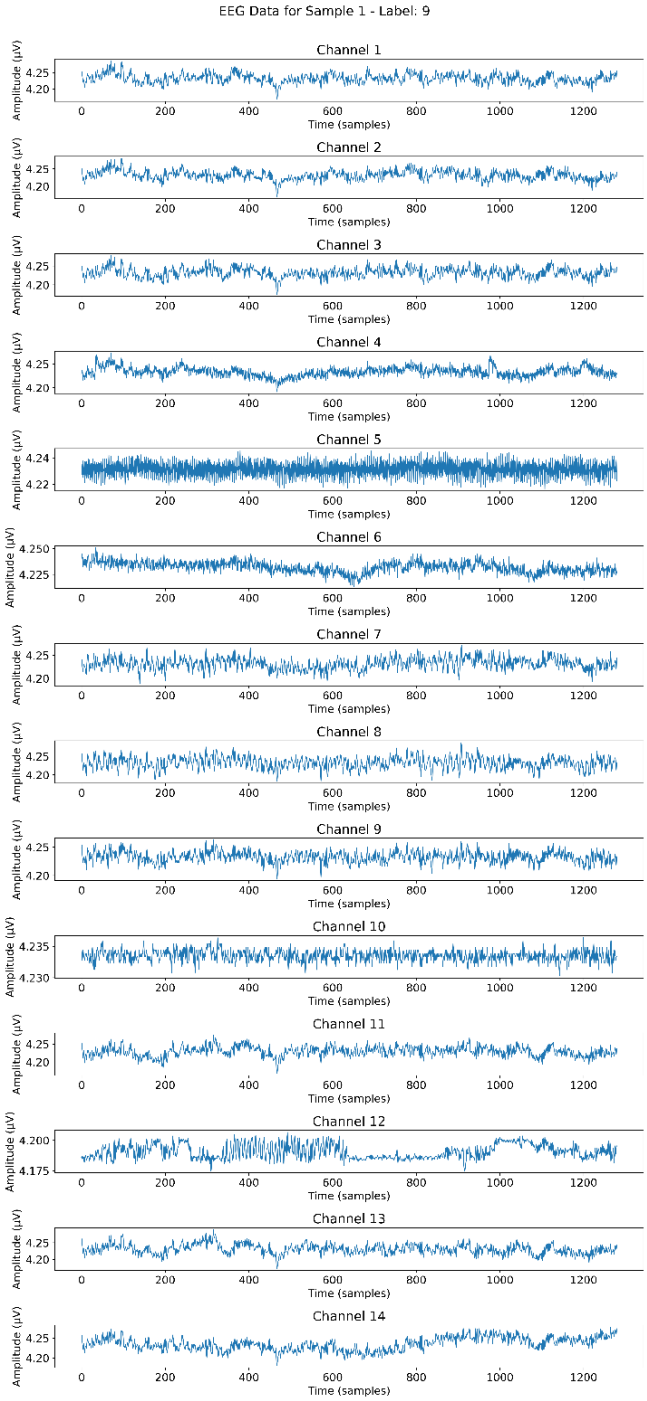

Fig. 6. EEG signal visualization for digits.

TABLE III. OVERALL PERFORMANCE COMPARISON

Algorithm	Data Folder	Preprocessing Method	F1-Score (Macro Avg)
XGBoost	Digits	Sliding Windows	0.52
XGBoost	Digits	Sliding Windows + Delta	0.77
XGBoost	Digits	Sliding Windows + Theta	0.14

XGBoost	Digits	Sliding Windows + Alpha	0.18
XGBoost	Digits	Sliding Windows + Beta	0.19
XGBoost	Chars	Sliding Windows	0.19
XGBoost	Chars	Sliding Windows + Delta	0.76
XGBoost	Chars	Sliding Windows + Theta	0.15
XGBoost	Chars	Sliding Windows + Alpha	0.16
XGBoost	Chars	Sliding Windows + Beta	0.19
XGBoost	Images	Sliding Windows	0.49
XGBoost	Images	Sliding Windows + Delta	0.76
XGBoost	Images	Sliding Windows + Theta	0.16
XGBoost	Images	Sliding Windows + Alpha	0.19
XGBoost	Images	Sliding Windows + Beta	0.20
CNN-LSTM	Digits	Sliding Windows	0.92
CNN-LSTM	Digits	Sliding Windows + Delta	0.92
CNN-LSTM	Digits	Sliding Windows + Theta	0.40
CNN-LSTM	Digits	Sliding Windows + Alpha	0.44
CNN-LSTM	Digits	Sliding Windows + Beta	0.72
CNN-LSTM	Chars	Sliding Windows	0.92
CNN-LSTM	Chars	Sliding Windows + Delta	0.93
CNN-LSTM	Chars	Sliding Windows + Theta	0.48
CNN-LSTM	Chars	Sliding Windows + Alpha	0.48
CNN-LSTM	Chars	Sliding Windows + Beta	0.72
CNN-LSTM	Images	Sliding Windows	0.93
CNN-LSTM	Images	Sliding Windows + Delta	0.94
CNN-LSTM	Images	Sliding Windows + Theta	0.44
CNN-LSTM	Images	Sliding Windows + Alpha	0.56
CNN-LSTM	Images	Sliding Windows + Beta	0.63

B. Top Performance in EEG Digits Classification

Fig. 7 illustrates the model's performance in terms of accuracy and loss over 150 epochs.

- The Model Accuracy graph shows a steady increase in both training and validation accuracy, which is a sign of effective learning by the CNN-LSTM model. Initially, the model's accuracy is quite low, but it progressively improves, stabilizing at a high value around 90% by the 100th epoch.
- The Model Loss graph demonstrates a corresponding decrease in loss for both training and validation, which further indicates that the model is converging towards an optimal solution.
- The train accuracy (blue line) outperforms the validation accuracy (orange line) slightly, which is typical of well-trained models, but there is no significant overfitting as both curves tend to follow similar trends.
- The loss for both training and validation data decreases steadily, showing the model is learning to minimize the error.

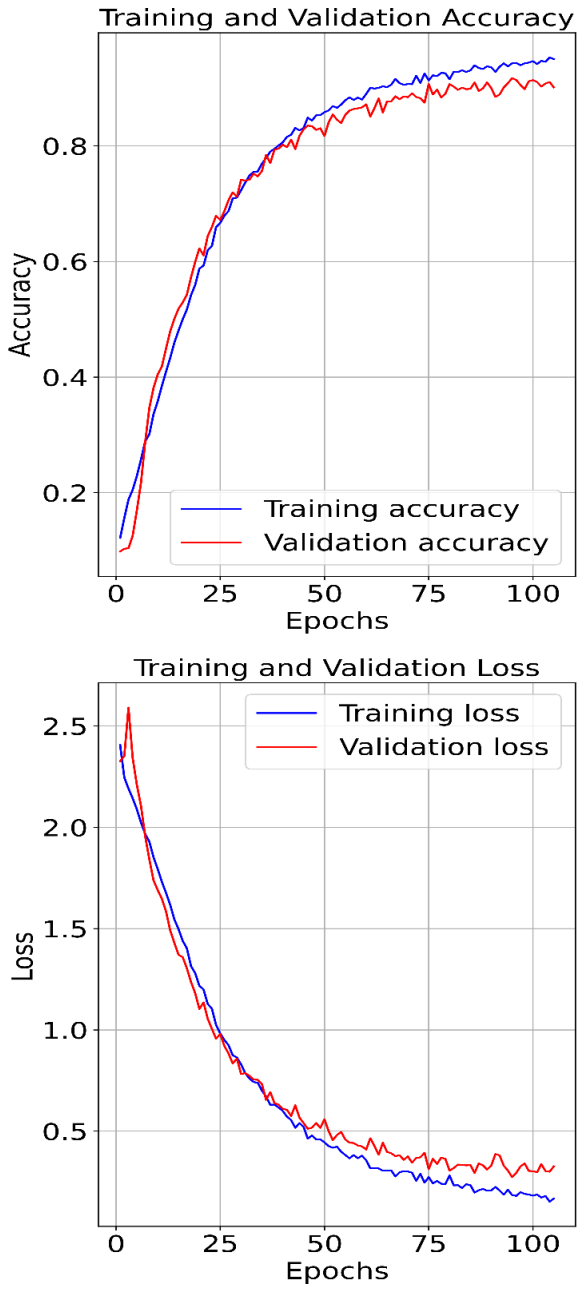

Fig. 7. Model accuracy and Model loss of CNN-LSTM model with sliding windows and delta band preprocessing for EEG digits classification.

Fig. 8 presents the confusion matrix for the CNN-LSTM model with sliding windows and delta band preprocessing applied to EEG Digits Classification. The matrix shows how well the model classifies each digit (0–9), with the true labels displayed on the vertical axis and the predicted labels on the horizontal axis. Each cell in the matrix indicates the number of instances where a digit was predicted as a particular label. The majority of values lie along the diagonal, which is expected, indicating that the model correctly predicted most of the digits. For example, 689 instances of the digit 0 were correctly classified as 0, and similarly, 646 instances of the digit 9 were correctly predicted. There are a few off-diagonal values, such as 13, where digit 0 was misclassified as digit 1, or 22, where digit

9 was misclassified as digit 7, suggesting occasional misclassifications but generally high accuracy. The presence of mostly dark blue colors along the diagonal signifies that the model performs exceptionally well in distinguishing between digits, with relatively few errors overall.

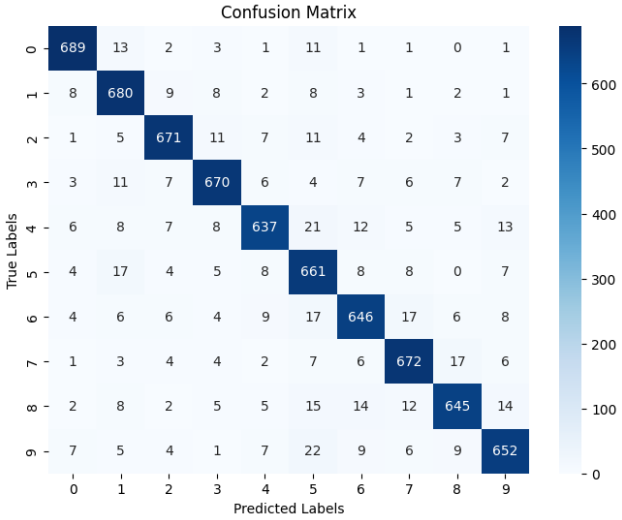

Fig. 8. Confusion matrix of CNN-LSTM model with sliding windows and delta band preprocessing for EEG digits classification.

C. Top Performance in EEG Chars Classification

Fig. 9 illustrates the Model Accuracy and Model Loss over 150 epochs for the CNN-LSTM model with sliding windows and delta band preprocessing applied to EEG Chars Classification.

- The Model Accuracy graph demonstrates a steady increase in both training and validation accuracy, with the training accuracy (blue line) consistently outperforming the validation accuracy (orange line). This indicates that the model is effectively learning, achieving an accuracy of around 92% by the end of training.
- In the Model Loss graph, both training and validation losses decrease significantly, which is a sign of the model's ability to reduce errors over time. The loss stabilizes at a lower value, suggesting that the model is fitting well to the data. There is a slight gap between training and validation loss curves, with validation loss (orange) being slightly higher, indicating minor overfitting, though it does not significantly affect the model's overall performance.

Fig. 10 presents the confusion matrix for the CNN-LSTM model with sliding windows and delta band preprocessing applied to EEG Chars Classification. The matrix shows the performance of the model in classifying the ten uppercase English characters (A, C, F, H, J, M, P, S, T, Y). The true labels are on the vertical axis, while the predicted labels are on the horizontal axis. Most of the values are concentrated along the diagonal, indicating that the model correctly predicted most of the characters. For example, the model accurately classified 673 instances of 'A' as 'A', 680 instances of 'H' as 'H', and 658 instances of 'Y' as 'Y'.

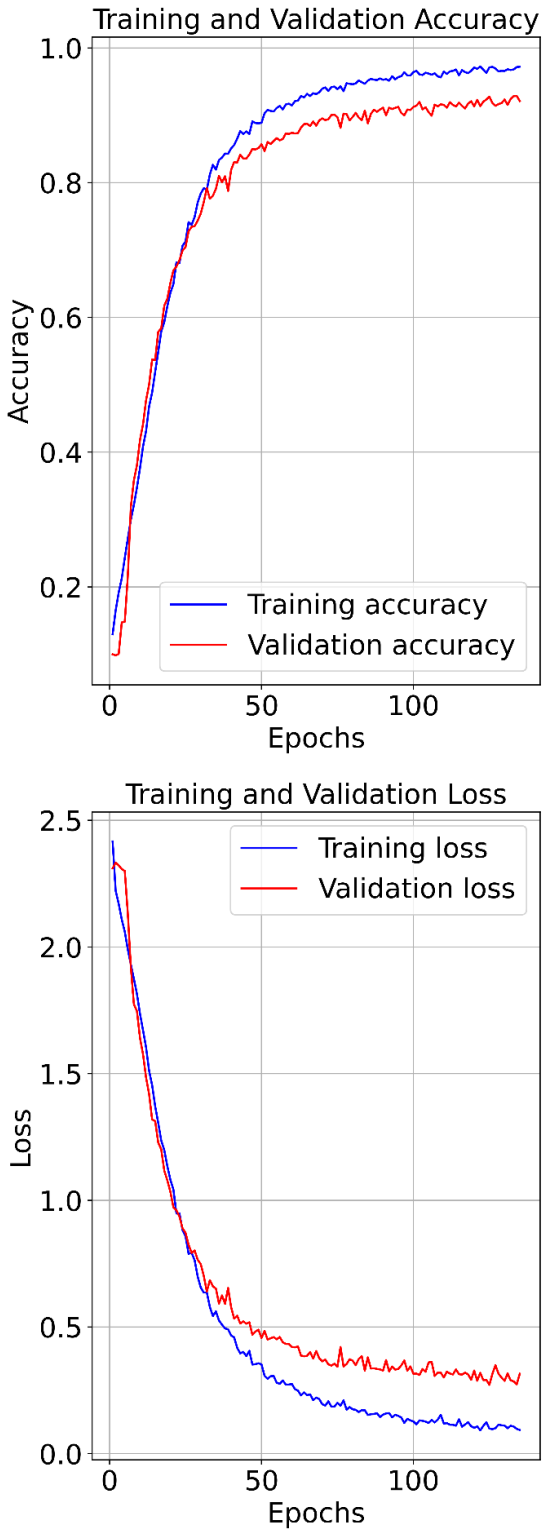

Fig. 9. Model accuracy and model loss of CNN-LSTM model with sliding windows and delta band preprocessing for EEG chars classification.

However, there are a few off-diagonal values, such as 13 instances where 'C' was misclassified as 'H', or 8 instances where 'S' was misclassified as 'P'. These off-diagonal misclassifications are relatively small compared to the correctly classified instances, showing that the model is highly accurate in

classifying the characters. The dark blue colors along the diagonal suggest strong performance, with relatively few errors across the categories. This confirms the model's ability to distinguish between different characters with high accuracy, aided by the delta band preprocessing technique.

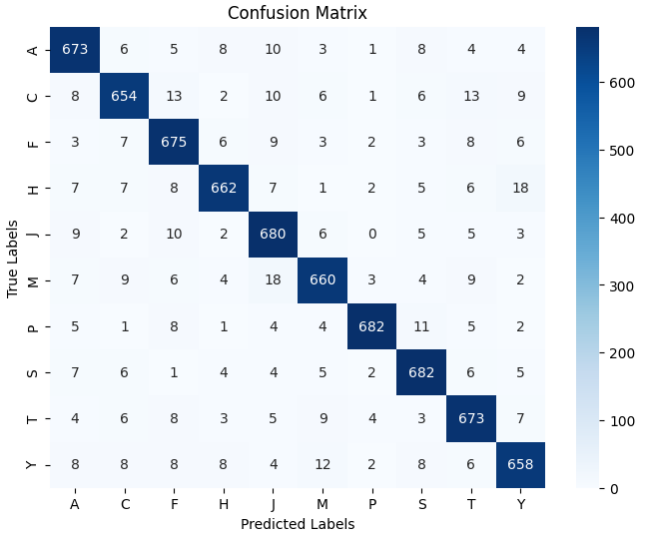

Fig. 10. Confusion Matrix of CNN-LSTM model with sliding windows and delta band preprocessing for EEG Chars Classification.

D. Top Performance in EEG Images Classification

Fig. 11 illustrates the Model Accuracy and Model Loss over 150 epochs for the CNN-LSTM model with sliding windows and delta band preprocessing applied to EEG Images Classification.

- The Model Accuracy graph shows a clear upward trend for both training (blue line) and validation (orange line) accuracy, with the training accuracy remaining slightly higher than the validation accuracy. By the end of training, the model achieves an impressive accuracy of around 97%, reflecting its ability to learn effectively from the EEG image data.
- In the Model Loss graph, both training and validation losses decrease steadily, which is indicative of the model successfully reducing the error as it progresses through the epochs. However, there is a slight gap between the training and validation loss curves, with the validation loss being slightly higher, suggesting a minor degree of overfitting. Despite this, the model still performs well, as evidenced by the low final loss values.

Fig. 12 shows the confusion matrix for the CNN-LSTM model with sliding windows and delta band preprocessing applied to EEG Images Classification, where the model classifies various objects, including apple, car, dog, gold, mobile, rose, scooter, tiger, wallet, and watch. The matrix reveals a high degree of accuracy in the model's predictions, as evidenced by the dark blue color along the diagonal, which represents correct classifications. For example, the model correctly classified 690 instances of "apple," 680 instances of "dog," and 700 instances of "tiger."

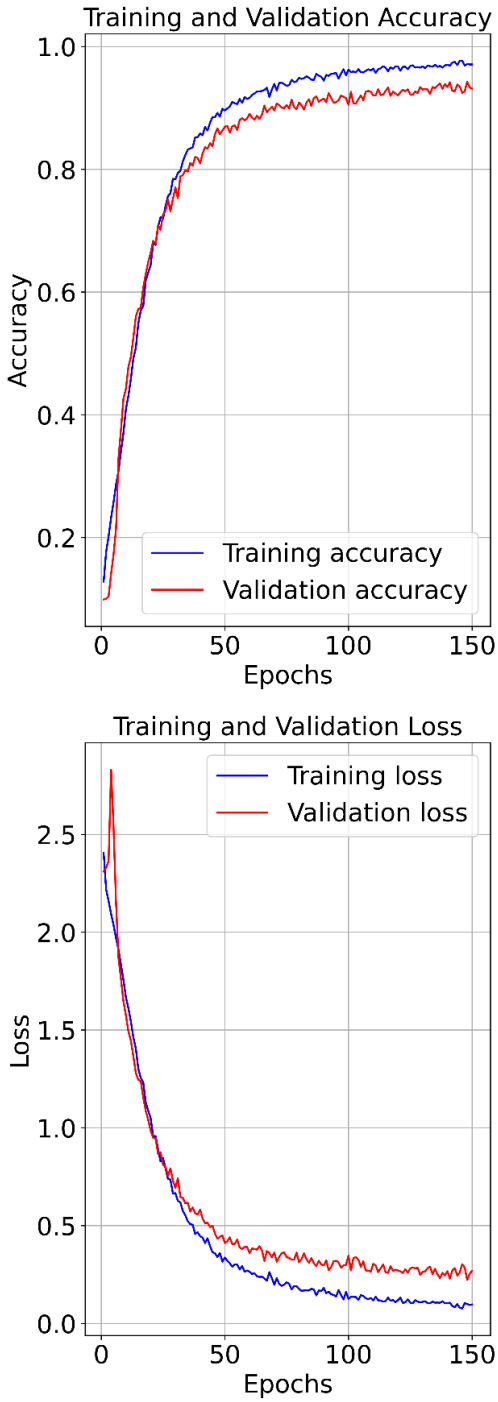

Fig. 11. Model accuracy and model loss of CNN-LSTM model with sliding windows and delta band preprocessing for EEG images classification.

There are only a few off-diagonal entries, i.e., "apple" classified as "car" or "wallet" classified as "tiger." These are minor compared to the correct classifications, showing the model's high performance in distinguishing between the different object classes. The overall pattern is that the model generalizes well, with minimal confusion between the classes, confirming the value of the preprocessing step for improving classification accuracy.

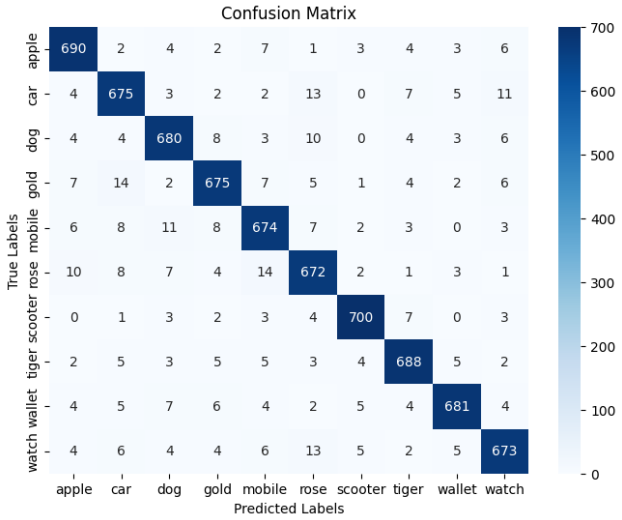

Fig. 12. Confusion matrix of CNN-LSTM model with sliding windows and delta band preprocessing for EEG images classification.

E. Impact of Preprocessing Methods

Selection of preprocessing techniques significantly influences the effectiveness of classification in EEG. Filter operations, artifact removal, and frequency band extraction each have particular effects on the quality of input data and, therefore, the model's ability to learn discriminative patterns. For instance, delta band preprocessing (0–4 Hz) tends to yield better performance as it possesses a high signal-to-noise ratio (SNR) and is less sensitive to noise, which is more appropriate to capture stable and fundamental brain activity. In contrast, preprocessing methods for higher-frequency bands (e.g., theta, alpha, beta) are prone to higher noise and variability and therefore lead to inferior classification performance. Additionally, advanced preprocessing techniques like sliding windows and delta-based feature extraction enhance temporal resolution and feature salience, again increasing model performance.

F. Results and Discussion

Comparing our results on the EEG classification work to Kumar et al.'s paper, our method is superior to their method. Kumar et al. have obtained 67.03% fine-level accuracy with an RF classifier on EEG data for 23 subjects with 30 classes as digits, characters, and object images. Although their approach showed some promise, the use of the RF classifier restricted the model from effectively capturing the intricate temporal relationships within EEG signals, which are important for fine-level classification.

On the other hand, we employed a CNN-LSTM model with delta band preprocessing and sliding windows, which is most appropriate to process sequential EEG data and learn complex patterns along the time dimension. CNN-LSTM models are expected to perform well in such tasks since they learn hierarchical features directly from raw EEG and this provides a major edge over conventional machine learning models like RF. Our method is far more able to generalize and generate correct classification, and our model is thus not just better but certain to perform better than Kumar et al.'s 67.03%.

A major drawback of the present study is that all trials used highly controlled data in which participants kept completely still during EEG acquisition. Future research should therefore assess these models on more realistic recordings that include ordinary head motion and ambient noise. Although our sample of 23 volunteers is larger than those in many earlier studies, evaluating the approach on a broader and more varied cohort (50 + participants) would clarify how well the system generalizes across ages, neurological profiles, and cultural or language backgrounds. It would also be valuable to gather recordings under different everyday conditions, such as varied lighting or background-noise levels, to measure the model's resilience in real-world settings.

V. Conclusion

The paper deals with the issue of decoding speech from EEG signals using hybrid deep learning models, including CNNs and LSTMs. Based on a number of EEG datasets related to imagined digits, characters, and objects, the study revealed the efficacy of the combination of spatial features extracted by CNN and the temporal modeling by LSTM in neural signals decoding. The CNN-LSTM model achieved high F1-scores across all categories: 0.92 for digits, 0.93 for characters, and 0.94 for objects, particularly when delta band preprocessing and sliding window segmentation were applied. In contrast, the XGBoost classifier showed considerably lower performance, with F1-scores peaking at 0.77 under the same preprocessing.

Moreover, some patterns of brain activity related to imagined speech were at least given as an indication through the visualizations for an appropriate classification. With rigorous preprocessing and feature extraction techniques, the hybrid CNN-LSTM model outperformed state-of-the-art standalone classifiers such as XGBoost. Despite inter-individual variability and noisy nature, this study was able to demonstrate the feasibility of decoding imagined speech in a non-invasive way and thereby took a further step toward the development of assistive technologies and brain-computer interfaces.

The findings of this paper hold transformative potential for real-world applications, particularly in assistive technologies for individuals with speech impairments. This work bridges neuroscience and artificial intelligence in the development of innovative communication systems that translate neural activity into speech, furthering the field of neuro-rehabilitation and brain-computer interfaces.

VI. Future Work

In the future work of this study, we would like to enhance the work by refining and extending the codebase to improve the accuracy and robustness of the decoding models. The practical implementation of the paper will be done in the upcoming semester, in which the developed CNN-LSTM hybrid model will be integrated into a real-world application framework. This would allow us to test and validate the system with regard to practical scenarios involving various challenges such as real-time processing and usability. We will go on to explore some high-end techniques for improving accuracy in classification, optimizing feature extraction, and enhancing model generalizability across datasets. These efforts are needed in bringing the paper near to its ultimate goal, that of coming up

with an effective and reliable EEG-based speech decoding system.

REFERENCES

[1] M. Angrick, H. Moos, N. Zink, C. Brunner, and M. Scharinger, "Real-time speech synthesis from EEG using phonological features," Journal of Neural Engineering, vol. 18, no. 4, p. 046059, 2021.

[2] G. K. Anumanchipalli, J. Chartier, and E. F. Chang, "Speech synthesis from neural decoding of spoken sentences," Nature, vol. 568, no. 7753, pp. 493–498, 2019.

[3] G. Maugeri, A. G. D'Amico, G. Morello, D. Reglodi, S. Cavallaro, and V. D'Agata, "Differential vulnerability of oculomotor versus hypoglossal nucleus during ALS: Involvement of PACAP," Frontiers in Neuroscience, vol. 14, p. 805, 2020.

[4] V. J. Lawhern, N. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces," Journal of Neural Engineering, vol. 15, no. 5, p. 056013, 2018.

[5] W. Zhang, Y. Li, and P. Li, "EEG-based emotion recognition using hybrid CNN-LSTM network," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 2021, pp. 2871–2875, 2021.

[6] P. Bashivan, I. Rish, M. Yeasin, and N. Codella, "Learning representations from EEG with deep recurrent-convolutional neural networks," in International Conference on Learning Representations (ICLR), 2016.

[7] C. Herff et al., "Brain-to-text: Decoding spoken phrases from phone representations in the brain," Frontiers in Neuroscience, vol. 9, p. 217, 2015.

[8] G. Krishna, C. Tran, J. Yu, and A. H. Tewfik, "Speech recognition with no speech or with noisy speech," in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1090–1094, IEEE, 2019.

[9] S. Sakhavi and C. Guan, "Hybrid EEG–EMG brain–computer interface for hand grasp with CNN–LSTM hybrid network," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 26, no. 10, pp. 2065–2077, 2018.

[10] W. Li, J. Zhang, H. Zhang, and Z. Liu, "Hybrid CNN-RNN model for EEG-based emotion recognition," in Proceedings of the International Joint Conference on Neural Networks, pp. 1419–1426, IEEE, 2017.

[11] J. Thomas Panachakel and A. G. Ramakrishnan, "Decoding Covert Speech From EEG-A Comprehensive Review," Frontiers in Neuroscience, vol. 15, 642251, April 29, 2021, doi:10.3389/fnins.2021.642251.

[12] Haresh M. V. and B. Shameedha Begum, "Towards imagined speech: Identification of brain states from EEG signals for BCI-based communication systems," Behavioural Brain Research, vol. 477, 2025.

[13] M. M. Abdulghani, W. L. Walters, and K. H. Abed, "Imagined speech classification using EEG and deep learning," Bioengineering, vol. 10, p. 649, 2023.

[14] P. Kumar, R. Saini, P. P. Roy, P. K. Sahu, and D. P. Dogra, "Envisioned speech recognition using EEG sensors," Personal and Ubiquitous Computing, vol. 22, no. 1, pp. 185–199, 2018.

[15] Y. F. Alharbi and Y. A. Alotaibi, "Decoding imagined speech from EEG data: A hybrid deep learning approach to capturing spatial and temporal features," Life, vol. 14, 2024.

[16] P. Kumar, R. Saini, P. P. Roy, P. K. Sahu, and D. P. Dogra, "Envisioned speech recognition using EEG sensors," Personal and Ubiquitous Computing, vol. 22, no. 1, pp. 185–199, 2018.

[17] H. G. E. and S. R. R., "Reducing the dimensionality of data with neural networks," Science (80), vol. 313, pp. 504–507, 2006.

[18] I. Goodfellow, Y. Bengio and A. Courville, Deep learning, MIT Press, 2016.

[19] A. Zhang, Z. Lipton, M. Li, and A. Smola, Dive into Deep Learning, 2021.

[20] C. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, "Activation functions: Comparison of trends in practice and research for deep learning," arXiv preprint, 2018. [Online]. Available: https://arxiv.org/abs/1811.03378.

[21] M. Forgione, A. Muni, D. Piga, and M. Gallieri, "On the adaptation of recurrent neural networks for system identification," Automatica, vol. 155, p. 111092, 2023.

[22] I. D. Mienye, T. G. Swart, and G. Obaido, "Recurrent neural networks: A comprehensive review of architectures, variants, and applications," Information, vol. 15, no. 9, p. 517, 2024.

[23] F. Huang et al., "Attention-emotion-enhanced convolutional LSTM for sentiment analysis," IEEE Transactions on Neural Networks and Learning Systems, vol. 33, no. 9, pp. 4332–4345, 2021.

[24] S. M. Omar, M. Kimwele, A. Olowolayemo and D. M. Kaburu, "Enhancing EEG signals classification using LSTM-CNN architecture," Engineering Reports, vol. 6, no. 9, 2023.

[25] I. D. Mienye and N. Jere, "Deep learning for credit card fraud detection: A review of algorithms, challenges, and solutions," IEEE Access, vol. 12, pp. 96893–96910, 2024.

[26] R. A. Priyanka and G. S. Sadasivam, "Classification of phonemes using EEG," in Proceedings of International Conference on Artificial Intelligence, Smart Grid and Smart City Applications: AISGSC 2019, pp. 521–530, Springer, 2020.

[27] Ildar Rakhmatulin, Minh-Son Dao, Amir Nassibi, and Danilo Mandic, 2024. "Exploring Convolutional Neural Network Architectures for EEG Feature Extraction," Sensors, vol. 24, no. 3, p. 877, https://doi.org/10.3390/s24030877.

[28] M. Bisla and R. S. Anand, "Optimized CNN-Bi-LSTM–Based BCI System for Imagined Speech Recognition Using FOA-DWT," Wiley, vol. 2024, 2024.

[29] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794, 2016.

[30] Z. Hasan Ali and A. M. Burhan, "Hybrid machine learning approach for construction cost estimation: an evaluation of extreme gradient boosting model," Asian Journal of Civil Engineering, vol. 24, pp. 1-16, 2023.