computation

MDPI

Article

EEG-Based Classification of Spoken Words Using Machine Learning Approaches

Denise Alonso-Vázquez [1], Omar Mendoza-Montoya [1], Ricardo Caraza [2], Hector R. Martinez [2] and Javier M. Antelis [1,*]

[1] Tecnologico de Monterrey, Escuela de Ingeniería y Ciencias, Monterrey 64849, Mexico; omendoza83@tec.mx (O.M.-M.)
[2] Tecnologico de Monterrey, Escuela de Medicina y Ciencias de la Salud, Monterrey 64849, Mexico; rcaraza@tec.mx (R.C.); hector.ramon@tec.mx (H.R.M.)
* Correspondence: mauricio.antelis@tec.mx

Abstract: Amyotrophic lateral sclerosis (ALS) is a neurodegenerative disease that affects the nerve cells in the brain and spinal cord. This condition leads to the loss of motor skills and, in many cases, the inability to speak. Decoding spoken words from electroencephalography (EEG) signals emerges as an essential tool to enhance the quality of life for these patients. This study compares two classification techniques: (1) the extraction of spectral power features across various frequency bands combined with support vector machines (PSD + SVM) and (2) EEGNet, a convolutional neural network specifically designed for EEG-based brain–computer interfaces. An EEG dataset was acquired from 32 electrodes in 28 healthy participants pronouncing five words in Spanish. Average accuracy rates of 91.04 ± 5.82% for *Attention* vs. *Pronunciation*, 73.91 ± 10.04% for *Short words* vs. *Long words*, 81.23 ± 10.47% for *Word* vs. *Word*, and 54.87 ± 14.51% in the multiclass scenario (*All words*) were achieved. EEGNet outperformed the PSD + SVM method in three of the four classification scenarios. These findings demonstrate the potential of EEGNet for decoding words from EEG signals, laying the groundwork for future research in ALS patients using non-invasive methods.

Keywords: electroencephalography; speech decoding; EEGNet

Citation: Alonso-Vázquez, D.; Mendoza-Montoya, O.; Caraza, R.; Martinez, H.; Antelis, J. EEG-Based Classification of Spoken Words Using Machine Learning Approaches. *Computation* **2023**, *11*, 225. https://doi.org/10.3390/computation11110225

Academic Editors: Juan Luis Crespo-Mariño and Marvin Coto-Jiménez

Received: 1 September 2023
Revised: 4 November 2023
Accepted: 7 November 2023
Published: 10 November 2023

1. Introduction

Language enables interaction and communication among individuals. It serves as a tool for thought in its symbolic and representational function, and plays a role in human action and interaction through its communicative and regulatory functions [1]. Speech is the oral expression of language, involving the physical production of sounds through mechanisms such as vocal cords, mouth, and tongue [2]. Numerous neurological disorders are characterized by the gradual degeneration of motor neurons, leading to a compromised or complete loss of speech capabilities. A prominent example of such conditions is amyotrophic lateral sclerosis (ALS). In ALS, it is estimated that approximately 85% of patients manifest symptoms indicative of bulbar dysfunction [3]. These symptoms, which include reduced verbal communication and impaired swallowing abilities, not only restrict the capacity for speech but also profoundly affect the overall quality of life of the individual.

In the realm of medical technology, an array of assistive devices has been developed to aid patients experiencing speech impairments. Particularly for ALS patients—who often encounter conditions such as tetraplegia and anarthria, the latter being a severe speech disorder marked by an inability to articulate sounds—the eye-tracking communication system stands out as a prevalent choice. Within this system, users direct and sustain their gaze on specific commands presented on a display. These commands are identified via infrared camera. Nonetheless, the substantial cost of these devices remains a significant barrier to widespread adoption [4]. Another extensively researched alternative is the use of brain–computer interfaces, known as BCIs. These interfaces detect and quantify features of

brain activity that indicate the intent of the user, translate in real-time these measurements into commands for the device, and provide feedback to the user simultaneously [5].

The main techniques for recording brain activity are (i) functional magnetic resonance imaging (fMRI), recognized for its non-invasive nature and remarkable spatial resolution; (ii) electroencephalography (EEG), which has excellent resolution over time; (iii) electrocorticography (ECoG), an invasive method with high temporal and spatial resolution; and (iv) magnetoencephalography (MEG), which stands out for its high temporal and spatial resolution, although it carries a high cost [6]. EEG is considered the most common method in developing BCIs since it is non-invasive, easy to use, safe, and affordable compared to other methods [7]. Several BCIs operate based on evoked potentials, such as the P300 potential or the steady-state visual evoked potential (SSVEP). Others rely on cognitive tasks like motor imagery (MI) [8,9]. In the P300 paradigm, participants typically gaze at a screen displaying flashing characters and select one by focusing their attention. For applications based on the SSVEP paradigm, visual stimuli or frames flash at varying frequencies. When the user concentrates on the desired item, a potential is generated in the visual cortex, matching the flashing frequency of the chosen image. In MI, the participant imagines moving a limb without actual movement; these signals are then decoded, and a command is sent to activate an external device [10]. While these paradigms have been widely used in a range of BCI applications [11–15], they are limited when the primary requirement for the user is direct communication through speech rather than control of movement. This is especially relevant for conditions such as bulbar ALS, dysarthria, anarthria, stroke, and Parkinson's disease, among other speech disorders. P300, SSVEP, and MI paradigms offer the capacity to select options that can subsequently be converted into auditory signals or text output [16,17]. However, the cognitive tasks engaged in by the user do not have a direct relation to speech processes; in other words, they do not decode responses that are inherently speech-related. Consequently, traditional BCI paradigms fall short in establishing a natural communication channel for the decoding and generation of speech.

Various studies have been conducted to decode speech from brain signals, whether it is spoken or overt (the most common form of human verbal communication with audible volume, clarity, intonation, and modulation), whispered (softer, less audible sound compared to normal speech), silent (merely gestured without producing sound), or imagined (internal pronunciation of the word without making any facial movement and without emitting any sound) [18]. In [19], attempted speech was decoded in a participant with anarthria resulting from a stroke. A 128-electrode array was implanted in the speech sensorimotor cortex, and a natural language model was employed to estimate the probability of the subsequent word given the preceding words in a sequence. In [20], recurrent neural networks were employed to first decode directly recorded cortical activity into representations of articulatory movement, and then these representations were transformed into speech acoustics. The study was conducted with five participants who had a high-density subdural electrode array implanted as part of their treatment for epilepsy. In [21], an approach is also presented to synthesize audible speech from imagined and whispered speech using a dictionary of 100 Dutch words and stereotactic deep electrodes implanted in a participant. Although these studies decode speech across different paradigms, they rely on invasive methods, posing health risks to the patient and escalating the costs compared to EEG. Using EEG signals, work has been done to investigate the neural mechanisms underlying the processing of heard words [22,23]. In terms of decoding the words produced, a system based on residual networks and gated recurrent unit models was used where the silent speech of eight Russian words and one pseudoword was decoded [24]. Some works classify grammatical classes, as seen in [25], where different deep learning methods and conventional classifiers were compared for recognizing spoken speech and distinguishing between decision adverbs and nouns. In [26], imagined speech was classified between nouns and verbs using a convolutional neural network (CNN). Spanish vowels have also been decoded, such as in [27], where they propose the architecture of a CNN, referred to as CNNeeg1-1. This has also been proposed for short and long words in English [28] using a

method based on covariance matrix descriptors, which are on the Riemannian manifold, in conjunction with a relevance vector machines classifier. In [29], pairs of Spanish words were classified using a traditional method and two CNNs. In [30], various traditional classification methods and three distinct CNNs are evaluated, aiming to find an optimal combination of hyperparameters for the classification of Spanish words. Most work on word decoding performs an intra-subject classification; in [31], they mapped the EEG signal across seven frequency bands over nine brain regions and used a deep long short-term memory (LSTM) network for inter-subject classification. Similarly, using an LSTM along with a wavelet sparsity transformation in [32], four English words were decoded. One growing area is the ability to gradually add new words to the vocabulary using incremental learning methods, similar to the approach used in [33].

Despite the continuous growth in the field of study, there remains a notable disparity between research focusing on online speech decoding using invasive methods and those employing non-invasive techniques like EEG. Furthermore, a majority of the studies in this domain have been primarily conducted on young and healthy participants or, in some cases, on patients with epilepsy who do not have any diagnosed speech disorders [6]. It is imperative to identify characteristics that enhance the differentiation between classes when classifying words and to develop methods for online decoding—initially in healthy participants and subsequently in the group of patients toward whom this technology is aimed. It is noteworthy that neural activities linked to language processing demand greater spatial and temporal resolutions for analysis compared to other cognitive functions, such as memory [34].

In light of this, as a preliminary step towards decoding attempted speech in ALS patients from EEG signals, this work aims to investigate the recognition of spoken words directly from the brain signals using machine learning algorithms. To achieve this, we designed and conducted an experiment where participants pronounced five words varying in connotation, syllable count, grammatical class, semantic meaning, and functional role within a sentence. In this paper, we contrast two machine learning approaches for classification: (1) a conventional method that involves feature extraction in the frequency domain and support vector machines, and (2) EEGNet, a convolutional neural network designed for EEG-based BCIs. These methods were evaluated across various intra-subject classification scenarios involving 28 healthy participants and five Spanish words: (i) attention vs. speech, (ii) short words vs. long words, (iii) word vs. word, and (iv) an all words scenario. The findings indicate that EEGNet outperformed the traditional method in three out of the four classification scenarios, achieving average accuracy values of 90.69 ± 5.21% for *Attention* vs. *Pronunciation*, 73.91 ± 10.04% for *Short words* vs. *Long words*, 81.23 ± 10.47% for *Word* vs. *Word*, and 54.87 ± 14.51% in the multiclass classification scenario (*All words*).

The innovation and scientific contribution of this work include (1) the acquisition of an EEG signal database comprising carefully selected spoken words; (2) a comparative evaluation of two machine learning techniques for the classification of spoken words based on EEG signals; (3) the employment of EEGNet, a convolutional neural network designed with the methodology of EEG-based BCIs, which, to the best of our knowledge, has not been previously applied to spoken speech signal classification; (4) a comprehensive statistical analysis to identify those words that are significantly more easily decoded. This paper is structured as follows. Section 1 provides an overview of previous work and outlines our research objectives. Section 2 details the experimental procedure and the classification methods employed. The values obtained in the classification performance metrics are presented in Section 3, while Section 4 offers an interpretation and analysis of the of these. Lastly, in Section 5, key findings are summarized, and implications and future directions of the study are highlighted.

2. Materials and Methods

This section describes the experimental protocol for data acquisition, processing, and preparation of electroencephalography signals and the classification scenarios and methods to be evaluated.

2.1. EEG Data Recording Experiment

The experiments were carried out with 28 healthy participants (P1–P28) with an age range between 20 and 47 years old (mean = 23.5 years, std = 5.5 years). From these participants, 7 were women, and 21 were man. All participants were right-handed and Spanish language native speakers, which was a requirement to be accepted as part of the experiment. None of the participants reported having any psychological, neurological, or speech-related disorders. All participants were duly informed about the experiment and freely agreed to be part of it; they signed an informed consent form that granted permission to use their data, including the photographic and video files. The experiments were conducted in accordance with the standards of the Helsinki Declaration after the approval of the ethics committee of the Zambrano Hellion Hospital, Monterrey, Mexico.

The experiment consisted of the pronunciation of the five Spanish words: *si*, *no*, *agua*, *comida*, and *dormir*, which translate into English as *yes*, *no*, *water*, *food*, and *sleep*, respectively. These words were chosen because they are basic for communication, especially for people with special communication limitations or needs. Additionally, they were carefully chosen so that among them, they had the following variations: connotation, syllable count, grammatical class, semantic meaning, and functional role within a sentence, so that we would be able to study certain characteristics of speech production that can help the decoding of words. The participant sat in front of a computer screen (18.5 inches) where a graphical user interface (GUI) guided the execution of the experiment. The GUI was used to present three visual stimuli that instructed the participants as to the specific action to take. The first visual stimulus consisted of a fixation cross that instructed the participant to relax and pay attention by focusing their gaze on it. The second visual stimulus consisted of one of the five Spanish words and indicated for the participant to pronounce the word at her/his normal pitch and volume. Note that the participants were duly instructed to move the mouth only while avoiding other body movements. The third visual stimulus was the image of a palm tree and indicated for the participants to take a break from the experiment, for instance, by blinking and moving the head and the body if necessary. These three visual stimuli were presented for 3 s each, resulting in a trial with a duration of 9 s. Figure 1 shows a snapshot of the experimental setup with a participant wearing the EEG electrodes, the amplifier, the screen, and the computer, while Figure 2 illustrates the timeline of a trial.

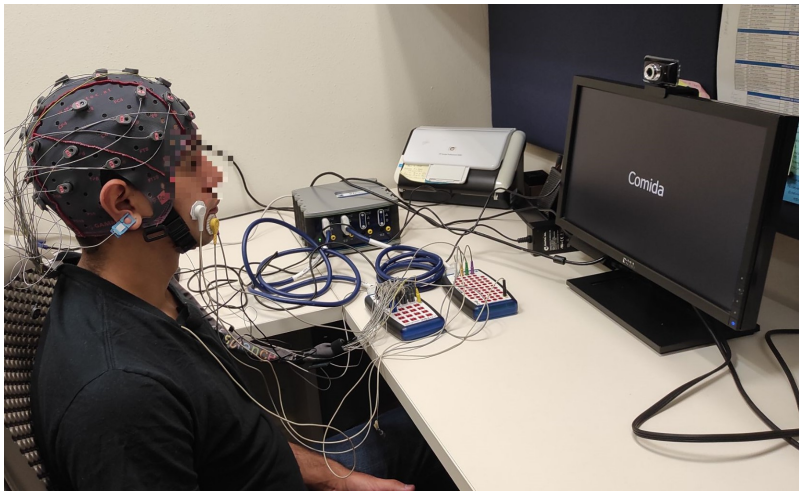

Figure 1. Photograph of the experimental setup. The participant is seated in front of the monitor with the EEG cap. The monitor shows the word *comida*, corresponding to the pronunciation stage.

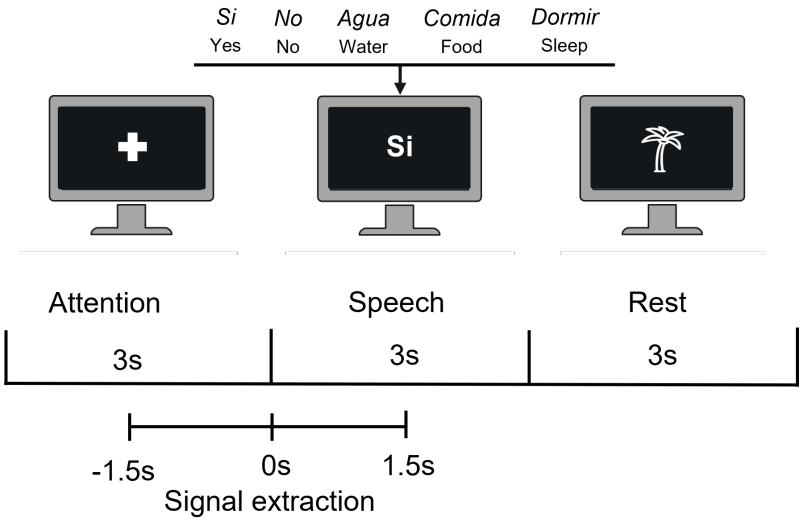

Figure 2. Graphical illustration of the temporal sequence of a trial. The trial starts with 3 s of attention (fixation cross), then for 3 s, the participant is shown one of the five words that will have to be pronounced only once, and finally 3 s of rest (palm). In addition, the lower part of the image shows the data segment from −1.5 s to 1.5 s, corresponding to the time window of interest.

The experiment was executed in four experimental blocks with a rest of at least 3 min between them. In an experimental block, a total of 50 trials, like the one presented in Figure 2, are recorded. The order of the words in a block is random, so the participant cannot anticipate the word to be pronounced. The duration of a block is 7.5 min and 10 trials per word are recorded in a block. Considering the four experimental blocks, for each participant, a total of 200 trials were obtained, with a total of 40 trials per word.

The EEG signals were recorded from 32 active electrodes uniformly distributed over the scalp according to the 10–10 international system. This system is an extended version of the 10–20 system. The ground electrode was placed at the AFz position, and the reference electrode was located on the right ear lobe. These biosignals were acquired and digitalized with the biosignal amplifier g.HIAMP 256 from g.tec medical engineering GmbH, Schiedlberg, Austria. A Butterworth-type band-pass filter from 0.5 to 500 Hz and a notch-type filter from 58 to 62 Hz were applied during the recording. The EEG signals were recorded at a sampling rate of 1200 Hz. After the execution of the experiments, the recorded EEG signals of each participant were subjected to the following processing steps.

2.2. EEG Data Processing and Preparation

The EEG signals were band-pass-filtered from 1 to 30 Hz with an eighth-order Butterworth-type digital filter and they were downsampled to 256 Hz. Epochs from −1.5 to 1.5 s with respect to the presentation time of the second visual stimulus (i.e., the time instant where the stimulus with the word to be pronounced is shown on the screen) were extracted. These epochs were selected because they include a pre-stimulus interval where the participant is only focusing attention on the screen with no pronunciation and a post-stimulus interval that encompasses the pronunciation of the word. Independent component analysis (ICA) was performed and then the EEG was reconstructed by zeroing components associated with artifacts such as blinking, heart activity, and muscle activity. Artifact components were selected based on visual inspection and four or less components were ruled out. A detailed description of the parameters and methods used in the implementation of ICA, as well as its limitations and challenges, is shown in Appendix A. Noisy epochs were then identified and discarded by applying a threshold-based algorithm according to the following conditions of the data distribution. An epoch was considered as noisy if, in at least one of the electrodes, the peak-to-peak voltage was greater than 150 μV or the standard deviation was greater than 20 μV.

2.3. Classification Scenarios

The recognition of words during overt speech from EEG signals was investigated in four classification scenarios, the first three being bi-class and the last multiclass (5 classes). These classification scenarios are described below.

- *Attention* vs. *Pronunciation*. The objective in this scenario was to discriminate between no pronunciation at all and pronounced words regardless of the words. For the first class, EEG signals in the time segment from −1.5 s to 0 s were used, while for the second class, EEG signals in the time segment from 0 to 1.5 s were used. In a participant where no trials were removed, the number of trials in this scenario was 200 for each class. This classification scenario is significant because of the ability to discriminate between segments where pronunciation is present and where it is not. It serves as the initial step in developing a system that decodes attempted speech in patients with speech impairments.
- *Short words* vs. *Long words*. In this scenario, we studied the classification between two groups of words according to their length: short and long. The short word group corresponded to the words of one syllable, *si* and *no*, while the long word group corresponded to two-syllable words, *agua* and *dormir*. In a participant where no trials were removed, the number of trials in this scenario was 80 for each class. The aim was to determine if it is possible to distinguish between short and long words since, in this case, the short words correspond to answers to binary questions.
- *Word* vs. *Word*. Here, we investigated the classification between all possible pairs of words (ten different bi-class situations, since we have 5 words). The objective was to determine which pairs of words can and cannot be discriminated from the EEG signals. In a participant where no trials were removed, the number of trials in this scenario was 40 for each class.
- *All words.* The objective in this scenario was to study the recognition between the five words; therefore, this is a multiclass classification with five classes. In a participant where no trials were removed, the number of trials in this scenario was 40 for each class.

Note that in the classification scenarios of *Word* vs. *Word*, *Short words* vs. *Long words*, and *All words*, only the time segment from 0 s to 1.5 s was used because it corresponds to the EEG signals of the pronunciation of the words.

2.4. Classification Methods

To investigate the classification of overt speech using EEG signals, we employed two machine learning-based approaches. In the first, we used time-domain and frequency-domain feature extraction techniques commonly used in EEG signals, in combination with classification methods. In this case, we explored different combinations of them, and here we report the results of the combination that provided the best performance: power spectral-based features and a support vector machine. In the second approach, we used a convolutional neural network model named EEGNet, which was specially designed for the classification of EEG signals.

2.4.1. Feature Extraction and Classifier

The power spectral density (PSD) of EEG frequency bands was used as features. PSD is a widely recognized feature extraction method in EEG-based classification. This is because power spectral variations occur in numerous scenarios, including motor tasks [9] and cognitive tasks [35]. In specific, we used the PSD of six frequency bands: delta (1–4 Hz), theta (4–7 Hz), alpha (8–13 Hz), low-beta (12–15 Hz), mid-beta (15–20 Hz), and high-beta (18–30 Hz). Using data windows of 1.5 s duration, the PSD was computed using the multitaper fast Fourier transform method, a technique designed to estimate the spectral representation of a system based on finite observations [36]. Multitaper methods are powerful for performing spectral analysis of short data segments [37]. The Hann (or Hanning) window operation was applied to each data window. The PSD was calculated

for each of the six frequency bands at a resolution of 0.5 Hz. Subsequently, the mean PSD was computed across the range of each band. Therefore, the resulting feature vector is $\mathbf{x} \in \mathbb{R}^{(N_{bands} \cdot N_{Channels}) \times 1}$, where $N_{bands} = 6$ is the number of frequency bands and $N_{Channels} = 32$ is the number of EEG channels, that is, the dimension of the feature vector is $N_{bands} \cdot N_{Channels} = 192$. To obtain a low-dimensional representation of the feature vector but with high discriminatory power, we use a feature selection method based on the F-statistic [38] to select the 50 best-ranked features. Therefore, the final dimension of the vector is $(1, 50)$. Hence, the final dimension of the feature vector provided to the classifier is 50. Data preprocessing and feature extraction were carried out in the FieldTrip toolbox of MATLAB R2022a.

The classification was performed with support vector machines (SVMs) implemented in scikit-learn in Python [39], using a linear kernel with a regularization parameter of 1. The SVM is a binary classifier; however, it can be extended by merging several types into a multiclass classifier by implementing the one-versus-one approach [40]. SVM is an algorithm that has shown effective results in EEG signal applications [14,15,22,23], and it is also one of the most popular machine learning algorithms used to decode speech from EEG signals [41]. Therefore, we define the method that uses feature extraction based on power spectral density, and for classification, it employs support vector machines, as the $PSD + SVM$ method.

2.4.2. EEGNet

EEGNet is a compact convolutional neural network architecture for EEG-based BCIs, which can be used for different BCI paradigms and trained with a very limited amount of data, capturing patterns in EEG signals in time and space [42].

Network Architecture and Hyperparameters

The EEGNet architecture (see Figure 3) is composed of three blocks; the first consists of a two-step convolutional sequence. It starts with a temporal convolution to learn frequency filters. For the first step of the block, we used $F1 = 6$ for bi-class classification and $F1 = 8$ for multiclass classification (5 classes), where $F1$ is the number of temporary filters with a kernel size of $(1, 128)$, that is, half of the sampling rate. The second step is a depthwise convolution to learn spatial filters for each temporal filter to extract frequency-specific spatial filters efficiently. The size is $(C, 1)$, with C defined as the number of channels; therefore, $C = 32$. The depth parameter D is set to $D = 2$, which is the number of spatial filters to learn within each temporal convolution. This part of the first block (the combination of temporal filtering with spatial filtering) is inspired by the filter-bank common spatial pattern (FBCSP) algorithm [43]. Subsequently, batch normalization is applied between the dimension of the feature maps before applying the exponential linear unit (ELU) nonlinearity and the dropout layer to regularize the model, with a probability set at 0.85. The block ends with an average pooling layer of size $(1, 4)$ to reduce the sampling rate of the signal to 64 Hz.

The second block is composed of a separable convolution formed by the combination of a deep convolution (size $(1, 16)$) and a point-wise convolution, with $F2 = F1 * D$, where $F2$ is the number of point filters to learn. As in the previous block, batch normalization is applied to the entire dimension of the feature maps obtained from the separable convolution. Then, the exponential linear unit (ELU) nonlinearity activation layer, and the dropout layer to regularize the model, with a probability set at 0.85, are used. Those layers are followed by average pooling, set to a size of $(1, 8)$ to reduce dimension. The third block corresponds to the classification made with the softmax function explained below. Detailed network information is available at [42].

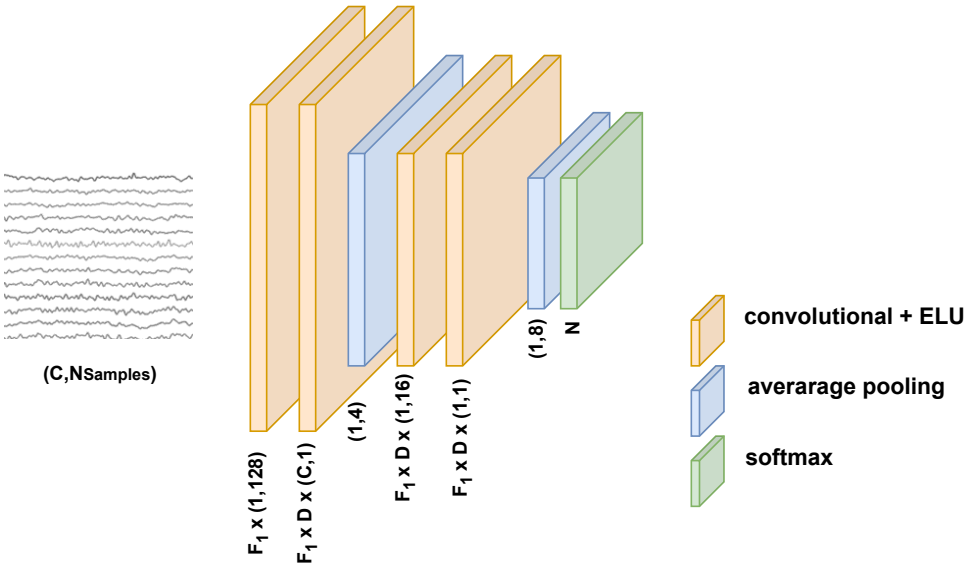

Figure 3. Architecture of the convolutional neural network EEGNet.

Network Training and Classification

The input vector to the EEGNet has dimension $\mathbf{x} \in \mathbb{R}^{(N_{Channels} \cdot N_{Samples})}$, where $N_{Channels} = 32$ is the number of EEG channels and $N_{Samples}$ is the number of samples in each time window. $N_{Samples} = 385$ is the number of samples in 1.5 s of the signal. The training was carried out intra-subject, which means that one model was made per subject. A softmax layer with N units is used in the classification block, where N is the number of classes; therefore, $N = 2$ and $N = 5$, respectively. The softmax function $g_n(T)$ assigns a probability to each class based on the values of the vector of outputs T [44], expressed as follows:

$$g_n(T) = \frac{e^{T_n}}{\sum_{l=1}^{N} e^{T_l}}. \quad (1)$$

We use Google Colab GPUs to train the network developed in Tensorflow using Keras API. The training was performed for each subject, and the number of epochs was 200 for bi-class classification and 300 for multiclass classification. These values were determined by visualizing the learning curves for both accuracy and loss. This curve illustrates how performance on the training and validation sets changes as one progresses through the epochs. The Adam optimizer and the categorical cross-entropy loss function were used to fit the model.

2.5. Validation Procedure and Performance Metrics

In each one of the four classification scenarios (i.e., *Attention* vs. *Pronunciation*, *Short words* vs. *Long words*, *Word* vs. *Word*, and *All words*), we assessed performance through the following procedure. Five cross-validation folds were implemented to train and test the classification methods. In detail, the set of features was split into five groups. Then, features from four groups were taken as the train dataset to tune the classification models, while features in the remaining group were used as the test dataset to compute the performance metrics. This procedure was repeated five times, taking a different group as training and testing datasets each time. Note that train and test datasets are always mutually exclusive. The procedure was carried out intra-subject, which means that the procedure was performed independently with the data of each participant.

To assess performance, we computed the following metrics:

i. The total classification accuracy represents the fraction of correctly classified instances in the test set [45], and it is defined as follows:

$$\mathrm{Acc}_T(f) = \frac{1}{|T|}\sum_{i=1}^{|T|} I(f(x_i) = y_i), \tag{2}$$

where $|T|$ is the total number of samples in the test dataset T, $I(a)$ is the indicator function, which is equal to 1 if the predicted class a is true, and 0 otherwise, $f(x_i)$ is the label predicted to the example x_i by the model f for the i-th sample, and y_i is the true label for the i-th sample.

ii. Recall by class or Recall_{class} is the number of elements correctly classified in that class divided by the total number of elements that should have been classified in that class, that is, the performance of the model when detecting that class. This is defined as follows:

$$\mathrm{Recall}_{class_c} = \frac{\sum_{i=1}^{|T|} I(y_i = c \wedge f(x_i) = c)}{\sum_{i=1}^{|T|} I(y_i = c)} \tag{3}$$

where c is the specific class for which the recall or precision is computed.

iii. Precision by class or $\mathrm{Precision}_{class}$ is the number of elements correctly classified in that class divided by the total number of elements classified as belonging to that class. It indicates how reliable our model is in predicting a specific class. It is defined as follows:

$$\mathrm{Precision}_{class_c} = \frac{\sum_{i=1}^{|T|} I(y_i = c \wedge f(x_i) = c)}{\sum_{i=1}^{|T|} I(f(x_i) = c)} \tag{4}$$

These metrics were computed for each fold, and then the mean and standard deviation can be computed.

The Wilcoxon signed-rank test was performed to assess whether there were significant differences between the results obtained in each of the methods. To determine if there was a class that was better recognized by the method, a Friedman test was performed between the recalls and the precisions of each class for each method. To determine between which classes these significant differences exist, the Nemenyi post hoc test was performed. The Wilcoxon signed-rank test serves as a non-parametric substitute for the paired *t*-test. It assesses the performance differences between two classifiers on each dataset by ranking these differences without considering their signs. Then, it evaluates the ranks of the positive and negative differences. The Friedman test is a non-parametric statistical test that compares three or more paired or related groups on an ordinal or continuous variable. It is a non-parametric counterpart to the repeated-measures ANOVA. For each dataset, it ranks algorithms, giving the highest-performing one a rank of 1 and assigning subsequent ranks to the others. The Nemenyi post hoc test is employed alongside the Friedman test to compare all classifiers against each other. It assesses whether the average rankings of two classifiers show significant differences [46].

3. Results

Here we present the classification results achieved with both classification methods, PSD + SVM and EEGNet, in the four classification scenarios, *Attention* vs. *Pronunciation*, *Short words* vs. *Long words*, *Word* vs. *Word*, and *All words*. The models are trained and tested per participant.

3.1. Bi-Class: Attention vs. Pronunciation

Figure 4 shows the across-all-participants distributions of classification accuracy, recall per class, and precision per class metrics achieved with both classification methods. These results show that the distributions of all performance metrics present median values around or even greater than 90%. As a summary of these classification results, the mean and standard deviation values derived from each metric and classification method are

presented in Table 1. Regarding the precision per class, the two methods have similar rates of correct positive predictions in relation to the total number of positive predictions the method makes for both classes (mean values around 91%).

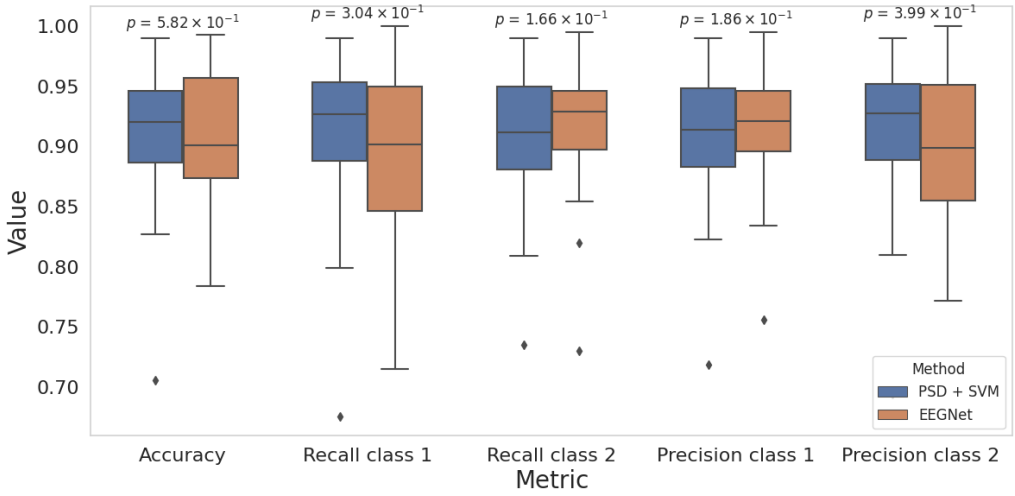

Figure 4. Distribution of classification accuracy, recall per class, and precision per class obtained with the PSD + SVM and EEGNet methods in the classification scenario *Attention* vs. *Pronunciation*. The diamonds represent outliers. In all performance metrics, no significant differences were found between the two classification methods (Wilcoxon signed-rank test, $p > 0.01$).

Table 1. Mean and standard deviation of classification accuracy, recall per class, and precision per class obtained with the PSD + SVM and EEGNet methods in the classification scenario *Attention* vs. *Pronunciation*.

Method	Accuracy	Recall_{class1}	Recall_{class2}	$\text{Precision}_{class1}$	$\text{Precision}_{class2}$
PSD + SVM	91.04 ± 5.82%	91.30 ± 6.42%	90.79 ± 5.64%	90.82 ± 5.75%	91.34 ± 6.10%
EEGNet	90.69 ± 5.21%	89.80 ± 7.16%	91.59 ± 5.69%	91.58 ± 5.24%	90.26 ± 6.35%

In addition, the classification metrics were also computed and analyzed separately for each participant, and the results showed that all of them achieved accuracies greater than 70%, with the baseline chance level being 50% (see Figure A1 in Appendix B). Similarly, the recall and precision per class also showed performance greater than 70% in all participants in the two methods.

3.2. Bi-Class: Short Words vs. Long Words

In this classification scenario, one-syllable words were categorized as class 1 (short words), while two-syllable words were designated as class 2 (long words). Figure 5 displays the distributions of classification accuracy, recall for each class, and precision for each class across all participants for both methods. Table 2 presents the mean and standard deviation for each metric.

In addition to analyzing metrics across all participants, we evaluated accuracy per participant. Using EEGNet, some participants achieved an accuracy rate of 90%, whereas, with PSD + SVM, they exceeded 80% (see Figure A2 in Appendix B). Employing the PSD + SVM method, only one participant (P10) registered an accuracy below this threshold. In the majority of cases, EEGNet demonstrated superior performance over the PSD + SVM method.

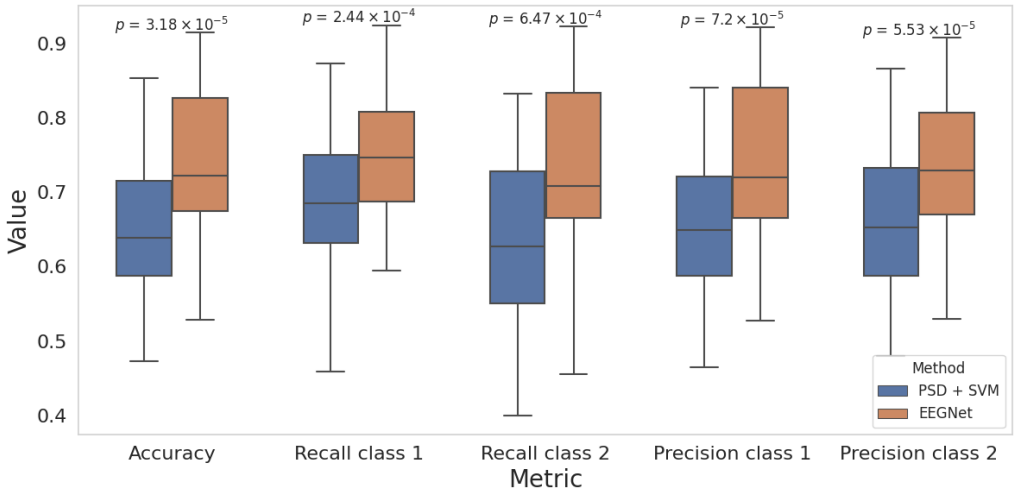

Figure 5. Distribution of classification accuracy, recall per class, and precision per class obtained with the PSD + SVM and EEGNet methods in the classification scenario *Short words* vs. *Long words*. In all performance metrics, significant differences were found between the two classification methods (Wilcoxon signed-rank test, $p < 0.01$).

Table 2. Mean and standard deviation of classification accuracy, recall per class, and precision per class obtained with the PSD + SVM and EEGNet methods in the classification scenario *Short words* vs. *Long words*.

Method	Accuracy	Recall_{class1}	Recall_{class2}	$\text{Precision}_{class1}$	$\text{Precision}_{class2}$
PSD + SVM	65.72 ± 9.56%	68.23 ± 9.58%	63.02 ± 10.92%	65.43 ± 9.27%	66.01 ± 10.02%
EEGNet	73.91 ± 10.04%	75.05 ± 9.57%	72.75 ± 12.57%	74.21 ± 10.88%	74.00 ± 10.03%

3.3. Bi-Class: Word vs. Word

In this classification scenario, we identify the word pairs that are easiest and most challenging to recognize. Figure 6 provides a comparative analysis of the accuracy distributions obtained between the EEGNet and PSD + SVM methods across different word pair combinations. Table 3 displays the mean results for accuracy, recall by class, and precision by class for all possible combinations among the five words. For the PSD + SVM method, the average accuracies ranged from 63.02 ± 11.13% to 73.97 ± 13.58%. With EEGNet, all average accuracies exceeded 70%, reaching up to 81.23 ± 10.47% in the *si* vs. *agua* pair.

3.4. Multiclass: All Words

The final classification scenario was multiclass, where each class corresponded to one of the five words under consideration. Figure 7 displays the distributions of classification accuracy, recall by class, and precision by class across all participants for each word, using both classification methods. Table 4 provides the average results for total accuracy, recall by class, and precision by class for each method. Using EEGNet, an average accuracy of 54.87 ± 14.51% was achieved, while PSD + SVM obtained 41.78 ± 13.34%, with 20% of baseline chance level. According to the Friedman test ($p > 0.01$), using EEGNet yielded no significant differences between the recalls and precisions of each class. However, with PSD + SVM, certain classes were more easily recognized than others (Friedman test, recall and precision per class: $p < 0.01$). To identify the classes between which these significant differences existed, the Nemenyi post hoc test ($p < 0.01$) was conducted. It was found that the word *si* had greater recall than *dormir*, and similarly, *agua* had greater recall than *dormir*. In terms of precision, *si* was detected better than the word *agua*, which in turn was detected with more precision than *no* and *dormir*. These findings suggest that with this method, certain words are more effectively captured by the model compared to others. Additionally, classification metrics were computed and analyzed individually for each participant. Using

EEGNet, accuracy rates ranged from 30% to 80%, whereas with PSD + SVM, they varied from approximately 20% to 70% (see Figure A3 in Appendix B).

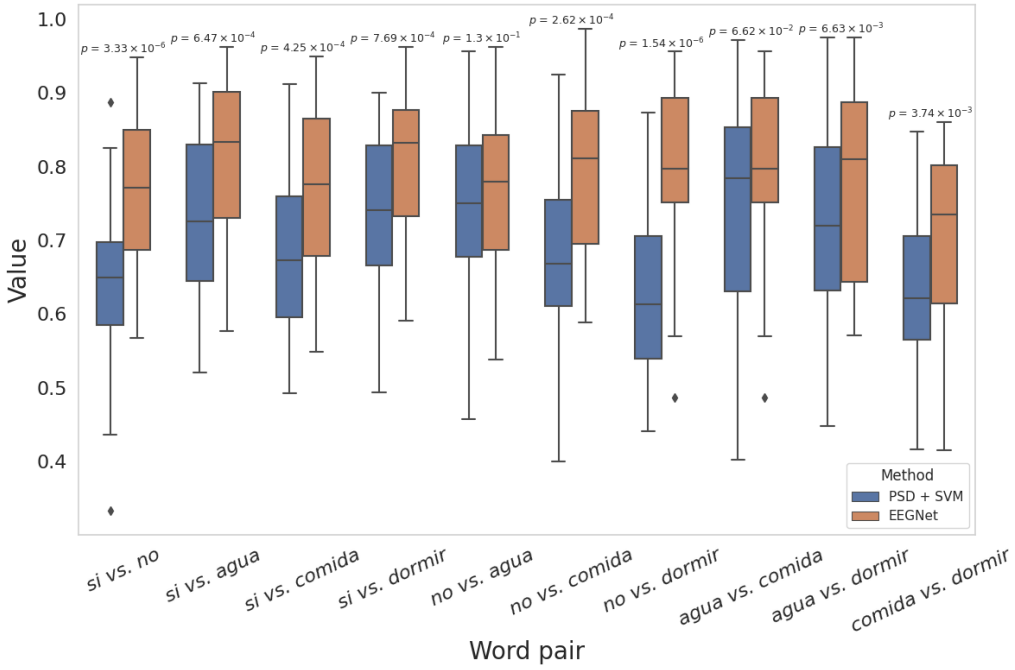

Figure 6. Distribution of classification accuracy for each word pair obtained with the PSD + SVM and EEGNet methods in the classification scenario *Word* vs. *Word*. The diamonds represent outliers. In eight out of the ten word pair classifications, significant differences were found between the two classification methods (Wilcoxon signed-rank test, $p < 0.01$).

Table 3. Mean and standard deviation of classification accuracy, recall per class, and precision per class obtained with the PSD + SVM and EEGNet methods in the classification of pairs of words. 1 = *si* vs. *no*, 2 = *si* vs. *agua*, 3 = *si* vs. *comida*, 4 = *si* vs. *dormir*, 5 = *no* vs. *agua*, 6 = *no* vs. *comida*, 7 = *no* vs. *dormir*, 8 = *agua* vs. *comida*, 9 = *agua* vs. *dormir*, and 10 = *comida* vs. *dormir*.

Method	Accuracy	Recall$_{class1}$	Recall$_{class2}$	Precision$_{class1}$	Precision$_{class2}$
PSD + SVM$_1$	64.56 ± 11.69%	65.12 ± 12.59%	63.97 ± 13.31%	64.95 ± 11.60%	64.37 ± 12.11%
EEGNet$_1$	76.61 ± 10.78%	77.18 ± 11.44%	76.01 ± 13.00%	77.07 ± 11.18%	76.71 ± 11.25%
PSD + SVM$_2$	73.13 ± 11.43%	75.10 ± 11.46%	70.95 ± 13.12%	73.26 ± 11.72%	73.31 ± 11.60%
EEGNet$_2$	81.23 ± 10.47%	81.96 ± 10.07%	80.26 ± 14.20%	82.26 ± 10.84%	80.61 ± 11.46%
PSD + SVM$_3$	68.32 ± 11.80%	69.11 ± 11.60%	67.60 ± 13.49%	68.24 ± 12.53%	68.63 ± 11.56%
EEGNet$_3$	76.24 ± 11.42%	76.37 ± 12.35%	76.06 ± 12.99%	76.38 ± 11.96%	76.69 ± 11.90%
PSD + SVM$_4$	73.85 ± 11.29%	75.36 ± 10.91%	72.08 ± 14.19%	74.47 ± 11.89%	73.38 ± 11.26%
EEGNet$_4$	80.76 ± 10.28%	81.69 ± 9.63%	79.76 ± 13.93%	81.57 ± 10.92%	80.24 ± 10.82%
PSD + SVM$_5$	73.97 ± 13.58%	74.94 ± 14.11%	72.92 ± 14.10%	73.00 ± 13.37%	74.05 ± 13.88%
EEGNet$_5$	76.54 ± 11.77%	76.63 ± 10.71%	76.47 ± 15.71%	78.08 ± 13.00%	75.71 ± 11.96%
PSD + SVM$_6$	68.51 ± 12.04%	67.87 ± 13.02%	68.98 ± 12.86%	68.28 ± 12.31%	68.88 ± 12.18%
EEGNet$_6$	80.13 ± 10.74%	78.63 ± 12.34%	81.44 ± 11.58%	80.53 ± 11.87%	80.02 ± 10.20%
PSD + SVM$_7$	63.08 ± 11.77%	63.10 ± 13.40%	62.99 ± 12.27%	62.66 ± 12.39%	63.65 ± 11.42%
EEGNet$_7$	78.74 ± 12.55%	76.82 ± 15.65%	80.06 ± 12.35%	78.38 ± 13.38%	79.38 ± 12.85%
PSD + SVM$_8$	73.48 ± 14.76%	71.14 ± 16.33%	75.49 ± 13.93%	73.26 ± 15.45%	73.57 ± 14.42%
EEGNet$_8$	78.74 ± 12.56%	76.82 ± 12.34%	81.44 ± 11.58%	80.53 ± 11.87%	80.02 ± 10.20%
PSD + SVM$_9$	73.10 ± 13.30%	70.28 ± 15.19%	75.60 ± 13.02%	73.24 ± 14.40%	72.96 ± 12.72%
EEGNet$_9$	78.41 ± 11.95%	77.95 ± 14.34%	78.79 ± 12.01%	77.71 ± 12.68%	79.44 ± 12.03%
PSD + SVM$_{10}$	63.02 ± 11.13%	62.97 ± 12.87%	63.06 ± 11.11%	63.11 ± 11.29%	63.06 ± 11.42%
EEGNet$_{10}$	71.20 ± 11.68%	71.32 ± 14.68%	71.01 ± 11.40%	71.04 ± 12.19%	71.69 ± 11.88%

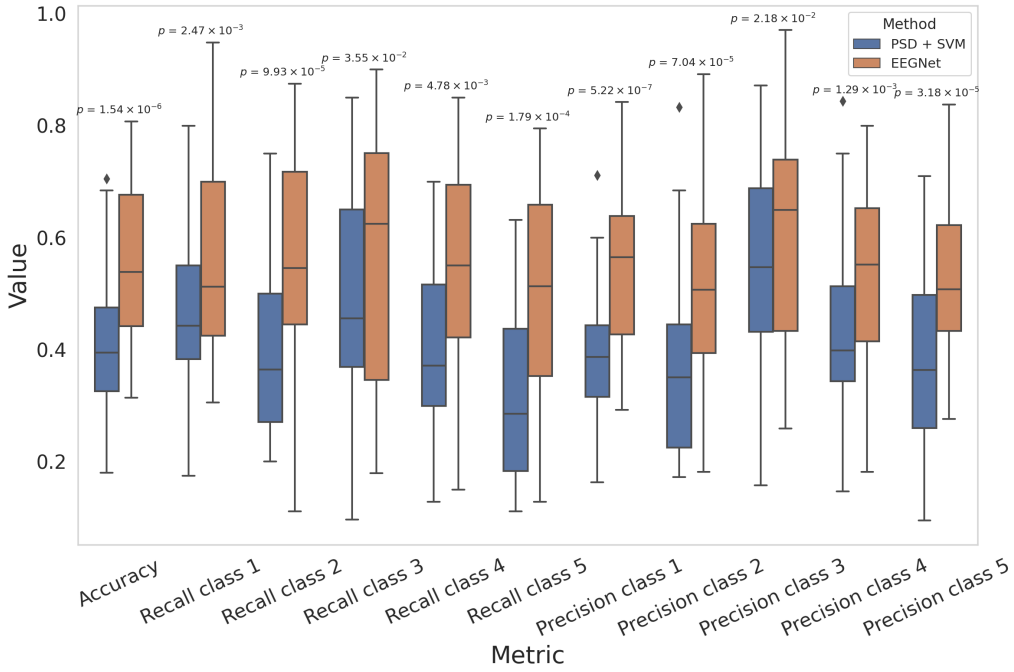

Figure 7. Distribution of classification accuracy, recall per class, and precision per class obtained with the PSD + SVM and EEGNet methods in the classification scenario *multiclass*. The diamonds represent outliers. In nine of the eleven computed performance metrics, significant differences were found between the two classification methods (Wilcoxon signed-rank test, $p < 0.01$): $class_1 = si$, $class_2 = no$, $class_3 = agua$, $class_4 = comida$, and $class_5 = dormir$.

Table 4. Mean and standard deviation of classification accuracy, recall per class, and precision per class obtained with the PSD + SVM and EEGNet methods in the multiclass classification: $class_1 = si$, $class_2 = no$, $class_3 = agua$, $class_4 = comida$, and $class_5 = dormir$.

Method	Accuracy	Recall_{class1}	Recall_{class2}	Recall_{class3}	Recall_{class4}
PSD + SVM	41.78 ± 13.34%	46.55 ± 14.20%	39.48 ± 15.62%	49.97 ± 20.06%	40.31 ± 16.44%
EEGNet	54.87 ± 14.51%	56.33 ± 17.68%	55.63 ± 20.57%	57.47 ± 22.22%	54.79 ± 19.40%
Recall_{class5}	$\text{Precision}_{class1}$	$\text{Precision}_{class2}$	$\text{Precision}_{class3}$	$\text{Precision}_{class4}$	$\text{Precision}_{class5}$
33.01 ± 16.30%	39.02 ± 11.88%	36.69 ± 16.51%	54.32 ± 20.05%	43.53 ± 17.25%	38.40 ± 16.06%
50.16 ± 18.31%	55.00 ± 14.06%	52.24 ± 19.13%	60.60 ± 20.71%	54.07 ± 16.19%	52.63 ± 14.78%

4. Discussion

To carry out this work, we acquired a database of EEG signals during the pronunciation of five words in the speech of 28 healthy participants. This study presents a comparative analysis of two classification methods employed in decoding spoken words from EEG signals. Additionally, EEG signal processing and conditioning techniques are also discussed. In the first approach, spectral power features across various frequency bands are calculated to construct a feature vector, which subsequently serves as the input for an SVM classifier. In contrast, the second approach utilizes a convolutional neural network designed for EEG-based BCIs, allowing the data to be directly inputted into the model without prior feature extraction. In three of the four classification scenarios, the EEGNet method demonstrated superior performance compared to PSD + SVM. Furthermore, specific words or pairs of words were identified that demonstrate better classification accuracy than others, providing valuable insights for the development of BCIs aimed at assisting individuals with speech impairments.

Classifying between the attention segment and the pronunciation segment is the initial step in developing a brain–computer interface that operates based on attempted speech. The best performance was achieved in the classification scenario *Attention* vs.

Pronunciation, attaining an average accuracy of 91.04 ± 5.82% using PSD + SVM (see Table 1). The performance between the two classification methods was not significantly different in any of the performance metrics, as assessed using the Wilcoxon signed-rank test (accuracy: $p = 0.582$; recall class 1: $p = 0.304$; recall class 2: $p = 0.166$; precision class 1: $p = 0.186$; and precision class 2: $p = 0.399$). The recall per class is similar for the two classes and the two methods with mean values around 90%, indicating that the models are equally capable of correctly identifying the instances of both classes compared to the total instances of those classes. Altogether, these results show the possibility of discriminating between no pronunciation and pronounced words from EEG signals regardless of the classification method. In [47], using an artificial neural network (ANN), an average accuracy of 75.7% ± 9.6 was achieved in the classification of trials for imagined speech vs. rest; although the task differed, the two works have the same study objective. The sample size is believed to have influenced performance across binary classification scenarios (with 200, 80, and 40 trials per class); however, to confirm this, additional data collection or trial reduction in specific scenarios would be necessary.

Starting from the *Short words* vs. *Long words* scenario, EEGNet significantly outperformed PSD + SVM by approximately 8% (Wilcoxon signed-rank test, $p < 0.01$). An average accuracy of 73.91 ± 10.04% was achieved with EEGNet, compared to 65.72 ± 9.56% with PSD + SVM (see Table 2). Statistically significant differences were observed between the performances of the two methods across all evaluated metrics (Wilcoxon signed-rank test, $p < 0.01$). Notably, only the PSD + SVM method demonstrated significant differences between the recalls of each class (Wilcoxon signed-rank test, $p < 0.01$), suggesting that this method more readily identifies class 1 (short words) compared to class 2 (long words). Although EEGNet outperformed PSD + SVM, the results indicate that, regardless of the chosen method, it is possible to distinguish between short and long words. In [26], they use a CNN to classify between verbs and nouns of imagined speech and obtain an average accuracy of 93.8%, indicating that there is still ample room for improvements in this two-class classification.

Classifying one word against another serves as a starting point for the development of a binary BCI, for example, where decoding between the words *si* and *no* would enable responses to binary questions. Additionally, this also identifies which pairs of words are better decoded. In the *Word* vs. *Word* classification scenario, our findings suggest that EEGNet consistently outperforms PSD + SVM across all word pair combinations. Using PSD + SVM, the medians of the accuracies for all word pairs were above or very close to 60% and, in half of the cases, exceeded 70%. In contrast, with EEGNet, the medians of the accuracies were above 70% and, in some instances, surpassed 80%. Significant differences were observed in the performance between the two classification methods for eight out of the ten word pair classifications, as determined by the Wilcoxon signed-rank test. However, these differences were not statistically significant in the pairs *no* vs. *agua* and *agua* vs. *comida* (accuracy *no* vs. *agua*: $p = 0.130$; accuracy *agua* vs. *comida*: $p = 0.066$). In [29], the highest accuracy achieved per word pair was 64.55% using a deep CNN. In contrast, using EEGNet, we achieved an average accuracy of 81.23 ± 10.47% for the word pair *si* vs. *agua*, while using PSD + SVM, an accuracy of 73.97 ± 13.58% was attained for the pair *no* vs. *agua* (see Table 3). Furthermore, using EEGNet, we obtained an average accuracy of 76.61 ± 10.78% in the classification of *si* vs. *no*, a notable improvement from 63.2% ± 6.4 reported in [47] for the classification of *yes* vs. *no*. It is also important to highlight that significant differences between recalls were found when using PSD + SVM, specifically between the word pairs *agua* vs. *comida* and *agua* vs. *dormir*, with better recognition for the words *comida* and *dormir*. This suggests that while both methods can classify between word pairs, EEGNet provides a more balanced performance than PSD + SVM.

The final classification scenario was multi-class, wherein each of the five words constituted a separate class. The best performance was achieved using EEGNet, with an average accuracy of 54.87 ± 14.51% (see Table 4), reaching up to 80% accuracy in some participants. In contrast, for the PSD + SVM method, the average accuracy obtained was 41.78 ± 13.34%,

with some participants achieving an accuracy of up to 70% (see Figure A3 in Appendix B). In this classification scenario, a broader range of accuracy values was observed compared to previous scenarios (30–80% for EEGNet and 20–70% for PSD + SVM). This is likely because the chance level is lower in this multi-class scenario compared to the binary class scenarios (20% and 50%, respectively), and the complexity increases with more classes. In nine out of eleven calculated metrics, EEGNet significantly outperformed PSD + SVM. Surprisingly, for the word *agua,* no significant difference in recall and precision was observed between the two methods (Wilcoxon signed-rank test, $p > 0.01$). As in the previous two binary class scenarios, the PSD + SVM method showed significant differences in class-wise recall and precision (Friedman test, recall and precision per class: $p < 0.01$). According to the Nemenyi post hoc test ($p < 0.01$), *si* and *agua* were recognized more effectively (higher recall) than *dormir*. Additionally, *si* was detected with greater precision than *agua*, which in turn had higher precision than both *no* and *dormir*. These collective results demonstrate the feasibility of discriminating between five Spanish words from EEG signals using a convolutional neural network such as EEGNet.

One of the limitations of this study is the low spatial resolution of EEG signals, which we aim to compensate for through increased signal processing, acquiring more data for training, or testing different classification algorithms. In general terms, regarding the classification results obtained, there is a need to improve the performance of the models to achieve online application in patients with ALS. Thus, for future work, we plan to evaluate additional feature extraction algorithms, such as common spatial patterns, to identify a set of spatial patterns (i.e., electrode combinations) that maximize the differences between classes. Subsequently, we intend to compare this performance with that of EEGNet, which incorporates this methodology in one of its stages, and with other deep learning algorithms. Although it is already possible to decode online attempted speech in a patient with anarthria [19] or imagined speech in a patient with epilepsy [21], we aim to accomplish this by using a non-invasive method like EEG. Given that the average life expectancy for ALS patients is 2 to 5 years from the onset of the disease [48], and in [19], the study took 81 weeks, this represents a significant portion of the time of the target group (ALS patients) in terms of average life expectancy, in addition to the risks and costs associated with invasive methods. By focusing on spoken speech instead of imagined speech, we aim to leverage the planning and execution mechanisms in the language process. While we are aware that patients may not be able to execute speech in the same way as healthy participants, the next step in our research is to test the models proposed here on patient data while they attempt speech and fine-tune them until acceptable performance is achieved. This work presents promising results in using convolutional neural networks like EEGNet for decoding words from EEG signals. This preliminary study provides a foundation for a subsequent study to perform the decoding of attempted speech in ALS patients using a non-invasive method such as EEG.

5. Conclusions

In this study, we acquired an EEG dataset from 28 healthy participants pronouncing five Spanish words. A comparative analysis of two classification methods, PSD + SVM and EEGNet, was presented. The latter is a convolutional neural network designed explicitly for EEG-based brain–computer interfaces. In three of the four classification scenarios, EEGNet demonstrated superior performance compared to the PSD + SVM method. In all classification scenarios, it was shown that it is possible to distinguish between the different classes using EEG signals.

We obtained the best performance by classifying the attention segment concerning the speech segment, exceeding 90% accuracy in both methods. Although EEGNet outperformed PSD + SVM in the multiclass classification, the task presented challenges due to increased classes. Nonetheless, the results indicate the feasibility of discriminating between five Spanish words using EEG and EEGNet. Significant differences between decoding different words or groups of words were found only using the PSD + SVM method.

One of the limitations of using EEG is the low spatial resolution of signals. To improve this, it is proposed as future work to explore additional feature extraction algorithms and compare the performance with EEGNet and other deep learning algorithms. While it is already feasible to decode online attempted speech in patients with anarthria or imagined speech in epilepsy patients, the aim is to achieve this using a non-invasive method like EEG. Given the limited average lifespan for ALS patients, it is crucial to develop efficient non-invasive methods. In summary, this preliminary study has demonstrated the feasibility and potential of using convolutional neural networks, such as EEGNet, to decode words from EEG signals. These findings lay the foundation for future research to develop non-invasive brain–computer interfaces based on attempted speech, especially for patients with ALS.

Author Contributions: Conceptualization, D.A.-V., O.M.-M., R.C., H.R.M. and J.M.A.; methodology, D.A.-V., O.M.-M. and J.M.A.; software, D.A.-V., O.M.-M. and J.M.A.; validation, D.A.-V., O.M.-M. and J.M.A.; formal analysis, D.A.-V., O.M.-M. and J.M.A.; investigation, D.A.-V.; data curation, D.A.-V.; execution of experiments, D.A.-V., R.C. and H.R.M.; writing—original draft preparation, D.A.-V.; writing—review and editing, O.M.-M. and J.M.A.; supervision, O.M.-M. and J.M.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Ethics Committee of the Zambrano Hellion Hospital.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The datasets generated and analyzed for this study are available upon request to the corresponding author.

Acknowledgments: To Consejo Nacional de Humanidades, Ciencia y Tecnología (CONAHCYT) for the scholarship granted to carry out postgraduate studies.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ALS	Amyotrophic lateral sclerosis
ANN	Artificial neural network
BCI	Brain–computer interface
EEG	Electroencephalography
MI	Motor imagery
PSD	Power spectral density
SSVEP	Steady-state visually evoked potential
SVM	Support vector machine

Appendix A. ICA Implementation

For the signal component decomposition, we chose the ICA implementation known as *runica* from the FieldTrip toolbox in MATLAB. This implementation is a variant of *InfomaxICA*, which is a popular and widely used technique for EEG signal decomposition. Table A1 summarizes the parameter values used in the implementation.

Components corresponding to artifacts were identified through visual inspection of topographies and power spectra. For instance, an eye movement artifact presents a strong frontal signal. According to [49], the highest voltage peak corresponding to EEG signal contaminated by muscular activity in gesture movements is around 30 Hz. Even though the data were filtered from 1–30 Hz and the motion-related signal was substantially reduced, muscular components were detected in the EEG signal. A muscular artifact was observed as increased power at high frequencies, primarily in the temporal channels and their vicinity. We acknowledge that visual identification of artifacts can carry a degree of

subjectivity. Although we followed a strict protocol, the inter-individual variability in EEG signals can pose challenges in accurately identifying artifacts.

Table A1. Parameters used for ICA implementation.

Parameter	Value
numcomponent	Number of channels
pca	No dimensionality reduction performed
approach	'extended'
g	'tanh'
stopping	1×10^{-6}
maxsteps	512
lrate	'auto'
weights	Initialized based on data
sphere	Initialized based on data

Appendix B. Classification Results per Participant

Below are the accuracies calculated for each of the participants with the two methods.

Appendix B.1. Bi-Class: Attention vs. Pronunciation

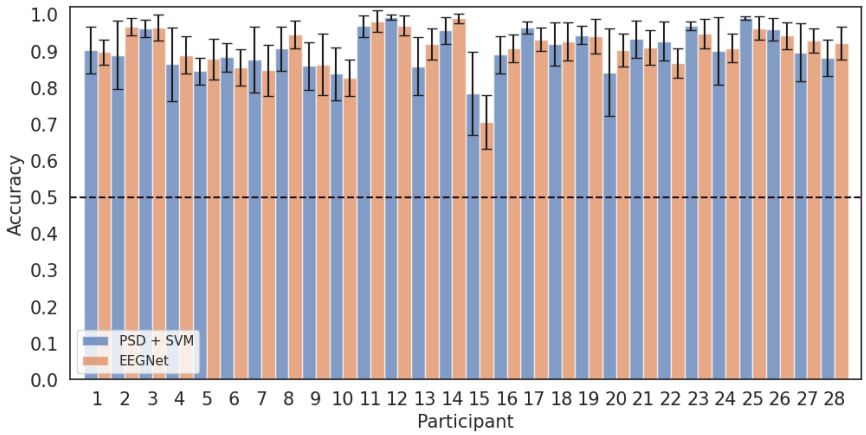

Figure A1. Accuracy results per participant by each method in classifying the attention vs. speech segment. The dotted black line corresponds to the level of chance (50%).

Appendix B.2. Bi-Class: Short Words vs. Long Words

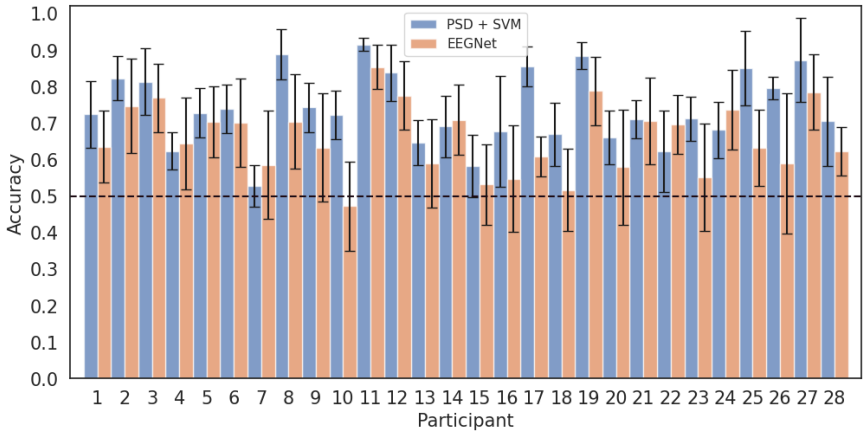

Figure A2. Accuracy results per participant by each method in classifying short words vs. long words. The dotted black line corresponds to the level of chance (50%).

Appendix B.3. Multiclass: All Words

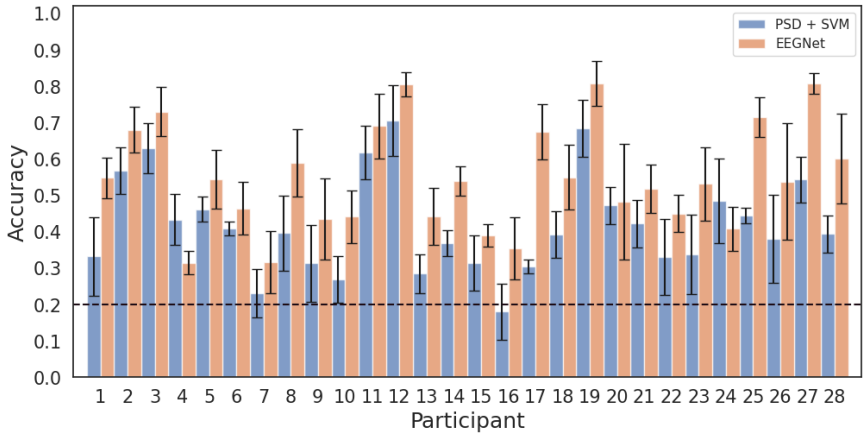

Figure A3. Accuracy results per participant by each method in multiclass classification. The dotted black line corresponds to the level of chance (20%).

References

1. Cuetos, F. *Neurociencia del Lenguaje: Bases Neurológicas e Implicaciones Clínicas*, 1st ed.; Editorial Médica Panamericana: Madrid, España, 2012; p. 111.
2. Ladefoged, P.; Johnson, K. *A Course in Phonetics*, 7th ed.; Cengage Learning: Boston, MA, USA, 2014; pp. 2–5.
3. Lee, J.; Madhavan, A.; Krajewski, E.; Lingenfelter, S. Assessment of dysarthria and dysphagia in patients with amyotrophic lateral sclerosis: Review of the current evidence. *Muscle Nerve* **2021**, *64*, 520–531. [CrossRef]
4. Caligari, M.; Godi, M.; Guglielmetti, S.; Franchignoni, F.; Nardone, A. Eye tracking communication devices in amyotrophic lateral sclerosis: Impact on disability and quality of life. *Amyotroph. Lateral Scler. Front. Degener.* **2013**, *14*, 546–552. [CrossRef]
5. Wolpaw, J.R. Brain–computer interfaces. *Handb. Clin. Neurol.* **2013**, *110*, 67–74.
6. Zhao, Y.; Chen, Y.; Cheng, K.; Huang, W. Artificial Intelligence Based Multimodal Language Decoding from Brain Activity: A Review. *Brain Res. Bull.* **2023**, *201*, 110713. [CrossRef]
7. Lotte, F.; Congedo, M.; Lécuyer, A.; Lamarche, F.; Arnaldi, B. A review of classification algorithms for EEG-based brain–computer interfaces. *J. Neural Eng.* **2007**, *4*, R1. [CrossRef]
8. Aggarwal, S.; Chugh, N. Review of machine learning techniques for EEG based brain computer interface. *Arch. Comput. Methods Eng.* **2022**, *29*, 3001–3020. [CrossRef]
9. Hernández-Rojas, L.G.; Montoya, O.M.; Antelis, J.M. Anticipatory detection of self-paced rehabilitative movements in the same upper limb from EEG signals. *IEEE Access* **2020**, *8*, 119728–119743. [CrossRef]
10. Rezeika, A.; Benda, M.; Stawicki, P.; Gembler, F.; Saboor, A.; Volosyak, I.Brain–computer interface spellers: A review. *Brain Sci.* **2018**, *8*, 57. [CrossRef]
11. Delijorge, J.; Mendoza-Montoya, O.; Gordillo, J.L.; Caraza, R.; Martinez, H.R.; Antelis, J.M. Evaluation of a p300-based brain-machine interface for a robotic hand-orthosis control. *Front. Neurosci.* **2020**, *14*, 589659. [CrossRef]
12. Hernandez-Rojas, L.G.; Cantillo-Negrete, J.; Mendoza-Montoya, O.; Carino-Escobar, R.I.; Leyva-Martinez, I.; Aguirre-Guemez, A.V.; Barrera-Ortiz, A.; Carrillo-Mora, P.; Antelis, J.M. Brain-computer interface controlled functional electrical stimulation: Evaluation with healthy subjects and spinal cord injury patients. *IEEE Access* **2022**, *10*, 46834–46852. [CrossRef]
13. Värbu, K.; Muhammad, N.; Muhammad, Y. Past, present, and future of EEG-based BCI applications. *Sensors* **2022**, *22*, 3331. [CrossRef]
14. Hekmatmanesh, A.; Azni, H.M.; Wu, H.; Afsharchi, M.; Li, M.; Handroos, H. Imaginary control of a mobile vehicle using deep learning algorithm: A brain computer interface study. *IEEE Access* **2021**, *10*, 20043–20052. [CrossRef]
15. Hekmatmanesh, A.; Wu, H.; Handroos, H. Largest Lyapunov Exponent Optimization for Control of a Bionic-Hand: A Brain Computer Interface Study. *Front. Rehabil. Sci.* **2022**, *2*, 802070. [CrossRef]
16. Kundu, S.; Ari, S. Brain-computer interface speller system for alternative communication: A review. *IRBM* **2022**, *43*, 317–324. [CrossRef]
17. Mannan, M.M.N.; Kamran, M.A; Kang, S.; Choi, H.S.; Jeong, M.Y. A hybrid speller design using eye tracking and SSVEP brain–computer interface. *Sensors* **2020**, *20*, 891. [CrossRef]
18. Nieto, N.; Peterson, V.; Rufiner, H.L.; Kamienkowski, J.E.; Spies, R. Thinking out loud, an open-access EEG-based BCI dataset for inner speech recognition. *Sci. Data* **2022**, *9*, 52. [CrossRef]
19. Moses, D.A.; Metzger, S.L.; Liu, J.R. Anumanchipalli, G.K.; Makin, J.G.; Sun, P.F.; Chartier, J.; Dougherty, M.E.; Liu, P.M.; Abrams, G.M.; et al. Neuroprosthesis for decoding speech in a paralyzed person with anarthria. *N. Engl. J. Med.* **2021**, *385*, 217–227. [CrossRef]

20. Anumanchipalli, G.K.; Chartier, J.; Chang, E.F. Speech synthesis from neural decoding of spoken sentences. *Nature* **2019**, *568*, 493–498. [CrossRef]
21. Angrick, M.; Ottenhoff, M.C.; Diener, L.; Ivucic, D.; Ivucic, G.; Goulis, S.; Saal, J.; Colon, A.J.; Wagner, L.; Krusienski, D.J.; et al. Real-time synthesis of imagined speech processes from minimally invasive recordings of neural activity. *Commun. Biol.* **2021**, *4*, 1055. [CrossRef]
22. Correia, J.M.; Jansma, B.; Hausfeld, L.; Kikkert, S.; Bonte, M. EEG decoding of spoken words in bilingual listeners: From words to language invariant semantic-conceptual representations. *Front. Psychol.* **2015**, *6*, 71. [CrossRef]
23. McMurray, B.; Sarrett, M.E.; Chiu, S.; Black, A.K.; Wang, A.; Canale, R.; Aslin, R.N. Decoding the temporal dynamics of spoken word and nonword processing from EEG. *NeuroImage* **2022**, *260*, 119457. [CrossRef] [PubMed]
24. Vorontsova, D.; Menshikov, I.; Zubov, A.; Orlov, K.; Rikunov, P.; Zvereva, E.; Flitman, L.; Lanikin, A.; Sokolova, A.; Markov, S.; et al. Silent EEG-speech recognition using convolutional and recurrent neural network with 85% accuracy of 9 words classification. *Sensors* **2021**, *21*, 6744. [CrossRef]
25. Rojas, S.J.B.; Ramírez-Valencia, R.; Alonso-Vázquez, D.; Caraza, R.; Martinez, H.R.; Mendoza-Montoya, O.; Antelis, J.M. Recognition of grammatical classes of overt speech using electrophysiological signals and machine learning. In Proceedings of the 2022 IEEE 4th International Conference on BioInspired Processing (BIP), Cartago, Costa Rica, 15–17 November 2022; pp. 1–6.
26. Datta, S.; Boulgouris, N.V. Recognition of grammatical class of imagined words from EEG signals using convolutional neural network. *Neurocomputing* **2021**, *465*, 301–309. [CrossRef]
27. Sarmiento, L.C.; Villamizar, S.; López, O.; Collazos, A.C.; Sarmiento, J.; Rodríguez, J.B. Recognition of EEG signals from imagined vowels using deep learning methods. *Sensors* **2021**, *21*, 6503. [CrossRef] [PubMed]
28. Nguyen, C.H.; Karavas, G.K.; Artemiadis, P. Inferring imagined speech using EEG signals: A new approach using Riemannian manifold features. *J. Neural Eng.* **2017**, *15*, 016002. [CrossRef]
29. Cooney, C.; Korik, A.; Raffaella, F.; Coyle, D. Classification of imagined spoken word-pairs using convolutional neural networks. In Proceedings of The 8th Graz BCI Conference, Graz, Austria, 19–20 September 2019; pp. 338–343.
30. Cooney, C.; Korik, A.; Folli, R.; Coyle, D. Evaluation of hyperparameter optimization in machine and deep learning methods for decoding imagined speech EEG. *Sensors* **2020**, *20*, 4629. [CrossRef]
31. Agarwal, P.; Kumar, S. Electroencephalography-based imagined speech recognition using deep long short-term memory network. *ETRI J.* **2022**, *44*, 672–685. [CrossRef]
32. Abdulghani, M.M.; Walters, W.L.; Abed, K.H. Imagined Speech Classification Using EEG and Deep Learning. *Bioengineering* **2023**, *10*, 649. [CrossRef]
33. García-Salinas, J.S.; Torres-García, A.A.; Reyes-Garćia, C.A.; Villaseñor-Pineda, L. Intra-subject class-incremental deep learning approach for EEG-based imagined speech recognition. *Biomed. Signal Process. Control.* **2023**, *81*, 104433. [CrossRef]
34. Fedorenko, E.; Thompson-Schill, S.L. Reworking the language network. *Trends Cogn. Sci.* **2014**, *18*, 120–126. [CrossRef]
35. Koizumi, K.; Ueda, K.; Nakao, M. Development of a Cognitive Brain-Machine Interface Based on a Visual Imagery Method. In Proceedings of the 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA, 18–21 July 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1062–1065.
36. Babadi, B.; Brown, E.N. A review of multitaper spectral analysis. *IEEE Trans. Biomed. Eng.* **2014**, *61*, 1555–1564. [CrossRef] [PubMed]
37. Mitra, P.P.; Pesaran, B. Analysis of dynamic brain imaging data. *Biophys. J.* **1999**, *76*, 691–708. [CrossRef] [PubMed]
38. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning with Applications in R*, 2nd ed.; Springer: New York, NY, USA, 2013; pp. 68—78.
39. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
40. Guler, I.; Ubeyli, E.D. Multiclass support vector machines for EEG-signals classification. *IEEE Trans. Inf. Technol. Biomed.* **2007**, *11*, 117–126. [CrossRef]
41. Panachakel, J.T.; Ramakrishnan, A. G. Decoding covert speech from EEG-a comprehensive review. *Front. Neurosci.* **2021**, *15*, 392. [CrossRef]
42. Lawhern, V.J.; Solon, A.J.; Waytowich, N.R.; Gordon, S. M.; Hung, C.P.; Lance, B.J. EEGNet: A compact convolutional neural network for EEG-based brain-computer interfaces. *J. Neural Eng.* **2018**, *15*, 056013. [CrossRef]
43. Ang, K.K.; Chin, Z.Y.; Zhang, H.; Guan, C. Filter bank common spatial pattern (FBCSP) in brain-computer interface. In Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, China, 1–8 June 2008; pp. 2390–2397.
44. Hastie, T.; Tibshirani, R.; Friedman, J. H. *The Elements of Statistical Learning Data Mining, Inference, and Prediction*, 2nd ed.; Springer: New York, NY, USA, 2009; pp. 392–395.
45. Japkowicz, N.; Shah, M. *Evaluating Learning Algorithms: A Classification Perspective*, 1st ed.; Cambridge University Press: New York, NY, USA, 2011; pp. 85–86.
46. Demšar, J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **2006**, *7*, 1–30.
47. Sereshkeh, A.R.; Trott, R.; Bricout, A.; Chau, T. EEG classification of covert speech using regularized neural networks. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 2292–2300. [CrossRef]

48. Masrori, P.; Van Damme, P. Amyotrophic lateral sclerosis: A clinical review. *Eur. J. Neurol.* **2020**, *27*, 1918–1929. [CrossRef]
49. Goncharova, I.I.; McFarland, D.J.; Vaughan, T.M.; Wolpaw, J.R. EMG contamination of EEG: Spectral and topographical characteristics. *Clin. Neurophysiol.* **2003**, *9*, 1580–1593. [CrossRef]