Cogent Arts & Humanities

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/oaah20

Investigation of brain response to acquisition and learning the second languages based on EEG signals and machine learning techniques

Talal A. Aldhaheri, Sonali B. Kulkarni, Pratibha R. Bhise & Baraq Ghaleb

To cite this article: Talal A. Aldhaheri, Sonali B. Kulkarni, Pratibha R. Bhise & Baraq Ghaleb (2024) Investigation of brain response to acquisition and learning the second languages based on EEG signals and machine learning techniques, Cogent Arts & Humanities, 11:1, 2416759, DOI: 10.1080/23311983.2024.2416759

To link to this article: https://doi.org/10.1080/23311983.2024.2416759

COGENT ARTS & HUMANITIES
2024, VOL. 11, NO. 1, 2416759
https://doi.org/10.1080/23311983.2024.2416759

LITERATURE, LINGUISTICS & CRITICISM | RESEARCH ARTICLE

OPEN ACCESS Check for updates

Investigation of brain response to acquisition and learning the second languages based on EEG signals and machine learning techniques

Talal A. Aldhaheri[a,b], Sonali B. Kulkarni[b], Pratibha R. Bhise[b] and Baraq Ghaleb[c]

[a]Faculty of Administrative and Computers Sciences, Albaydha University, Albaydha, Yemen; [b]Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Maharashtra, Aurangabad, India; [c]School of Computing at Edinburgh, Napier University, Edinburgh, UK

ABSTRACT

Brain-computer interfaces (BCI) and neurolinguistics have become vital areas of scientific inquiry, focusing on neural mechanisms in language acquisition. While studies have examined brain activity during language learning, there's a need for validated data on cognitive and functional effects of acquiring new languages, especially Arabic and Hindi. This study addresses this gap by recording and analyzing EEG data related to learning Arabic and Hindi as second languages, comparing linguistic differences during the process. EEG signals were recorded from eight participants (four Indian, four Yemeni) as they learned words in Arabic and Hindi. The data was pre-processed, cleaned, and analyzed to extract language learning-related features. To validate the approach and demonstrate cognitive and functional differences in brain activity during second language acquisition, various machine learning classification models were applied: Random Forest, Support Vector Machine, Decision Tree, Xgboost, and Catboost. The classifiers were trained and tested on the extracted features, achieving the following accuracies: RF 71.62%, SVM 68.41%, DT 64.12%, Xgboost 72.17%, and Catboost 74.56%. These results provide insights into neural mechanisms underlying second language acquisition. By comparing brain activity patterns between Arabic and Hindi, this study contributes to neurolinguistics and offers data that can be used to develop more effective language learning strategies and interventions.

ARTICLE HISTORY
Received 19 April 2024
Revised 30 September 2024
Accepted 11 October 2024

KEYWORDS
Brain-computer interfaces (BCI), electroencephalography (EEG), neurolinguistics; language acquisition; brain signals; spectroscopy

SUBJECTS
Artificial Intelligence; Behavioral Neuroscience; Language & Linguistics; Language Teaching & Learning

Introduction

Brain-computer interfaces (BCI) have emerged as one of the most critical areas of modern science, with researchers and developers working to create an interactive environment that implements all commands and instructions generated by the brain with no effort or muscle activity. This field depends entirely on advanced external hardware and software that simulates what the brain and the human body do to carry out specific tasks, such as movement, thinking, learning, practicing certain hobbies and many more (Lotte et al., 2018; Millán et al., 2010).

One of the sciences within this field is neurolinguistics, where researchers have been focusing in recent years to study and understand how languages are acquired and learned in the human brain. There are many techniques for recording and monitoring brain activity in both functional and cognitive states, including the electroencephalography (EEG) technique, which depends on placing a group of electrodes on the scalp (Aldhaheri et al., 2020; Rashid et al., 2020).

EEG is a non-invasive technique for monitoring and recording the functional and cognitive activity of the brain. It is a very old technique, but it remains highly effective in various fields, such as medicine (most famously for patients with epilepsy), education, marketing and cognitive and motor control for people with special needs. It provides these individuals with external devices and equipment

CONTACT Talal A. Aldhaheri talalalthahri@gmail.com Faculty of Administrative and Computers Sciences, Albaydha University, Albaydha, Yemen; Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Maharashtra, Aurangabad, India

related to their brain activity, and one of its fields is neurolinguistics (Jamil et al., 2021; Parvizi & Kastner, 2018).

The importance of this technique lies in its effectiveness in studying and monitoring many neurological disorders resulting from brain activity, such as epilepsy, Alzheimer's, stuttering and aphasia of both kinds (Alexander et al., 1990; Park et al., 2015). However, the handling and analysis of this type of signal are very complex and inflexible. This is primarily because this type of signal is non-linear, unstable, and described as chaotic to a great extent. Additionally, the presence of several recording channels that cover the entire brain activity during recording makes focusing on one aspect for analysis more complicated (Bhise et al., 2020; Tzimourta et al., 2021).

To address these challenges, scientists have developed open-source tools and programs to deal with and analyze such types of signals recorded in the electroencephalogram device. The most famous and important of these is EEGLAB, an open-source tool that works with the MATLAB (R2022a) program (Delorme & Makeig, 2004). Another significant tool is MNE-Python version 1.4.0 (Boston, MA), which runs on the Python platform. These two tools are considered among the most important available for signal analysis to extract all the required features from brain signals with the necessary accuracy in various fields. Consequently, researchers and scholars often combine the output from these tools with algorithms and techniques from machine learning and deep learning (Alexandre et al., 2013; Vaid et al., 2015). In this study, the research team utilized this software to conduct their experiments in a laboratory at the Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, located in Aurangabad, Maharashtra, India.

The complexity of the human brain is staggering, with the number of neurons estimated to outnumber the stars in the universe. These neurons interact and communicate in several intricate ways to conduct brain activities or complete processes that science has not yet been able to decipher accurately. It is known that the brain makes more than a million actual connections every day between neurons to practice daily routine activities such as sleeping, thinking, eating, learning and many more, including learning or acquiring a new language (Liao et al., 2012; Ofner et al., 2019; Rashid et al., 2020; Rosenfeld & Wong, 2017).

The data resulting from the learning process is a mixture of several functions, like other routine functions of the brain, which mix the resulting data with each other, such as thinking, listening, meditating, seeing, bodily movements and noise from the brain or from outside it. Given this complexity, researchers and scholars have focused on analyzing and reprocessing the resulting brain signals to obtain as much required and accurate data as possible for the subject of research or study in fields, such as neurolinguistics (Haufe et al., 2014; Ryan et al., 2017; Schultz et al., 2017).

Many people in the world speak at least two or more languages fluently (Hebb & Ojemann, 2013). However, there is a lack of clear scientific studies that show how the vocabulary of each language is stored in the brain and how it is represented cognitively. Questions remain about the extent to which the mother tongue affects other acquired languages or vice versa, and whether it is easy to switch between the vocabulary of acquired languages in the brain equally or if there are functional and cognitive representations that each language follows separately (Batterink & Neville, 2013; Moses et al., 2021; Rabbani et al., 2019).

This work is considered unique in studying and analyzing brain signals to learn or acquire one of the Arabic or Hindi languages. This study is regarded as the first of its kind among researchers and those interested in the field of neurolinguistics to study the extent of functional and cognitive effects and differences by EEG in attempting to learn and acquire one of two languages: Arabic or Hindi.

Our research focused on recording and collecting EEG signals from participants' brains during the acquisition and learning of a second language, with a specific emphasis on Hindi and Arabic for non-native speakers. We aimed to pre-process and analyze the brain signals recorded in an electroencephalogram device, cleaning and removing unwanted noise and traffic from the signals and extracting the relevant features of the study subject.

The purpose of our study was to collect clear and specific data that can be built upon in the field of neurolinguistics, particularly in acquiring and learning a second language. We aimed to contribute to this field by recording and analyzing the EEG data of volunteers' cues as they attempted to learn and acquire a language that was not their native language. To ensure the accuracy of our approach, we utilized several machine-learning techniques.

The results of our study provide valuable insights into the neural mechanisms underlying the process of learning and acquiring a second language. These findings can be used to develop more effective language-learning strategies and interventions, ultimately benefiting both the scientific community and society at large.

The article is structured as follows: Section 2 reviews related works. Section 3 covers materials and methods, including participants, data recording, recording procedure and raw data reduction and segmentation. Section 4 discusses EEG features extraction, encompassing distribution and primary spectroscopy, separation of multi-signals into components and final analysis of the power spectrum. Section 5 describes data splitting, while Section 6 details classification analysis. Section 7 outlines performance evaluation methods. Section 8 presents the results and discussion, and Section 9 concludes the article with a summary of the research and its implications.

Related works

Soman et al. (2019) investigated the use of overt speech and EEG signals for speech recognition. They employed a BESS F-32 amplifier with 32 passive electrodes to collect data from 17 participants speaking Indian languages (English, Japanese and Hindi). The study applied Support Vector Machines (SVMs) with a linear kernel for classification, achieving an accuracy of 63.32%. This study highlighted the potential for using EEG signals in speech recognition across multiple languages.

Hashim et al. (2018) focused on imagined speech recognition using EEG signals. They utilized an Emotiv Epoc sensor with 14 channels to collect data from 4 male participants. The study employed a k-Nearest Neighbors (k-NN) classifier for the recognition task, resulting in an accuracy of 58%. This study demonstrated the feasibility of using EEG signals for imagined speech recognition, although the accuracy is lower compared to overt speech recognition.

Sereshkeh et al. (2017a) investigated the use of imagined speech and EEG signals for speech recognition. They used a BrainAmp EEG system with 64 electrodes and a DC amplifier to collect data from 12 fluent English speakers. The study applied a Multi-Layer Perceptron Neural Network (MLPNN) classifier for the recognition task, achieving an accuracy of $54.1 \pm 9.7\%$. This study highlights the challenges associated with imagined speech recognition using EEG signals.

In a subsequent study, Sereshkeh et al. (2017b) used the same data collection method as their previous work but applied a linear SVMs classifier for the recognition task. This approach resulted in an improved accuracy of 69.27% compared to the MLPNN classifier used in their earlier study. This research demonstrated the importance of selecting an appropriate classifier for imagined speech recognition using EEG signals.

González-Castañeda et al. (2017) studied imagined speech recognition using EEG signals. They employed an EEG system with 14 high-resolution channels to collect data from 27 native Spanish speakers. The study compared the performance of SVM and Naive Bayes classifiers, with the latter achieving an accuracy of 83.34%. This study highlighted the potential for using high-resolution EEG channels and probabilistic classifiers for imagined speech recognition.

Nguyen et al. (2018) investigated imagined speech recognition using EEG signals. They used a BrainProducts ActiCHamp EEG system with 64 electrodes and an amplifier to collect data on three different types of speech (long, short words, and vowels). The study applied a Relevance Vector Machines (RVM) classifier for the recognition task, achieving accuracies of 95% for 2 words and 70% for 3 words. This study demonstrated the potential for using EEG signals and RVM classifiers for imagined speech recognition, particularly for a limited set of words.

Kang et al. (2015) studied imagined speech recognition using a combination of eye-tracking and EEG signals. They employed a Tobii 1750 eye-tracker and a Brainno device with 2 EEG channels to collect data from 63 English speakers. The study used an SVM classifier for the recognition task, achieving an accuracy of $80.16 \pm 0.14\%$. This study showed the potential for using a multimodal approach, combining eye-tracking and EEG signals, for imagined speech recognition.

Kumar et al. (2018) focused on imagined speech recognition using EEG signals. They utilized an Emotiv EPOC+ with 14 electrodes to collect data from 23 participants on images, numbers and texts. The study

applied a Random Forest (RF) classifier for the recognition task, achieving accuracies ranging from 67.03% to 85.20%. This study demonstrates the potential for using EEG signals and RF classifiers for imagined speech recognition across different types of stimuli.

Prat et al. (2016) investigated overt speech recognition using quantitative EEG (qEEG) signals. They employed a wireless 16-channel EPOC qEEG to collect data from 16 French-English speakers. The study applied the Mini-Mental State Examination (MMSE) for the recognition task, achieving an accuracy of 60%. This study highlighted the potential for using qEEG signals and cognitive assessment tools for overt speech recognition in bilingual speakers.

Rahma and Nurhadi (2017) studied overt speech recognition using EEG signals. They used 4 EEG channels with a maximum impedance of 15 Ω to collect data from 16 English speakers using the openBCI platform. The study conducted Power Spectral Density (PSD) analysis and found that males had higher accuracies than females. This study demonstrates the potential for using a limited number of EEG channels and PSD analysis for overt speech recognition while highlighting gender differences in accuracy.

Dave et al. (2018) investigated overt speech recognition using EEG signals. They employed a SCAN (Compumedics Neuroscan) EEG system with 29 tin electrodes and an elastic cap (ElectroCap International) to collect data from 60 young adults and 36 older adults speaking English. The study performed PSD analyses but did not report specific accuracy rates. This study highlights the potential for using EEG signals and PSD analysis for overt speech recognition across different age groups.

Grundy et al. (2019) focused on overt speech recognition using EEG signals. They utilized a 64-channel Biosemi ActiveTwo EEG system and E-prime version 2.0 software (Pittsburgh, PA) to collect data from 40 participants (20 monolingual and 20 bilingual). The study applied Multi-scale Entropy Analysis (MSE) for the recognition task, achieving accuracies of 0.94 for monolinguals and 0.93 for bilinguals. This study demonstrated the potential for using high-density EEG signals and MSE analysis for overt speech recognition in both monolingual and bilingual speakers.

Liu et al. (2017) investigated overt speech recognition using EEG signals. They employed a 7-channel EEG headset called Muse to collect data from eleven English-speaking volunteers. The study used logistic regression (LR) with local and global models and multi-task learning for the recognition task, achieving accuracies ranging from 54.97% to 55.01%. This study highlighted the potential for using consumer-grade EEG headsets and advanced machine-learning techniques for overt speech recognition. More detail in Table 1.

This study extends previous research in EEG-based language processing by focusing on the neural mechanisms of language acquisition, specifically for Arabic and Hindi. Unlike prior studies that primarily examined speech recognition in established languages (e.g. Soman et al., 2019; Hashim et al., 2018), this research investigates brain activity during the active learning of new languages with distinct scripts and phonological systems.

Methodologically, this study employs a comprehensive approach, combining EEG data collection during word learning tasks, sophisticated signal processing using EEGLAB and multiple machine learning algorithms for classification. The achieved classifier accuracies (64.12%–74.56%) are comparable to or exceed those reported in previous studies, despite the increased complexity of the language acquisition task.

Furthermore, this study provides novel insights into the formation and evolution of neural representations for newly learned words, offering a dynamic, longitudinal perspective on language acquisition. This contrasts with many previous studies that focused on single-session recordings (e.g. González-Castañeda et al., 2017; Prat et al., 2016).

By addressing these aspects, this study contributes significantly to the field of neurolinguistics and EEG-based language research, with potential implications for language education and cognitive rehabilitation. It advances our understanding of the neural processes underlying new language acquisition, an area not extensively explored in previous works.

Materials and methods

In this section, the data recorded and collected by us has been studied, analyzed and pre-processed in order to remove noise and traffic that weaken its value and accuracy. To eventually become standard data that can be relied upon by scholars and researchers in the field of neurolinguistics related to EEG.

Table 1. Summary of previous research studies on neurolinguistics by EEG.

Author and year	Type of speech	Recording technique	Data acquisition device	Data collection	Classifier	Accuracy%
Soman et al. (2019)	Overt	EEG	BESS F-32 amplifier – 32 passive electrodes	17 Indian (English, Japanese and Hindi)	SVM and linear kernel	63.32
Hashim et al. (2018)	Imagined	EEG	Emotiv Epoc sensor –14 channels	4 Males	k-NN	58
Sereshkeh et al. (2017a)	Imagined	EEG	BrainAmp – 64 electrodes and DC amplifier	12 (Fluently English)	MLPNN	54.1 ± 9.7
Sereshkeh et al. (2017b)	Imagined	EEG	BrainAmp – 64 electrodes and DC amplifier	12 (Fluently English)	linear SVM	69.27
González-Castañeda et al. (2017)	Imagined	EEG	High-resolution channels (14 electrodes)	27 (Native Spanish)	- SVM Naive Bayes is a probabilistic classifier	83.34
Nguyen et al. (2018)	Imagined	EEG	BrainProducts ActiCHamp Includes 64 electrodes with amplifier	3 Different types (long, short words and vowels)	Relevance Vector Machines classifier (RVM)	(2 words 95) (3 words 70)
Kang et al. (2015)	Imagined	EEG	Tobii 1750 eye-tracker, Brainno device-2 channels	63 (English)	SVM	80.16 ± 0.14
Kumar et al. (2018)	Imagined	EEG	Emotiv EPOC+ 14 electrodes	23 (Images, numbers, and texts).	RF	67.03–85.20
Prat et al. (2016)	Overt	qEEG	wireless EPOC 16-channel	16 Speaking (French-English)	Mini-Mental State Examination (MMSE)	60
Rahma and Nurhadi (2019)	Overt	EEG	4-EEG channels with a max impedance of 15Ω.	16 (English) by openBCI	PSD analysis	Males higher than females
Dave et al. (2017)	Overt	EEG	SCAN (Compumedics Neuroscan) consists of 29 tin-electrodes Included an elastic cap (ElectroCap International)	60 Young adults and 36 older adults speaking (English).	PSD analyses	–
Grundy et al. (2019)	Overt	EEG	64-channel Biosemi ActiveTwo and E-prime 2.0	Forty (20 monolingual and 20 bilingual)	Multi-scale Entropy Analysis (MSE)	Mono/0.94 and bili/0.93
Liu et al. (2017)	Overt	EEG	headset called Muse has 7-channels	Eleven volunteer speaking (English)	LR (Local + Global models) And Multi-task Learning	54.97–55.01

Figure 1 represents the main steps of the research methodology. Each step leads to the next, showing the sequential process of the study from data collection through to final performance evaluation.

3.1. Participants

The nine participants were five males and four females, aged 18–39 (mean: 29.87), equally holding Indian and Yemeni nationalities, and none of them suffered from any chronic diseases or neurological disorders as per the previous written consent they signed so that signals from their brains could be recorded. One participant's data was ignored due to the loud noise within the signals recorded from his brain. This could be due to an error related to the recording device or a malfunction. Table 2 shows more details about the volunteers in this study. The subjects are the participants in our data recording, so each participant is denoted by a letter (S) and a number (1, 2, … 8).

3.2. Data recording

The data were collected and fully recorded in the Medicover Hospital in Aurangabad, India, for more than two months of training, preparation, and final registration. We used the Virgo EEG device produced by Allengers, which includes 40 active electrodes placed according to the 10–20 system with

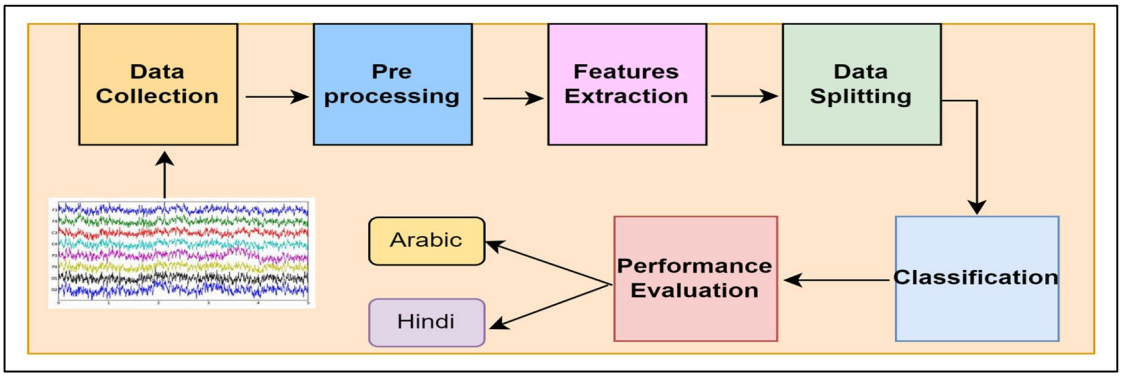

Figure 1. Our proposed methodology.

Table 2. Information of participants in detail note the number of sessions for each of the participant.

Subjects	S1	S2	S3	S4	S5	S6	S7	S8
Nationality	Indian	Yemeni	Yemeni	Indian	Indian	Yemeni	Indian	Yemeni
Gender	M	M	M	F	F	M	F	F
Age	31	30	34	26	29	39	18	32
Native language	Hindi	Arabic	Arabic	Hindi	Hindi	Arabic	Hindi	Arabic
Recording device	EEG	EEG	EEG	EEG	EEG	EEG	EEG	EEG
Sessions	12	12	12	12	12	12	12	12

a resistance of less than 10 kOhm at a frequency of 256 Hz, confirming the presence of electrodes on both sides of the brain (Allengers Global, 2022). All data registration sessions were subject to all required procedures in accordance with the research and scientific standards followed (Kawala-Sterniuk et al., 2020), whether in general or the procedures followed within the hospital under registration, as shown in Figure 2. The volunteers were placed in a closed environment free from noise or loud influences in order to prevent the largest amount of noise associated with brain signals (only the minimum limit associated with the recording device or generated from the brain itself). The duration of recording and data collection took a long time due to the authors' difficulty in acquiring an expensive EEG device. Each participant recorded twelve full recording sessions for the two languages under study. The best six sessions were adopted after taking the highest and best average for all sessions in terms of recording accuracy and the amount of noise and damage within the recorded signals.

The group of stimuli in each recording session included written and shown words associated with a picture that expresses the content of the word. where the words were chosen almost uniformly in the period, with 12 Arabic words (the language familiar to Yemenis) and 12 Hindi words (the language familiar to Indians) as in Table 3, with an interval of 5 s for each word.

3.3. Procedure of recording

After taking the written consent of all the volunteers and arranging the procedures used to record brain signals, each experiment was divided into several sessions, according to Figure 3. Each session consisted of three basic stages. In the first stage, the participant takes a rest period of 5 s, during which he is ready to focus on the words and start the recording. This is followed by the stage called the baseline, which marks the beginning of the actual time of the recording with a period of 2.5 s. The last stage, which is the longest in terms of time, is known as the stage of the actual recording of the words displayed on the computer screen in an equal number of Arabic and Hindi words as a language unfamiliar to the participants equally. At this stage, each word is displayed for an amount of time equal to 5 s. The total number of words presented for registration in each session was 24 for both languages. components of the experiments: a computer for displaying words, an EEG device for recording brain signals and monitoring functional and cognitive activity, and a table and chair in an isolated environment.

Figure 2. Three of the participants in the recording sessions during the acquisition of data.

Table 3. The set of recorded words in both languages.

Arabic	Hindi	English	Duration (s)
سيارة	गाड़ी	Car	5.00
أم	माां	Mother	5.00
بيت	मकान	House	5.00
كتاب	पुस्तक	Book	5.00
معلم	अध्यापक	Teacher	5.00
الوالد	पपता	Father	5.00
يوم	पनि	Day	5.00
ينام	सो रहा	Sleeping	5.00
بحر	समुद्र	Sea	5.00
أسبوع	सप्ताह	Week	5.00
أخت	बहन	Sister	5.00
لا	ना	No	5.00

Arabic and Hindi are phonetic languages. English words to clarify the meaning.

3.4. Reduction and segmentation of raw rata

The data generated after the recording stage is raw data represented by dynamic signals. In our study, we saved each session's data for each participant in European Data Format (EDF) without any pre-processing (Kemp et al., 2010). Using the EDFBROWSER application, it was then divided into several files with a predetermined time depending on the language to be studied (Gurumoorthy et al., 2020). We then labeled each file associated with the participant with a specific number to distinguish between the data of the Arab and Indian participants. This stage, known as 'labelling RAW EEG,' is where the final separation occurs between all registration sessions regarding the participant's language type and gender. All the previous steps are in preparation for the pre-processing of the recorded data using the MATLAB

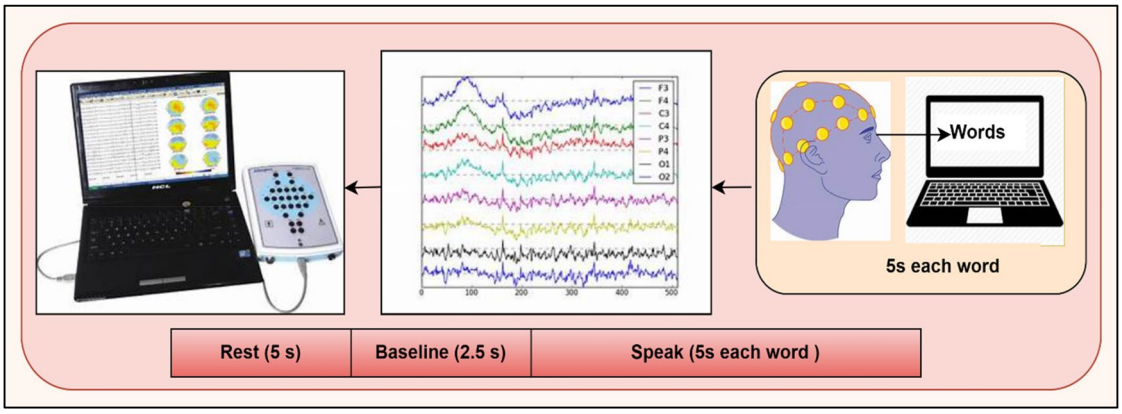

Figure 3. Experimental design used for data collection.

program, specifically the open-source program EEGLAB (Delorme & Makeig, 2004). We applied several sequential steps to clean and purify the signals (unprocessed raw data) from noise and traffic, as shown in Figure 4. The data went through two cases of internal and external purification using many available and proven techniques and methods for processing, such as Independent Component Analysis (ICA) (Bugli & Lambert, 2006; Subasi & Gursoy, 2010). The data was cleaned of signals with frequencies outside the required range to distinguish the spoken words in both languages at a given frequency (0.1 and 0.75 Hz) and purify the flat channel signals that recorded zero voltage. Additionally, certain parts of the generated brain signals were cut off due to the intensity of the noise at specific time intervals.

The PSD of the EEG data was visualized for each channel, with the theta (4–8 Hz), alpha (8–12 Hz) and beta (13–30 Hz) bands highlighted. The red dashed line indicates the 50 Hz frequency, typically associated with power line noise. From Figure 5, we observed the distribution of power across frequencies for each channel. The highlighted areas help in identifying dominant power in the theta, alpha and Beta bands, which are consistent with signals relevant to language processing. The presence of a peak at or around 50 Hz would indicate power line noise.

Given the presence of 50 Hz power line artifacts, we applied a narrow notch filter to remove this narrow-band noise and its harmonics. The notch filter successfully eliminated the 50 Hz peak with minimal effects on the broader EEG spectrum. A notch filter targeted 50 Hz to remove power line noise from the EEG data. The PSD of the filtered data showed reduced 50 Hz power, indicating the filter worked. The broader EEG spectrum remained unchanged aside from the filtered 50 Hz noise. A 50 Hz notch filter was applied to the EEG data to remove power line noise, evident in the PSD visualization of the filtered data. The PSD as shown in Figure 6 highlighted the theta (4–8 Hz), alpha (8–12 Hz) and beta (13–30 Hz) bands. A dashed line at 50 Hz indicated the notch filter's target frequency. Compared to the unfiltered PSD, the 50 Hz power was markedly reduced, demonstrating effective attenuation of power line noise. The remaining EEG spectrum stayed largely intact, with the notch filter specifically targeting and removing the 50 Hz interference.

After completing all the pre-processing stages, we obtained filtered data from most forms of noise. Additionally, we selected the data channels related to the areas of learning and language acquisition in the left hemisphere of the brain. This step reduced the number of recorded channels in the data and the size and number of data samples, preparing them for training and testing using machine learning techniques. Table 4 shows the channels used and the final samples. The data was converted and saved in.csv format after being processed from .edf and .mat formats to allow for accuracy training and testing.

EEG features extracting

Purification and filtering of signals from all forms of unwanted noise and frequencies is critical to initializing an EEG signal to extract specific features about the subject of study. There are many techniques and algorithms for pre-processing and analysis to extract specific features. We are not here to discuss it in detail. To extract features from the EEG signals used in our study, we used some of them. Figure 7 shows the steps involved in the feature extraction process.

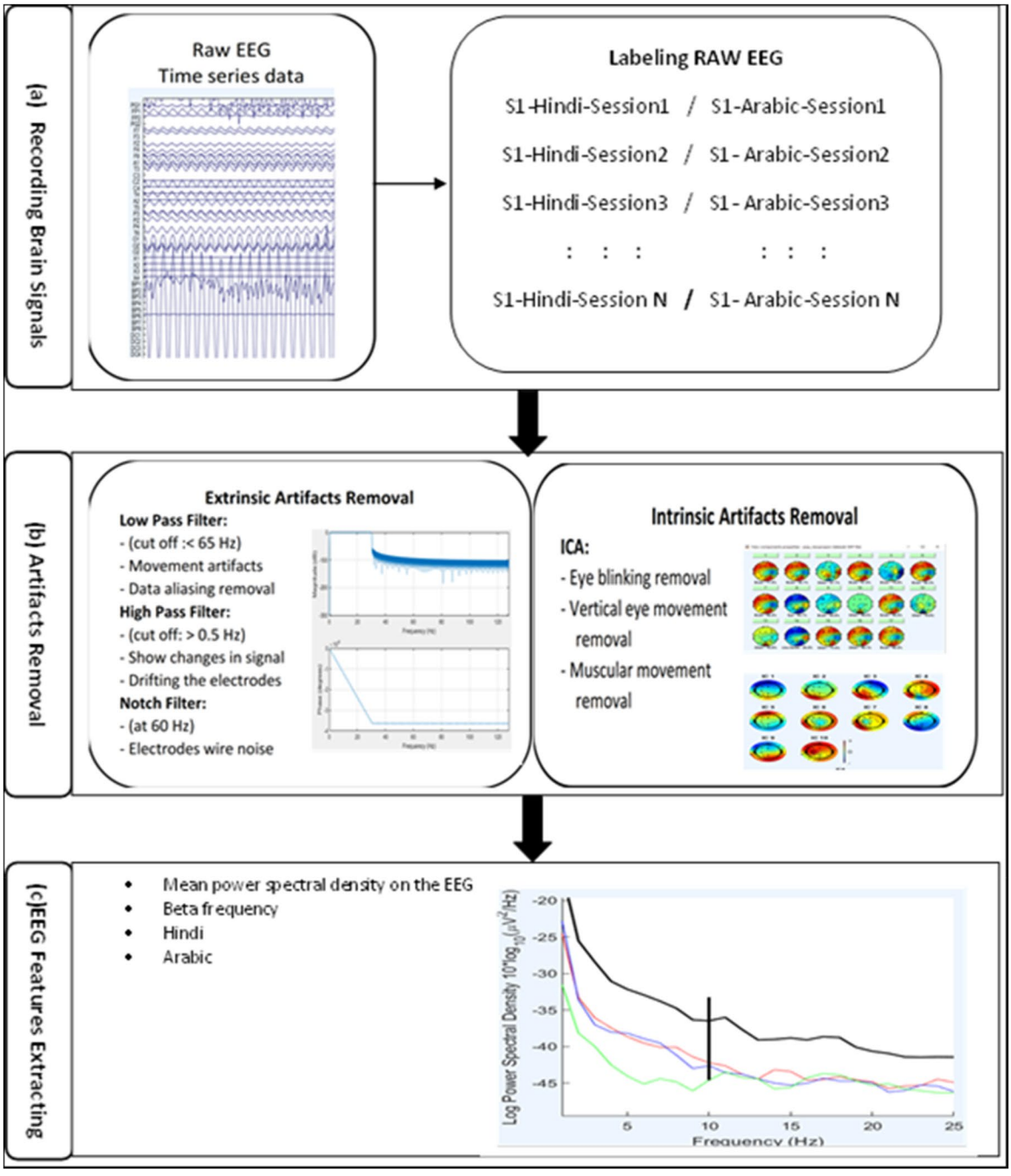

Figure 4. Raw data Segmentation and Reduction: (a) Recording across 40 electrodes and data labelling; (b) Steps to remove extrinsic and Intrinsic artifacts; (c) EEG features extracting.

4.1. Distribution and primary spectroscopy

In the initial analysis or primary processing of the recorded signals by EEG technique, we note the contrast, difference and overlap for all power spectra between channels as in Figure 8. This discrepancy results in difficulty in analyzing and inferring the constant frequencies of beta, alpha, theta and delta waves. with high-frequency spectral energies for some waves, such as delta and beta. The impact of this problem or difficulty can be reduced by using ICA, whereby the power spectrum of overlapping or dissimilar signals is separated into separate and individual components that facilitate analysis and processing.

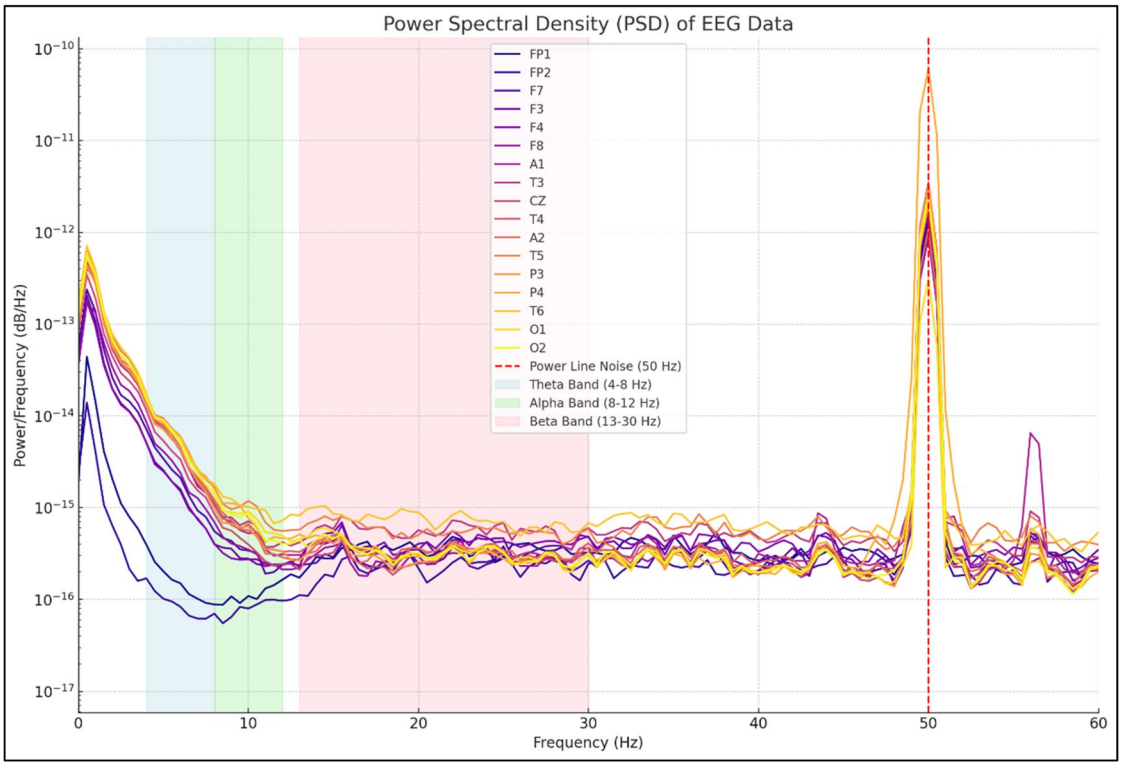

Figure 5. PSD for the original EEG data Theta, alpha and beta bands highlighted. Red dashed line indicates power line noise at 50 Hz.

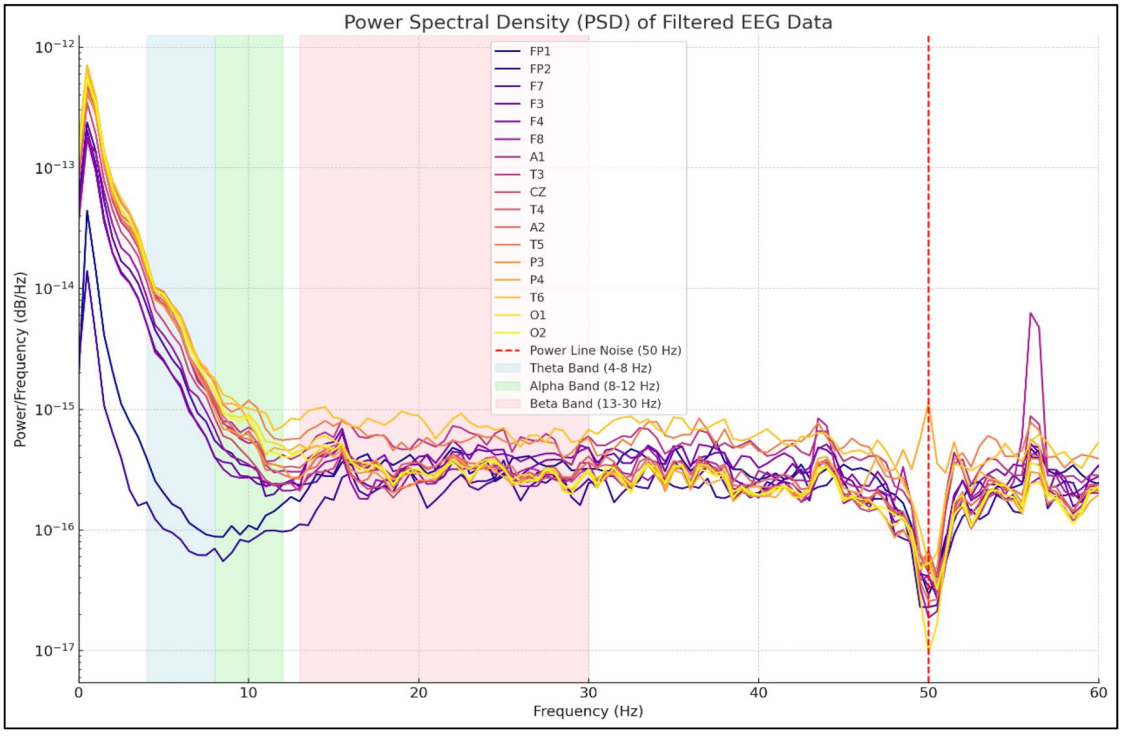

Figure 6. Power spectral density (PSD) of EEG data after applying notch filter at 50 Hz. The filtered data has attenuated 50 Hz power while leaving the theta, alpha and beta bands intact, demonstrating effective removal of power line artifacts by the notch filter. A dashed line marks the 50 Hz target frequency.

Table 4. Data properties before and after processing.

	During recording	After processing
Subjects	S1, S2, S3, S4, S5, S6, S7, S8 and S9	S1, S2, S3, S4, S5, S6, S7 and S8
Sessions	108 Sessions	48 Sessions
Number of samples	216 Files as raw data	96 Files as processing data
N/Name of channels	'(40) PG1, FP1, FP2, PG2, F7, F3, FZ, F4, F8, A1, T3, C3, CZ, C4, T4, A2, T5, P3, PZ, P4, T6, O1, OZ, O2, X1, X2, X3, X4, BP1, BP2, BP3, BP4 BP5, BP6, BP7, BP8, DC1, DC2, DC3 & DC4'	(17) FP1, FP2, F7, F3, F4, F8, A1, T3, CZ, T4, A2, T5, P3, P4, T6, O1, O2
Brain regions	Frontal, temporal, parietal and occipital lobes	Frontal and temporal lobes
Data type	Raw data (dynamic signals)	Numeric data

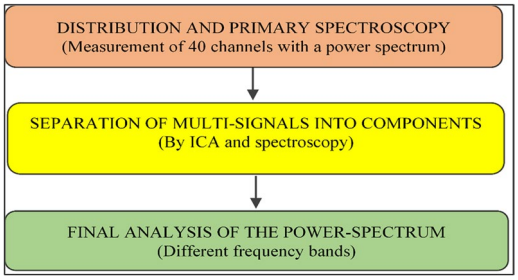

Figure 7. Diagram of the steps involved in the feature extraction.

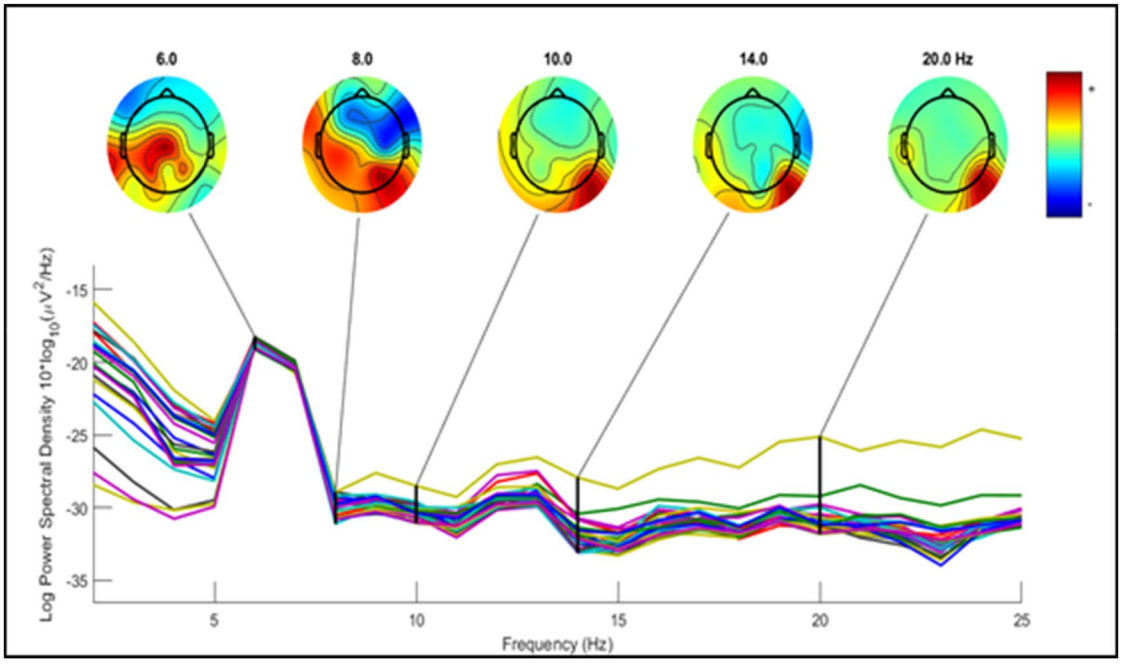

Figure 8. Overlapping and dissimilar channels in the power spectra.

4.2. Separation of multi-signals into components

The greater the number of electrodes or channels for recording brain signals, the more valuable and unhelpful information is generated, and this makes it difficult to study a specific functional or cognitive part of the brain, for example, diseases (such as epilepsy and Alzheimer's), movement or stuttering, including learning and acquiring a second language, and many other functions resulting from the brain. Fourier analysis and wavelet transform are two methods that are commonly used and best suited for analyzing EEG data generated from each channel separately. Still, it is not appropriate or ideal for

predicting data from multi-channel recordings due to its random and non-stationary properties (Haufe et al., 2014; Ryan et al., 2017). Here, the ICA method emerges due to its distinctive statistical characteristics, such as the statistical separation of the signals of independent sources from the multi-channel signals as shown in Figure 9(a).

It can also detect all subsequent changes in the brain by recording the electroencephalogram and estimating and distinguishing each signal separately, as shown in Figure 9(b). The extended ICA algorithm or the logistic Infomax ICA algorithm with natural scaling feature is optionally used in the ICA process (Hebb & Ojemann, 2013).

4.3. Final analysis of the power spectrum

EEG signals are passed into digital bandpass filters for analysis and separation into fixed and known frequencies, such as the delta-band (1–4 Hz), the theta-band (4.1–8 Hz), the alpha-band (8.1–13 Hz) and the beta-band (13.1–30 Hz) as in Figure 10.

When comparing the resulting energy levels for the same words in the two languages, we notice that there are different and varying frequency bands. This is evidence of functional and cognitive differences when trying to learn or acquire new language words, indicating that the learning and memorization mechanisms in the brain are not equal. In general, the results showed that the highest power range reached was 21 dB, and the lowest was −20 dB, as shown in Table 5.

When calculating the frequency domain characteristics of the signals, we applied the PSD, which is the basis for this purpose, as shown in Figure 8(a,b). We used non-parametric methods, such as Fourier transform, regularly computed through 'Short Time Fourier-Transform (STFT)' and the 'Fast Fourier-Transform (FFT)' algorithm for the signals, as shown in Figure 11(a,b). In general, spectrograms are the most effective tool for analyzing and processing speech signals.

Spectroscopy showed for random sample (S2 and S6) the highest brain activity occurred between 0 and 700 ms, in Figure 12(a) within the frequency band of 10–26 Hz. And the highest brain activity

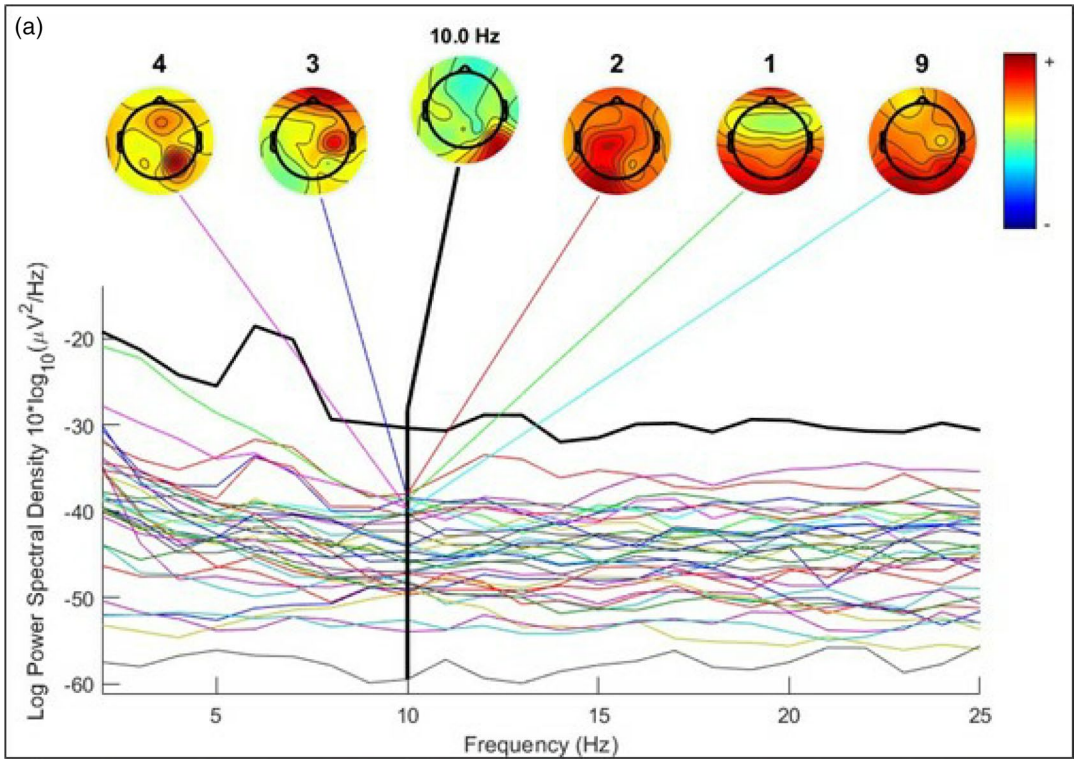

Figure 9a. By means of independent component analysis, multiple and intertwined signals at any frequency level were separated into their components by statistically independent sources. This example illustrates separation at 10 Hz, where we clearly notice a variation in the amount of power for each component.

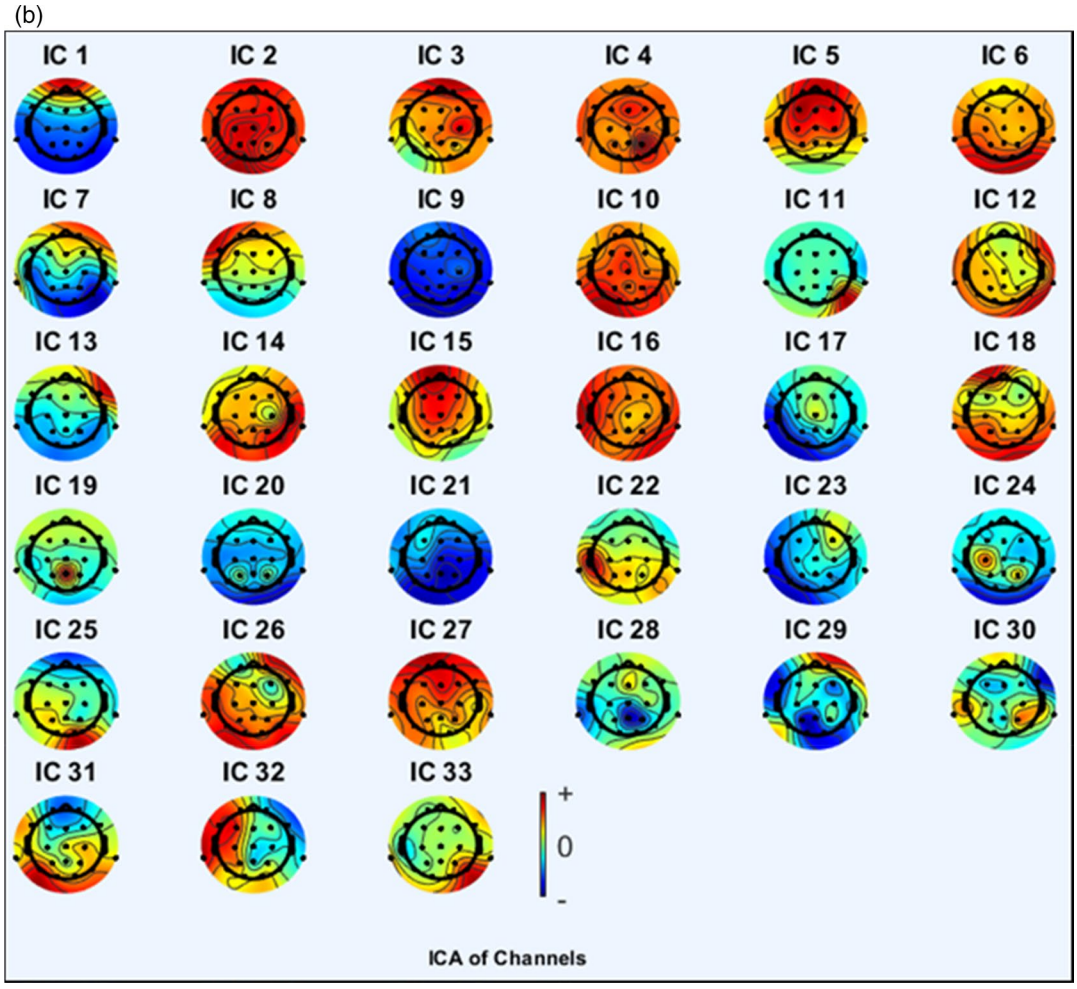

Figure 9b. Signals recorded in several channels whose independent sources are statistically separated by the independent components analysis.

occurred between 0 and 700 ms, in Figure 12(b) within the frequency band of 10–30 Hz. These ranges of frequencies are the same as the alpha and beta bands, which oversee all language activity.

Data splitting

The pre-processed EEG data was divided into training and testing sets using an 80/20 split. In total, we get processed 96 files of data, as detailed in Table 4 of our manuscript. The 20% (8602 data points) testing set comprises approximately 19 files (calculated as 96 * 0.2 = 19.2, rounded down to 19). The remaining 80% (34,408 data points), amounting to 77 files, was utilized for training the model, while 20% was reserved for evaluating performance on unseen data. To ensure that the class distribution was preserved in both the training and testing sets, we employed stratified sampling during the splitting process.

Classification analysis

Our aim is to discover and identify the possible relationship between two languages through brain activity, whether functional or cognitive. We have developed a model to observe the potential correlation between the two language signals, depending on the training set used to train the proposed model. We prepared and tested different classifications in machine learning (Lotte et al., 2018), such as RF, SVM, DT, XGBoost and CatBoost, to compare the accuracy of the proposed model for gathering as much information as possible about the brain during the process of learning and acquiring a second language. To optimize the performance of each classifier, we carefully tuned their respective hyperparameters. For the RF classifier,

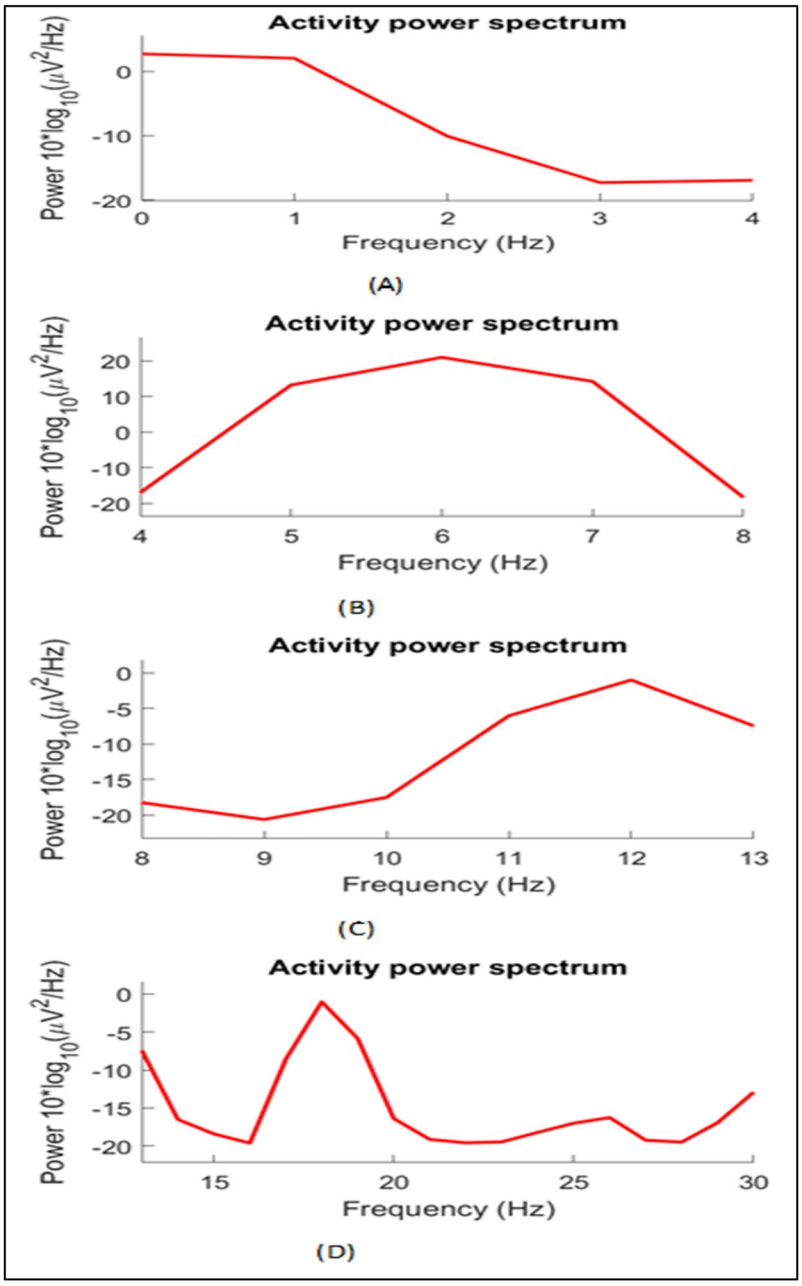

Figure 10. Checking the popular spectra of powers: (a) delta frequency band, (b) theta frequency band, (c) alpha frequency band and (d) beta frequency band.

Table 5. The set of recorded words in both languages.

Frequency-bands	DELTA (0.1–4.0)	THETA (4.1–8.0)	ALPHA (8.1–13.0)	BETA (13.1–30.0)
Max. power (dB)	0–5	20	−1–0	−3 to −1
Min. power (dB)	−18 to −20	−19	−20	−19 to −20
Max. power Freq (Hz)	0–1	5.8–6.2	12	17.5–18.5
Min. power Freq (Hz)	3–4	8	9	16

Arabic and Hindi are phonetic languages. English words to clarify the meaning.

we set the number of trees (n_estimators) to 100 and the maximum depth of each tree (max_depth) to None, allowing the trees to grow until all leaves contain less than the minimum number of samples required to split an internal node (min_samples_split), which was set to 2. In the case of the SVM classifier, we employed the Radial Basis Function (RBF) kernel, which enables the model to capture non-linear relationships between the features. The regularization parameter (C) was set to 1.0, striking a balance between

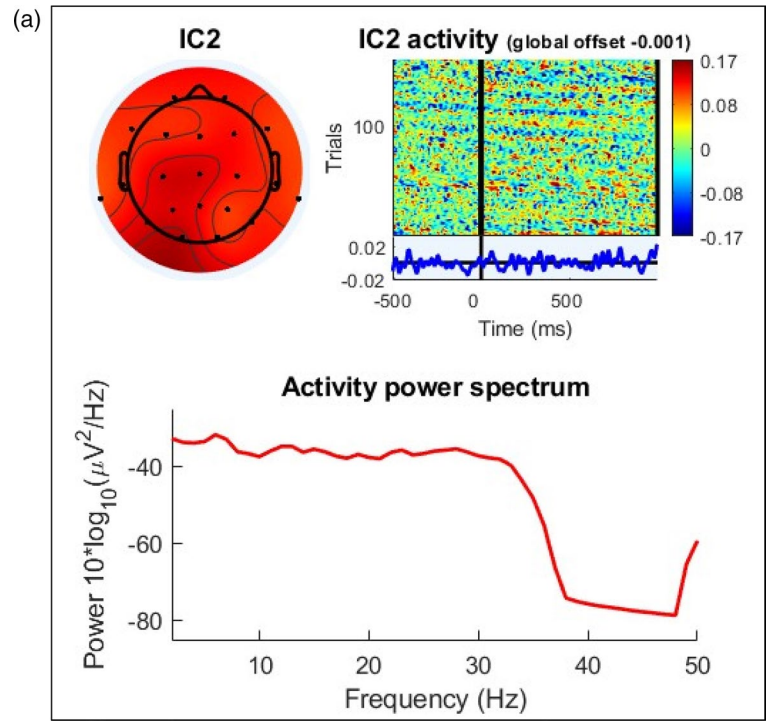

Figure 11a. The power spectrum density (PSD) of Channel 2 by participant 3 when he pronounced Arabic words.

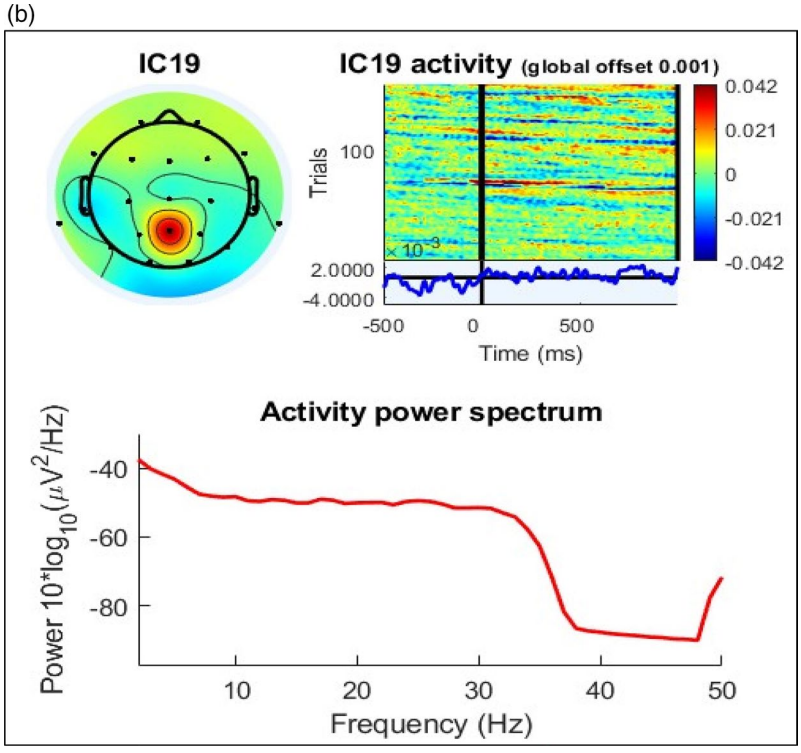

Figure 11b. The power spectrum density (PSD) of Channel 19 by participant 3 when he pronounced Hindi words.

achieving low training error and maintaining model simplicity. The kernel coefficient (gamma) was set to 0.1, determining the influence of individual training examples on the decision boundary. For the DT classifier, we set the maximum depth (max_depth) to None, allowing the tree to grow until all leaves are pure or contain less than the minimum number of samples required to split a node (min_samples_split), which

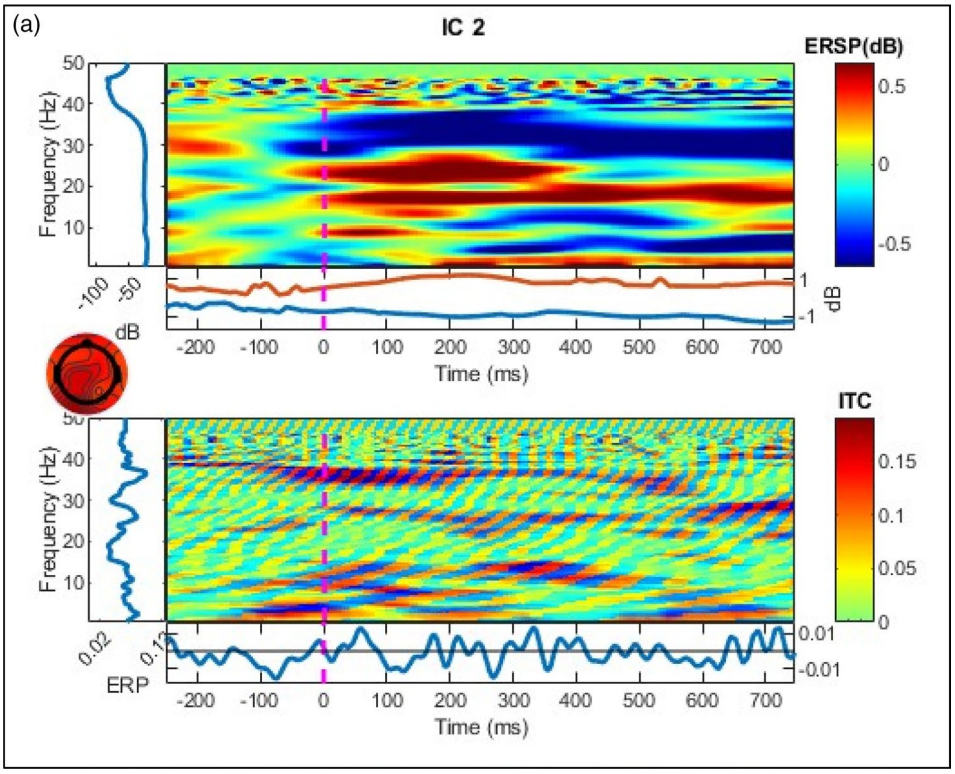

Figure 12a. STFT of Channel 2 by participant 3 when he pronounced Arabic words.

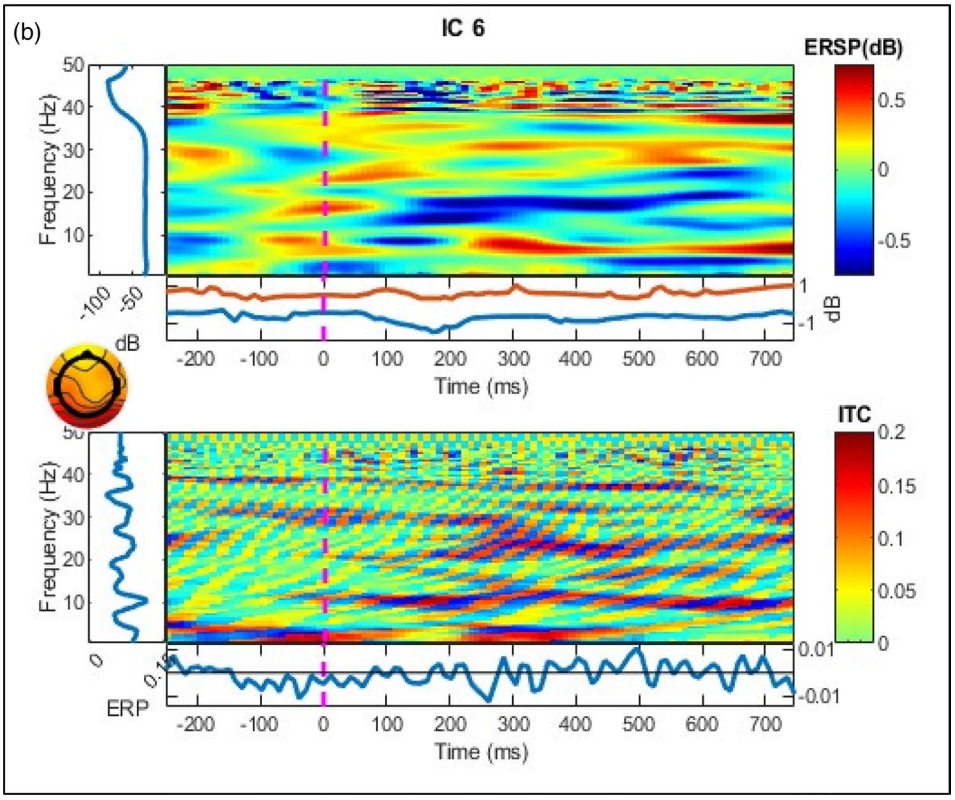

Figure 12b. STFT of Channel 6 by participant 3 when he pronounced Hindi words.

Table 6. Performance accuracy was measured for all models applied during the two recording sessions in Arabic and Hindi for all volunteers.

Model	Multi-features	S1	S2	S3	S4	S5	S6	S7	S8	Accuracy
RF	PSD, STFT, ICA	65.00	66.00	85.00	73.00	68.00	69.00	75.00	72.00	71.63
SVM	–	62.00	64.00	79.05	71.34	63.00	67.03	71.02	69.88	68.42
DT	–	56.00	60.00	79.00	65.00	60.00	63.00	69.00	61.00	64.13
CatBoost	–	63.99	65.89	85.99	73.33	79.51	70.41	79.89	77.52	74.57
XGBoost	–	64.60	65.46	85.38	73.24	68.00	69.96	79.07	71.70	72.18

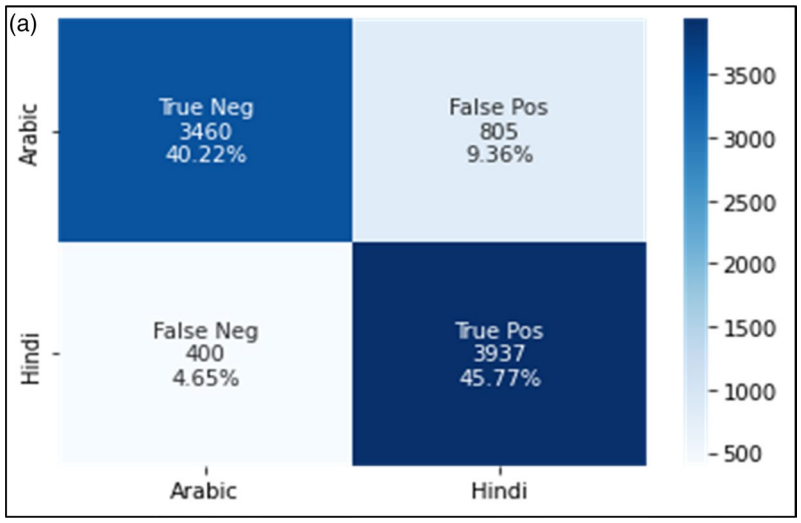

Figure 13a. Confusion matrix for Cat boost.

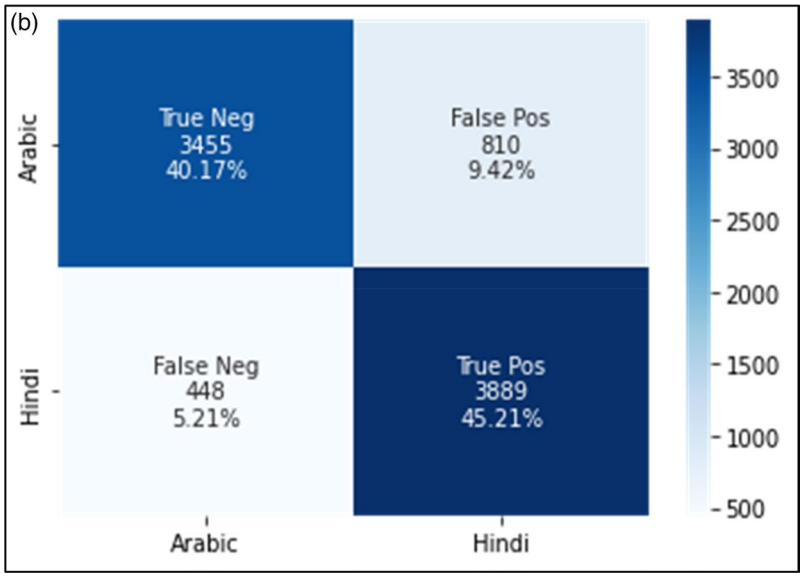

Figure 13b. Confusion matrix for XG boost.

was set to 2. The minimum number of samples required to be at a leaf node (min_samples_leaf) was set to 1. Regarding the XGBoost classifier, we used a learning rate (eta) of 0.1, which controls the step size at each boosting iteration. The maximum depth of each tree (max_depth) was set to 6, limiting the complexity of the model. The subsample ratio (subsample) was set to 1.0, indicating that all training instances were used for fitting each tree. Lastly, for the CatBoost classifier, we set the learning rate (learning rate) to 0.1, governing the step size at each boosting iteration. The maximum depth of each tree (max_depth) was set to 6, controlling the model's complexity. The L2 regularization term (l2_leaf_reg) was set to 3, helping to

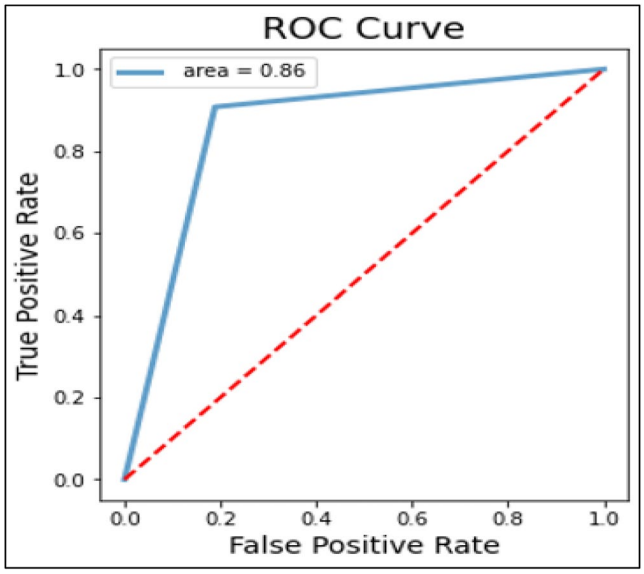

Figure 14. Receiver operator characteristic curve for XG boost.

Table 7. Recall results for each classifier on Arabic and Hindi language classification.

Classifier	Recall (Arabic)	Recall (Hindi)
RF	0.85	0.88
SVM	0.82	0.84
DT	0.79	0.81
XGBoost	0.87	0.89
CatBoost	0.89	0.91

prevent overfitting by adding a penalty term to the objective function. These hyperparameter settings were chosen based on a combination of domain knowledge, empirical results and best practices commonly employed in the field of machine learning. By carefully tuning these hyperparameters, we aimed to optimize the performance of each classifier and ensure a fair comparison of their accuracy in capturing the relationship between brain activity and language learning.

Performance evaluation

To determine the extent of second language acquisition or learning by the human brain of the volunteers in this study, we resorted to determining and choosing the optimal performance of the model, where we applied four practical sequential steps as criteria for evaluating the proposed model. Among these criteria were accuracy in evaluating the performance of the classifier in depth and accuracy in general, as well as specific measures to evaluate the performance of classifiers. Moderate machine learning was proposed. All these criteria played an effective role in proving and evaluating the results achieved by the techniques used in distinguishing and determining the brain's ability to acquire and learn the two languages under study (Hindi and Arabic). Here, we review the general formulas for the criteria used in the evaluation as in the following formulas:

$$\text{Accuracy} = \frac{TP + TN}{FP + FN + TP + TN} \times 100 \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \times 100 \tag{2}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{3}$$

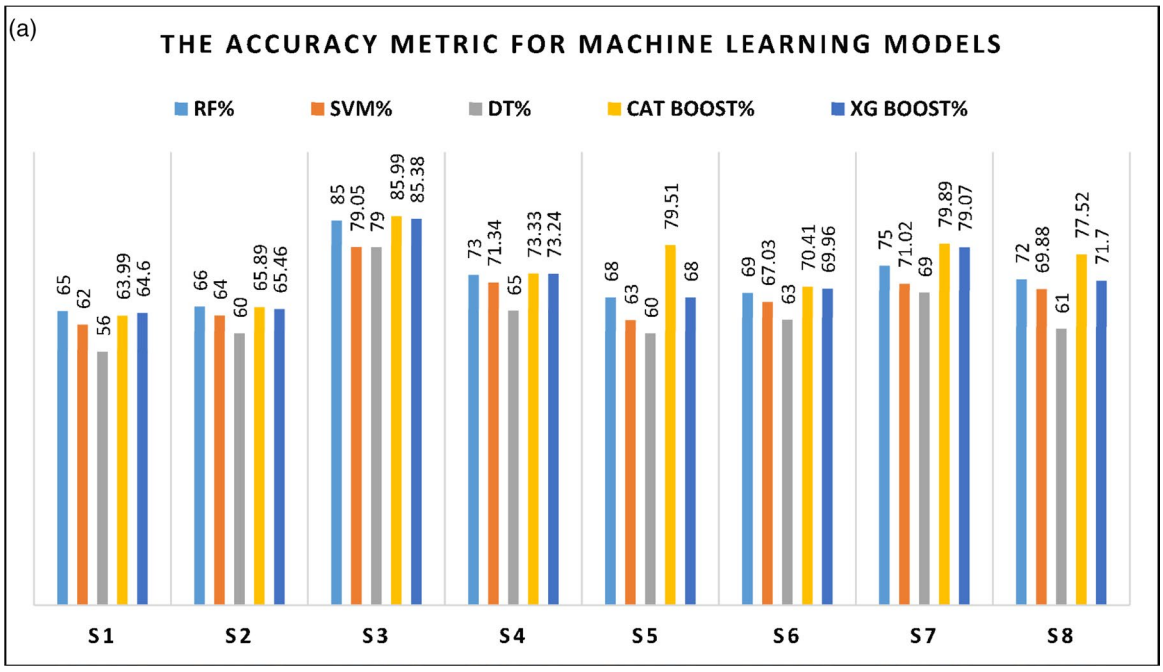

Figure 15a. The performance metric for machine learning models for all participants.

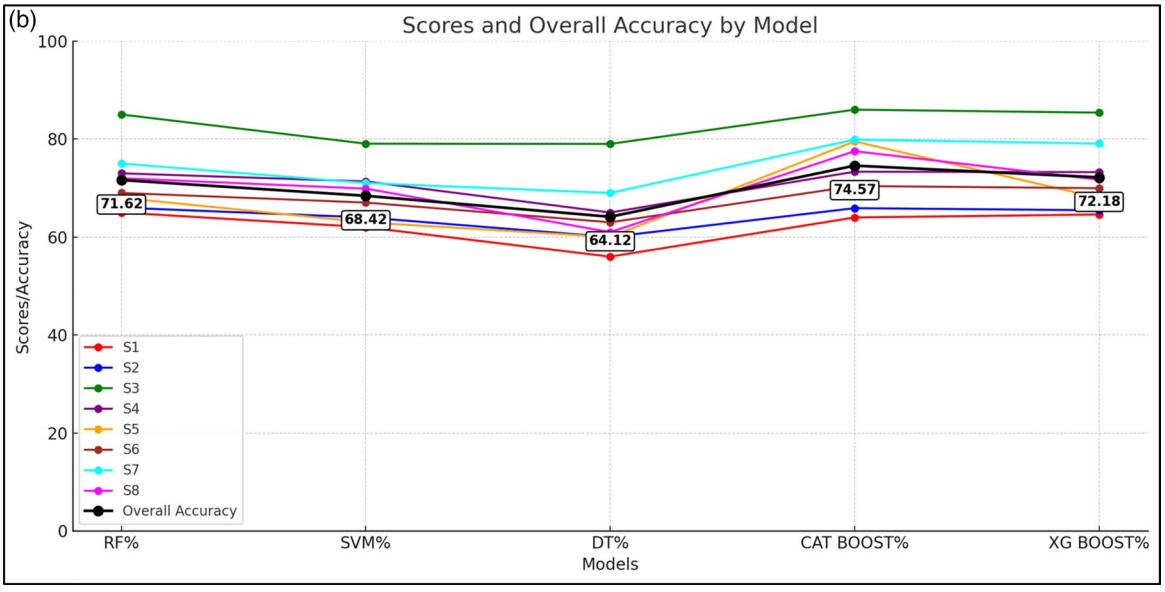

Figure 15b. Preferential comparison of classification models The Cat Boost model is the best with an accuracy of 74.57%, the XG Boost model comes in second with an accuracy of 72.18% and the DT model is the worst in terms of accuracy at 64.12%.

$$\text{Specificity (Spe)} = \frac{TN}{TN + FP} \tag{4}$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \tag{5}$$

Results and discussion

It is widely recognized among researchers and those interested in the field of linguistics, particularly neurolinguistics, that there are differences in the phonetics associated with language learning. The

Table 8. The performance metric for machine learning models for all participants during two recording sessions for both Arabic and Hindi.

		SVM%	RF%	DT%	XG BOOST%	CAT BOOST%
S1	Sensitivity%	75	65	68	67	68
	Specificity%	49	64	44	61	64
	f1-score	57	65	50	63	62
	f1-score	67	65	61	66	66
S2	Sensitivity%	50	63	44	58	59
	Specificity%	77	69	75	72	73
	f1-score	68	67	65	68	69
	f1-score	59	65	52	63	66
S3	Sensitivity%	94	87	83	89	90
	Specificity%	63	81	73	81	85
	f1-score	75	84	77	85	85
	f1-score	82	85	80	86	87
S4	Sensitivity%	83	75	74	78	78
	Specificity%	59	69	56	68	73
	f1-score	67	72	62	72	72
	f1-score	75	74	68	75	75
S5	Sensitivity%	79	70	72	72	73
	Specificity%	47	65	48	62	67
	f1-score	56	67	54	66	65
	f1-score	68	69	65	69	69
S6	Sensitivity%	45	64	55	62	62
	Specificity%	89	75	70	77	70
	f1-score	73	71	65	72	73
	f1-score	58	68	60	68	68
S7	Sensitivity%	57	66	54	66	66
	Specificity%	63	67	58	65	64
	f1-score	61	67	57	65	65
	f1-score	59	67	56	66	66
S8	Sensitivity%	80	72	58	69	78
	Specificity%	53	69	70	65	73
	f1-score	56	70	63	68	72
	f1-score	71	75	60	62	75

pronunciation of the alphabet varies between languages, and information related to spoken or imagined speech cannot be perceived directly based on pronunciation. Processing the signals of spoken or imagined speech is necessary to extract this information, and one method for doing so is to process electroencephalographic signals to recognize the characteristics of the language being learned or acquired. In our experiment, we utilized various classifiers to identify language from the features we extracted from the participants, as shown in Table 1. The performance of the classifiers is presented in Table 6, with CatBoost and XGBoost achieving the highest accuracy. The confusion matrix for the best model is depicted in Figure 13(a,b), and the ROC curves are shown in Figure 14.

The recall results demonstrate the ability of each classifier to correctly identify instances of Arabic and Hindi languages based on the brain activity patterns. CatBoost achieved the highest Recall for both Arabic (0.89) and Hindi (0.91) languages, indicating its superior performance in minimizing false negative predictions. XGBoost also exhibited strong performance, with Recall values of 0.87 for Arabic and 0.89 for Hindi (Figure 15(a,b)). Table 7 presents the recall results for each classifier. Also, Table 8 presents all values of other performance metrics.

Many scientific studies and research in various fields related to neurolinguistics are gaining more interest nowadays. Some of these studies have dealt with the stages of linguistic development (Brown, 1973a, 1973b; Byers-Heinlein et al., 2017; Culbertson & Newport, 2015; Fernald et al., 2006; Gleason, 1958; Rosch, 1978; Roy et al., 2015), while others dealt with language disorders in all their forms, such as aphasia (Breitenstein et al., 2017; Cherney & Patterson, 2017; Garcia & Alves, 2020; Van Ewijk et al., 2017) and studies that dealt with changes associated with language in the brain (Goucha et al., 2017; John W.& Michel, 2019; Sliwinska et al., 2017; Tremblay & Dick, 2016; Xiang et al., 2012). Furthermore, some studies focused on the stages of infant or child brain acquisition of the mother tongue or second language at a young age (Kuhl, 2004; Perani et al., 2011; Werker & Hensch, 2015). There have not been adequate studies or standard data regarding human brain learning and second language acquisition in terms of function and cognition by adults, which is what this study is looking for. Table 9 summarizes the most important studies on the human brain's ability to learn and acquire a second language. It is noticeable

Table 9. Comparison of related works to our work when brain try learning and acquisition second language.

Reference	Scope of work	Recording technique	Participants	Native language	language to be learned	Brain lobe	Feature extraction	The classifier	Average accuracy
Soman et al. (2019)	Investigate the neural mechanisms that underlie a new language learning	EEG	12 subjects/Indian nationals	Hindi	Japanese	Temporal and frontal regions	Spectrogram	(SVM)/ linear kernel/ Gaussian	65.09% and 63.71%
(Berthelsen et al., 2020)	Investigates the neural mechanisms underlying the rapid acquisition of words with grammatical tone by learners from tonal and non-tonal language backgrounds	EEG	23 native speakers of Swedish 23 native speakers of German	Swedish and German	Swedish	left anterior, left central, left posterior	ERP analysis and independent component analysis (ICA)	Statistical analysis/ one-tailed Pearson correlations	–
Hashim et al. (2018)	2 words/yes/&/no/	EEG (Emotive Epoc 14-channels sensor device)	4 volunteers, native speakers of English	English	–	the left side of the brain (motor cortex, Broca's and Wernicke's areas)	the Mel Frequency Cepstral Coefficients (MFCC)	k-NN classification	58%
Prat et al. (2016)	The predicts the rate of second language learning based on individual differences in brain activity during the resting state.	qEEG (wireless EPOC 16-channel headsets (Emotiv, Australia))	19 volunteers, native speakers of English	English	French	Frontal, temporal, parietal and occipital	power spectrum using the Fast Fourier Transform, log-transforming	The Pearson's r correlation	–
Our work	Effects and differences associated with the human brain when trying to learn and acquire a second language	EEG (Virgo EEG device produced by Allengers, which includs 40 active electrodes)	Eight participants, equally holding Indian and Yemeni nationalities	Arabic Hindi	Hindi Arabic	Frontal and Temporal lobes	• Short Time Fourier Transform (STFT) and the Fast Fourier-Transform (FFT). • The logistic Infomax ICA • Power Spectrum.	• Random Forest (RF) • Support Vector Machine (SVM). • Decision Tree • Xgboost • Catboost	74.56%

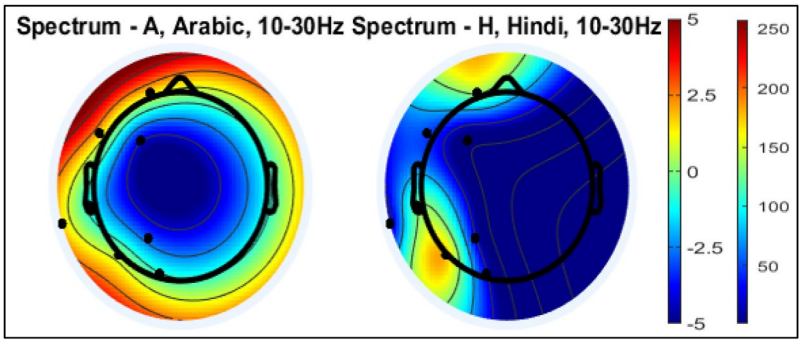

Figure 16. Scalp plots indicating the channels with higher difference for Arabic, Hindi and the difference between the mean inter-trial distance in EEG signals of the two languages.

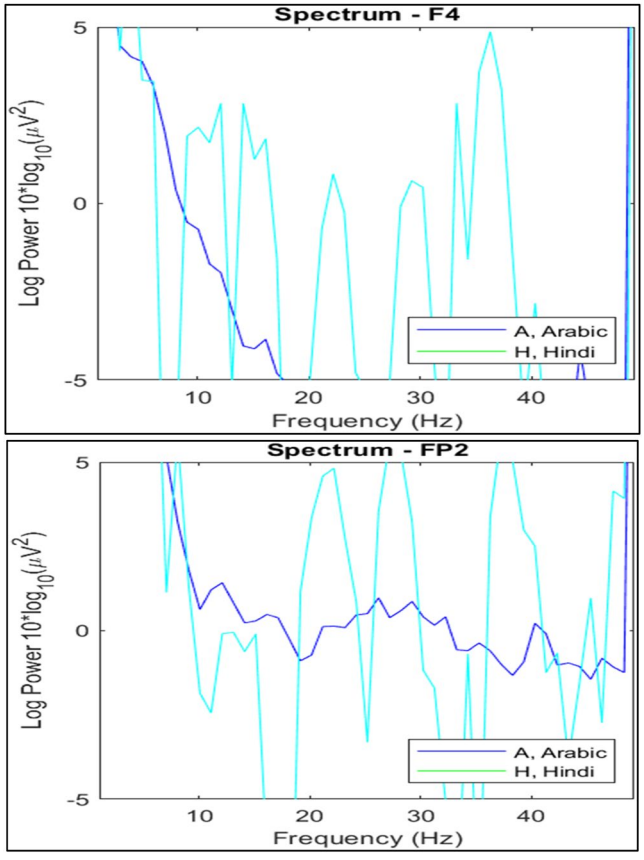

Figure 17. Some of the most active and beat-frequency channels from the language area of the brain. Here the two channels (F4, FP2) are better when pronouncing the Arabic language.

that the studies focusing on learning and acquiring a second language are still in their infancy. This is one of the reasons why we conducted this study, which focused in part on neurolinguistics and brain activity during the second language learning stage. This work shows better performance than previous studies.

Discussion

All participants have never learned or spoken both languages. This is restricted to their native languages, Arabic or Hindi. Figure 16 shows the extent of distances and differences over time for a Yemeni

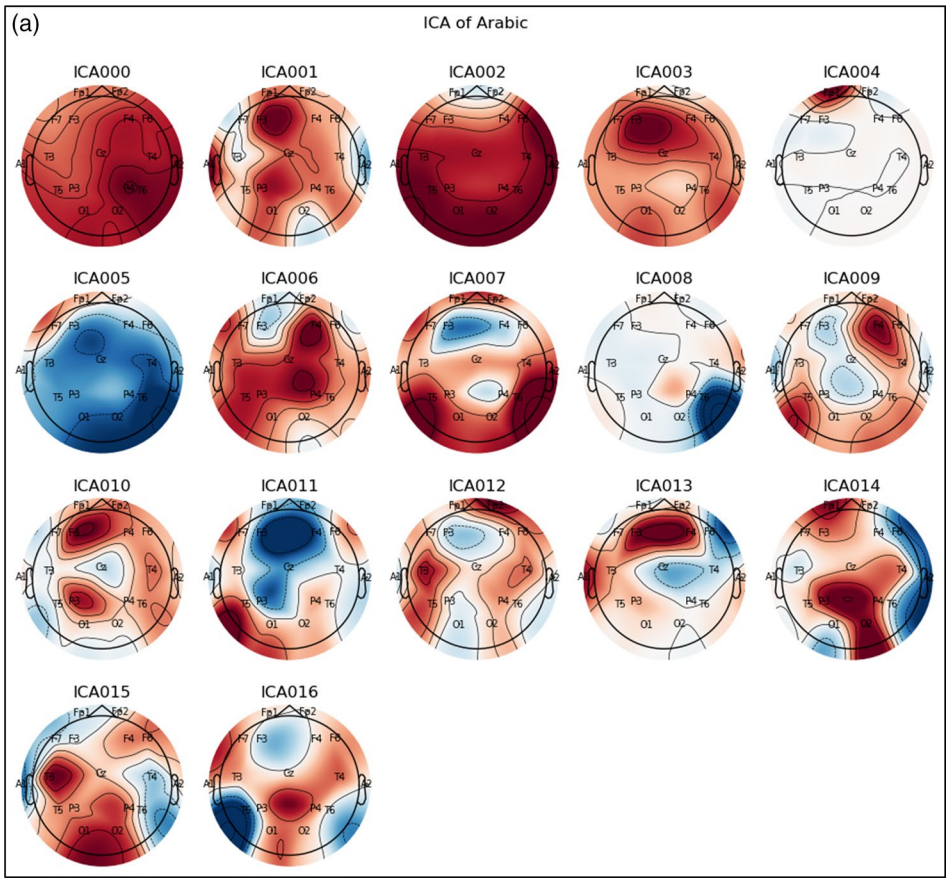

Figure 18a. The neural representations when an Arabic participant pronounce a term in Arabic.

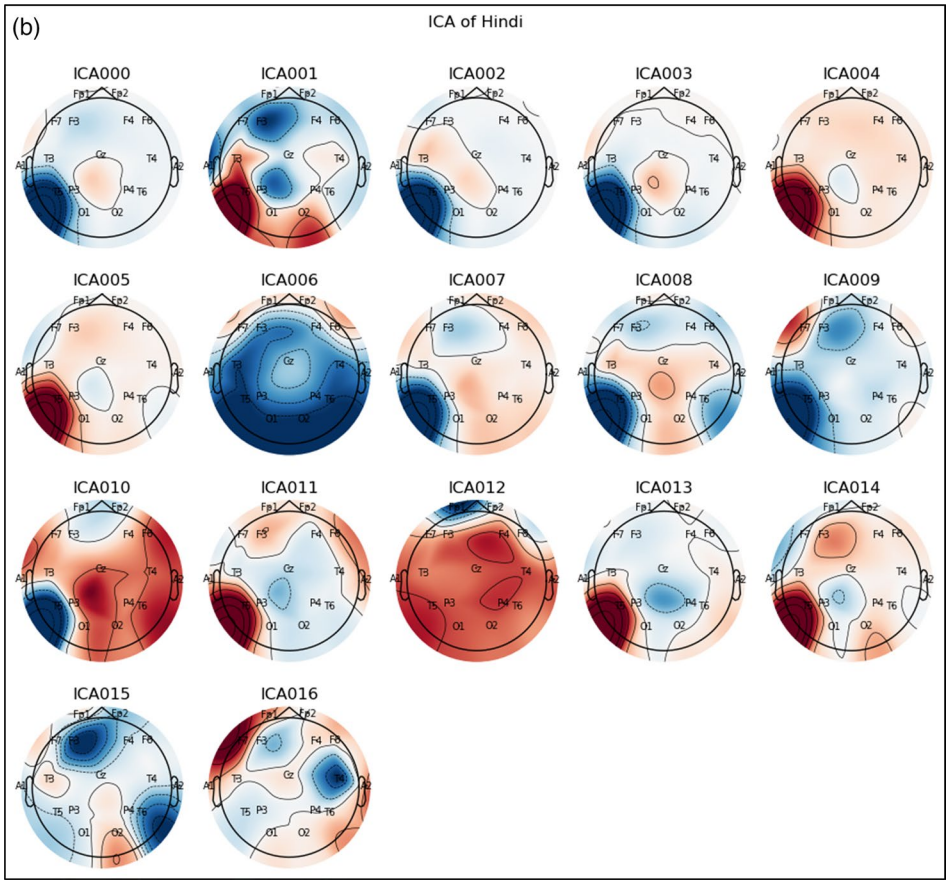

Figure 18b. The neural representations when an Arabic participant pronounce a term with the identical meaning in Hindi.

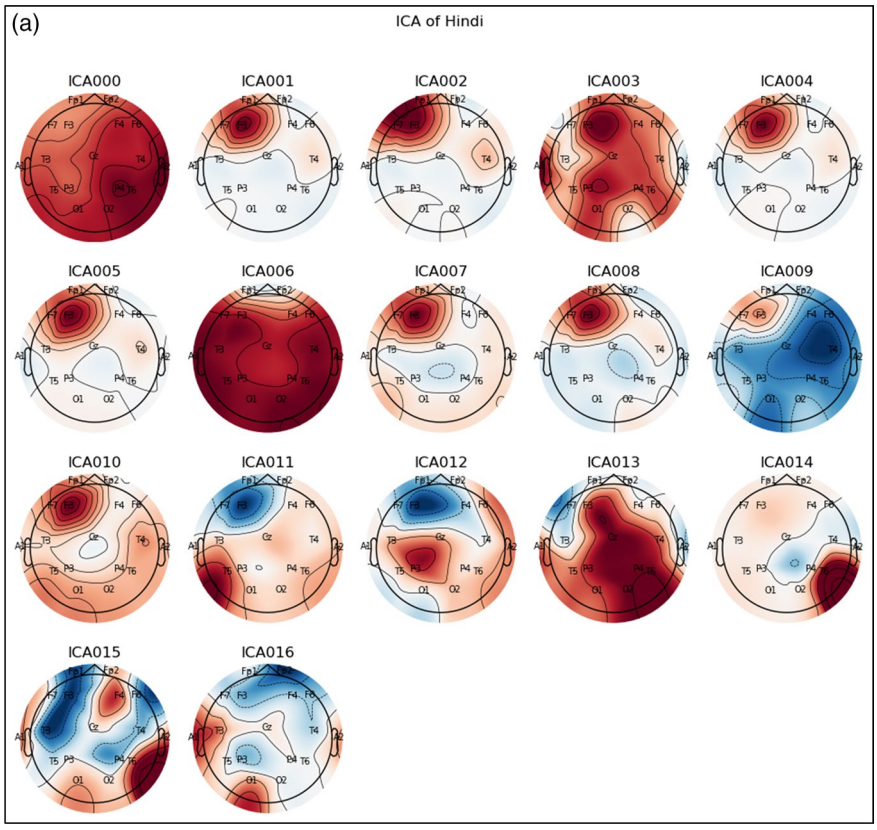

Figure 19a. shows the Indian participant's neural representations of brain activity during the pronunciation of a Hindi word.

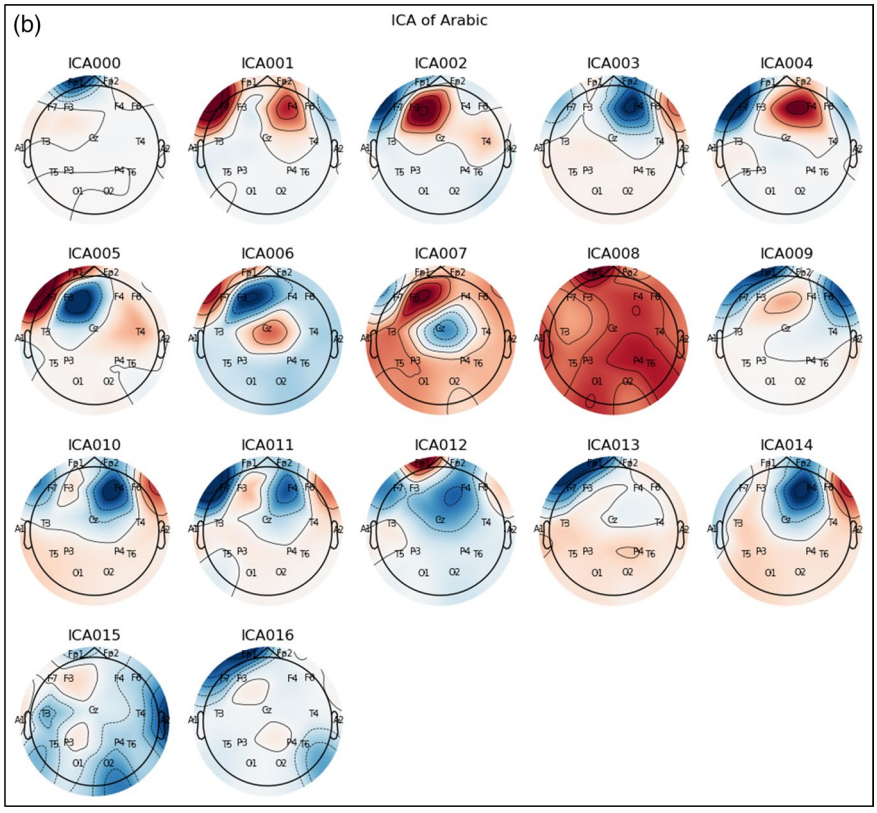

Figure 19b. shows the Indian participant's neural representations of his brain activity during the pronunciation of a word with the same meaning by Arabic.

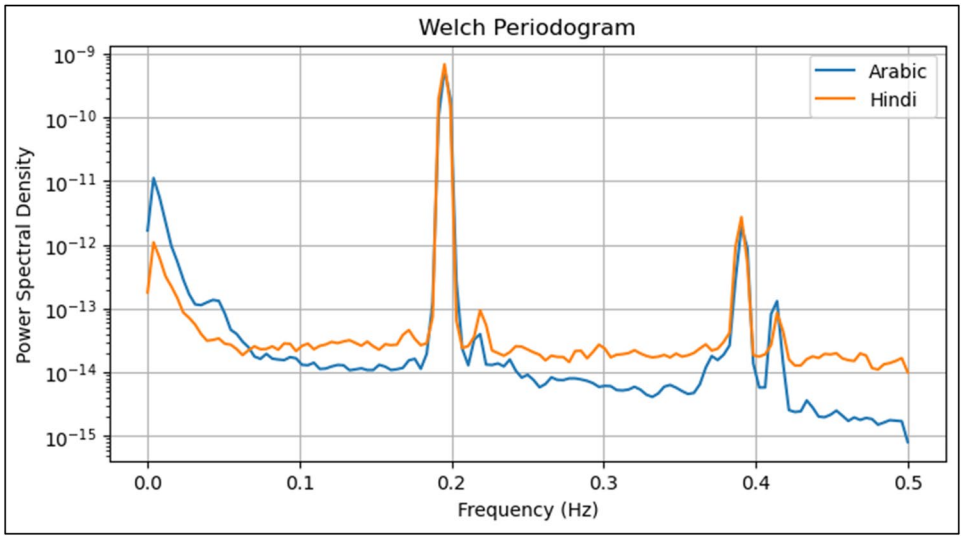

Figure 20. The magnitude of the difference in brain activity when a Yemeni pronounced the same word in Arabic and Hindi.

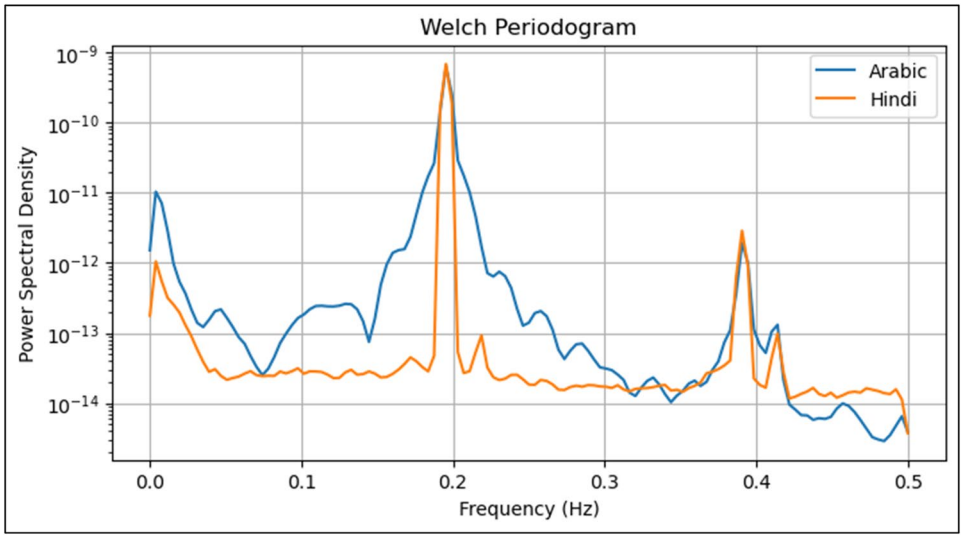

Figure 21. The magnitude of the difference in brain activity when an Indian participant pronounced the same word in Arabic and Hindi.

participant who speaks Arabic and Hindi, where high brain activity is found on the left hemisphere (frontal region) in the pronunciation of the Arabic language due to his prior knowledge of the words under study. In contrast, at the same stimuli, the experiments show a semi-low brain effort in pronouncing Hindi.

The process of learning a new language can be hard for a number of reasons. Here, we'll talk about the most important ones. First, the brain tends to construct a neural representation of unfamiliar stimuli, such as Hindi or Arabic words, which can take time and effort to establish. This can be particularly challenging for non-native speakers who are attempting to learn a second language. Second, the letters used in the two languages, such as Hindi and Arabic, are almost entirely different. This can make it difficult for learners to recognize and remember the new letters and their corresponding sounds, which can further slowdown the process of language acquisition. Third, the length of words in different languages can vary significantly, which can affect the neural representation of words in the brain as detected through EEG. This difference in neural representation can contribute to difficulties when language learning and acquisition.

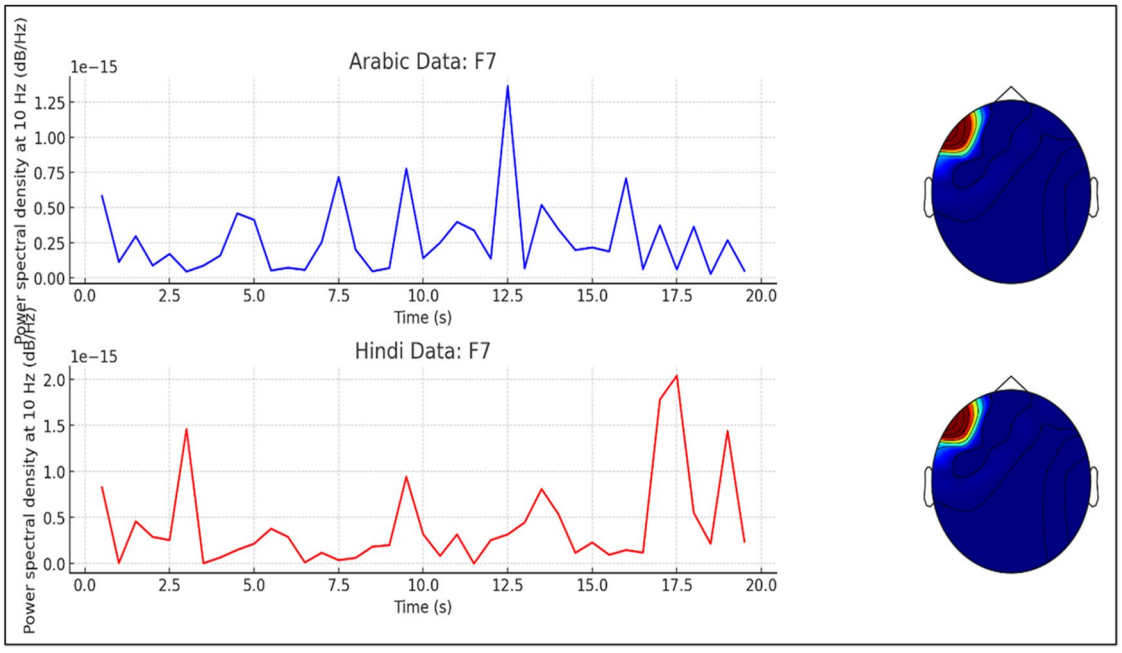

Figure 22. The most distinct difference between the Arabic and Hindi data in power at 10 Hz in the 'F7' channel which is the closest to the Broca's area.

In general, the study finds that through repetition and practice, the human brain creates a distinct neural representation for each word it learns. The neural representation of words in a native language is well-established and familiar to the brain, which makes it easier to process. On the other hand, learning new words in a different language requires building a new and unfamiliar neural representation. This makes the brain less precise and less active in the beginning stages of learning.

The experiment conducted an analysis to discover the channels that are most accurate in capturing the signals responsible for language. The following channels (FP1, F7, F3, T3, A1, P3, T5 and O1) were identified in the brain's frontal region, where they showed the highest difference compared to the other channels, as depicted in Figure 17.

The Broca region and the Wernicke region of the brain, located in both the left frontal and the temporal lobes, are involved in language formation and learning. Figure 18(a,b) shows neural representations of brain activation when the Arabic participant spoke a word meaning the same in Hindi and Arabic (independent components analysis). Arabic is his mother tongue and the language he learned at an early age. Arabic has a high energy level in his pronunciation. On the other hand, as shown in Figure 18(b), brain activity decreases or disappears in several areas of the brain after learning and mastering a word in Hindi.

In Figure 19(a,b), ICA-based neural representations are obtained when the Hindi participant repeated the same phrase to convey the same meaning in both Arabic and Hindi. Since Hindi is his native language and something he was taught at an early age, he displays a high level of vitality when speaking it. Brain activity decreases or is almost non-existent in some regions of the brain when trying to learn and understand the same term in Arabic, as shown in Figure 19(b).

To estimate the PSD of the signal, we used Welch's method after dividing it into overlapping segments. Using this method, we were able to identify and compare with high precision the spectral characteristics of each signal of the Yemeni or Indian participant and learn the difference in the energy distribution of different frequency components between them when acquiring new words. In Figure 20, we notice differences in the power distribution between different frequency components when a Yemeni pronounces the same word in Arabic and Hindi. If it shows a clear difference in the amount of energy produced across the frequency and this is proof of the difference in the nervous system of the same word in the human brain.

Conversely, Figure 21 also shows differences in power distribution between different frequency components when an Indian participant pronounces the same word in both Hindi and Arabic. If the amount of energy produced over the frequency is clearly different and this is proof of the difference in the nervous system of the same word in the human brain.

Figure 22 illuminates a prominent contrast between the Arabic and Hindi datasets regarding power at 10 Hz in the 'F7' channel, nearest to Broca's area. The first row represents the Arabic data. The initial sub-figure exhibits the 'F7' channel's power at 10 Hz over time, with PSD in dB/Hz on the y-axis and time in seconds on the x-axis. The succeeding sub-figure portrays the channel's topographic map, where color denotes average PSD at 10 Hz across the 20-s interval. The second row displays the Hindi data. The first sub-figure illustrates power at 10 Hz over time for the 'F7' channel, with PSD on the y-axis and time in seconds on the x-axis. In sum up here the most distinct difference between the Arabic and Hindi data in power at 10 Hz in the 'F7' channel which is the closest to the Broca's area.

Conclusions

Due to the scarcity of studies on brain activity during learning and acquiring a new language, especially the Arabic language. This study covers various studies on EEG data and explains the different methods used to record EEG signals during language learning and processing. We provide a detailed description of the EEG data we collected and processed in order to investigate the effects of learning a foreign language on the brain and to demonstrate the activity, specifically for Hindi and Arabic language learning. We analyzed the brain signals using the EEGLAB software tool and extracted characteristics associated with learning some words before thoroughly comparing the brain signals when acquiring new words in both languages. We also explored the relationship between the potential data vector resulting from the recorded brain signals during language acquisition and various classifiers through the implementation of multiple machines used to record EEG signals during language learning and processing. We provide a detailed description of the EEG data we collected and processed in order to investigate the effects of learning a foreign language on the brain and to demonstrate the activity, specifically for Hindi and Arabic language learning. We analyzed the brain signals using the EEGLAB software tool and extracted characteristics associated with learning some words before thoroughly comparing the brain signals when acquiring new words in both languages. We also explored the relationship between the potential data vector resulting from the recorded brain signals during language acquisition and various classifiers through the implementation of multiple machine-learning algorithms. The obtained classifier accuracy ranged from 64.12% to 74.56%. The study found that through repetition and practice, the human brain creates a distinct neural representation for each word it learns. The neural representation of words in a native language is well-established and familiar to the brain, which makes it easier to process. On the other hand, learning or acquiring new words in a different language requires building a new and unfamiliar neural representation, leading to less precise and less active brain function at the initial stages of learning, as illustrated in this work. In the future, we hope to identify the brain activity centers involved in learning more than just languages.

Acknowledgments

We extend our sincere thanks and gratitude to a MediCover hospital in Aurangabad, India, for providing us with the facilities and capabilities for recording and collecting data related to the subject of the study by allowing us to use their electroencephalography device without a fee. It also provided us with an integrated environment for registration in accordance with the procedures followed in so field. We also assisted his technicians in the processes of training, preparation and registration, according to the latest procedures and available expertise.

Disclosure statement

There is no conflict of interest regarding the publication of this article.

About the authors

Talal A. Aldhaheri is a lecturer at Faculty of Administrative and Computers Sciences, Albaydha University, Yemen. Currently working as a Ph.D. Research Scholar in the Department of Computer Science and Information Technology at Dr. Babasaheb Ambedkar Marathwada University (BAMU), Chhatrapati Sambhajinagar (formerly Aurangabad),

India. He holds a master's degree in computer information systems (2013) from Middle East University, Amman, Jordan. Aldhaheri has published many papers in national and international venues. His areas of interest are Brain-Computer Interface, Neurolinguistics, Electroencephalography, Deep Learning and Machine Learning.

Dr. Sonali B. Kulkarni is an Associate Professor in the Department of Computer Science and IT at Dr. Babasaheb Ambedkar Marathwada University, Chhatrapati Sambhajinagar (formerly Aurangabad), India. She holds a master's degree in computer science (2002) and a Ph.D. in Computer Science from the same university. With 22 years of teaching and research experience, Dr. Kulkarni's research interests include Remote Sensing & GIS, Brain Computer Interface, Natural Language Processing, and Computational Linguistics. She has published extensively in international journals and conferences and is a life member of several professional organizations including IETE and ISCA. Currently Ph.D. and post-graduation students are working under her guidance.

Pratibha R. Bhise is a Ph.D. Research Scholar in the Department of Computer Science and Information Technology at Dr. Babasaheb Ambedkar Marathwada University (BAMU), Chhatrapati Sambhajinagar (formerly Aurangabad), India. She completed her Post Graduation and M.Phil from the same department. Bhise has published over 25 papers in national and international venues and authored 4 books. Her research interests include Remote Sensing, Brain-Computer Interface, and Machine Learning.

Baraq Ghaleb (PhD, MSc, BSc, FHEA, IEEE) is an Associate Professor at Edinburgh Napier University's Centre for Distributed Computing, Networks, and Security. He leads and teaches several Cyber Security and Systems Engineering modules at both undergraduate and postgraduate levels. He earned his PhD from Napier in June 2019 and has over 8 years of research experience in Cyber Security, IoT, Blockchain, and Machine Learning. He has co-authored 35+ papers and led research projects totaling £435,000, while also mentoring master's and PhD students.

Data availability statement

The data that support the findings of this study are not publicly available due to some restrictions imposed by the supporting party. The data is still incomplete, and we are still working on it. We will share some samples upon your request.

References

Aldhaheri, T. A., Kulkarni, S. B., & Bhise, P. R. (2020, August). *Brain computer interface and neuro linguistics: A short review* [Paper presentation]. 2nd International Conference on Sustainable Communication Networks and Application (ICSCN 2020). Surya Engineering College, Erode, India. Springer. https://doi.org/10.1007/978-981-15-8677-4_54

Alexander, M. P., Naeser, M. A., & Palumbo, C. (1990). Broca's area aphasias: Aphasia after lesions including the frontal operculum. *Neurology*, *40*(2), 353–362. https://doi.org/10.1212/WNL.40.2.353

Alexandre, G., Martin, L., Eric, L., Denis, A. E., Daniel, S., Christian, B., … Matti, H. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, *7*, 267. https://doi.org/10.3389/fnins.2013.00267

Allengers Global. (2022). *EEG virgo*. https://www.allengersglobal.com/eeg-virgo

Batterink, L., & Neville, H. J. (2013). ERPs recorded during early second language exposure predict syntactic learning. *Journal of Cognitive Neuroscience*, *25*(6), 936–951. https://doi.org/10.1162/jocn_a_00354

Berthelsen, S. G., Horne, M., Shtyrov, Y., & Roll, M. (2020). Different neural mechanisms for rapid acquisition of words with grammatical tone in learners from tonal and non-tonal backgrounds: ERP evidence. *Brain Research*, *1729*, 146614. https://doi.org/10.1016/j.brainres.2019.146614

Bhise, P. R., Kulkarni, S. B., & Aldhaheri, T. A. (2020). Brain computer interface-based EEG for emotion recognition system: A systematic review. *Proceedings of the Second International Conference on Innovative Mechanisms for Industry Applications (ICIMIA 2020)* (pp. 327–334). https://doi.org/10.1109/icimia48430.2020.9074921

Breitenstein, C., Grewe, T., Flöel, A., Ziegler, W., Springer, L., Martus, P., Huber, W., Willmes, K., Ringelstein, E. B., Haeusler, K. G., Abel, S., Glindemann, R., Domahs, F., Regenbrecht, F., Schlenck, K.-J., Thomas, M., Obrig, H., de Langen, E., Rocker, R., … Bamborschke, S. (2017). Intensive speech and language therapy in patients with chronic aphasia after stroke: A randomised, open-label, blinded-endpoint, controlled trial in a health-care setting. *The Lancet*, *389*(10078), 1528–1538. https://doi.org/10.1016/S0140-6736(17)30067-3

Brown, R. (1973a). A first language: The early stages. In L. Bloom (Ed.), *Studies in language and language behavior* (Vol. 3, pp. 257–285). Wiley.

Brown, R. (1973b). A first language: The early stages. In P. Fletcher & M. Garman (Eds.), *Language development: Form and function in emerging grammars* (pp. 221–263). Harvard University Press.

Bugli, C., & Lambert, P. (2006). Comparison between principal component analysis and independent component analysis in electroencephalograms modeling. *Biomedical Journal*, *49,* 312–327. https://doi.org/10.1002/bimj.200510285

Byers-Heinlein, K., D. A. Behrend, L. M. Said, H. Girgis, and D. Poulin-Dubois. (2017). Monolingual and bilingual children's social preferences for monolingual and bilingual speakers. *Developmental Science,* 20(4), e12392. https://doi.org/10.1111/desc.12392.

Cherney, L. R., & Patterson, J. P. (2017). Evidence-based practice guidelines for individuals with moderate-to-severe primary progressive aphasia. *American Journal of Speech-Language Pathology*, *26*(3), 913–927. https://doi.org/10.1044/2017_AJSLP-16-0167

Culbertson, J., & Newport, E. L. (2015). Harmonic biases in child learners: In support of language universals. *Cognition*, *139*, 71–82. https://doi.org/10.1016/j.cognition.2015.02.007

Dave, S., Brothers, T., & Swaab, T. (2018). 1/f neural noise and electrophysiological indices of contextual prediction in aging. Brain Research, 1691, 34–43. https://doi.org/10.1016/j.brainres.2018.04.007

Delorme, A., & Makeig, S. (2004). EEGLAB: An open-source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Fernald, A., Perfors, A., & Marchman, V. A. (2006). Picking up speed in understanding: Speech processing efficiency and vocabulary growth across the 2nd year. *Developmental Psychology*, *42*(1), 98–116. https://doi.org/10.1037/0012-1649.42.1.98

Garcia, A. C., & Alves, G. S. (2020). Language therapy for aphasia: A systematic review and meta-analysis of randomized controlled trials. *Aphasiology*, *34*(10), 1183–1218. https://doi.org/10.1080/02687038.2019.1697949

Gleason, J. B. (1958). The child's learning of English morphology. *Word*, *14*(2–3), 150–177. https://doi.org/10.1080/00437956.1958.11659661

González-Castañeda, E., García, A., García, C., & Villaseñor-Pineda, L. (2017). Sonification and textification: Proposing methods for classifying unspoken words from EEG signals. *Biomedical Signal Processing and Control*, *37*, 82–91. https://doi.org/10.1016/j.bspc.2016.10.012

Goucha, T., Emiliano Z., and A. D. Friederici. (2017). A revival of Homo Loquens as a builder of labeled structures: Neurocognitive considerations. *Neuroscience & Biobehavioral Reviews*, 81, 213–224. https://doi.org/10.1016/j.neubiorev.2017.01.036.

Grundy, J., Anderson, J., & Bialystok, E. (2017). Bilinguals have more complex EEG brain signals in occipital regions than monolinguals. *NeuroImage*, *159*, 280–288. https://doi.org/10.1016/j.neuroimage.2017.07.063

Gurumoorthy, S., Muppalaneni, N. B., Sekhar, C., & Sandhya Kumari, G. (2020). Epilepsy analysis using open source EDF tools for information science and data analytics. *International Journal of Communication Systems*, *33*(13), e4095. https://doi.org/10.1002/dac.4095

Hashim, N., Ali, A., & Mohd-Isa, W. N. (2018). Word-based classification of imagined speech using EEG. In R. Alfred, Y. Lim, A. Ibrahim, & P. Anthony (Eds.), *Computational science and technology* (Vol. 488, pp. 195–204). Springer.

Haufe, S., Kim, J. W., Kim, I. H., Sonnleitner, A., Schrauf, M., Curio, G., & Blankertz, B. (2014). Electrophysiology-based detection of emergency braking intention in real-world driving. *Journal of Neural Engineering*, *11*(5), 056011. https://doi.org/10.1088/1741-2560/11/5/056011

Hebb, A. O., & Ojemann, G. A. (2013). The thalamus and language revisited. *Brain and Language*, *126*(1), 99–108. https://doi.org/10.1016/j.bandl.2012.06.010

Jamil, N., Belkacem, A. N., Ouhbi, S., & Lakas, A. (2021). Noninvasive electroencephalography equipment for assistive, adaptive, and rehabilitative brain-computer interfaces: A systematic literature review. *Sensors (Basel, Switzerland)*, *21*(14), 4754. https://doi.org/10.3390/s21144754

Kang, J., Ojha, A., & Lee, M. (2015). Development of intelligent learning tool for improving foreign language skills based on EEG and eye tracker. *Proceedings of the 3rd International Conference on Human-Agent Interaction (HAI '15)* (pp. 53–56). ACM. https://doi.org/10.1145/2814940.2814951

Kawala-Sterniuk, A., Podpora, M., Pelc, M., Blaszczyszyn, M., Gorzelanczyk, E. J., Martinek, R., & Ozana, S. (2020). Comparison of smoothing filters in analysis of EEG data for the medical diagnostics purposes. *Sensors*, *20*(3), 807. https://doi.org/10.3390/s20030807

Kemp, B., van Beelen, T., Stijl, M., van Someren, P., Roessen, M., & van Dijk, J. G. (2010). A DC attenuator allows common EEG equipment to record fullband EEG and fits fullband EEG into standard European Data Format. *Clinical Neurophysiology*, *121*(12), 1992–1997. https://doi.org/10.1016/j.clinph.2010.05.006

Kuhl, P. K. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience*, 5(11), 831–843. https://doi.org/10.1038/nrn1533

Kumar, P., Saini, R., Roy, P., Sahu, P., & Dogra, D. (2018). Envisioned speech recognition using EEG sensors. *Personal and Ubiquitous Computing*, *22*(1), 185–199. https://doi.org/10.1007/s00779-017-1083-4

Liao, L. D., Lin, C. T., McDowell, K., Wickenden, A. E., Gramann, K., Jung, T. P., et al. (2012, March 12). Biosensor technologies for augmented brain–computer interfaces in the next decades. In *Proceedings of the IEEE, 100 (Special Centennial Issue)*, 1553–1566. IEEE. https://doi.org/10.1109/JPROC.2012.2184829

Liu, X., Tan, P., Liu, L., & Simske, S. (2017). *Automated classification of EEG signals for predicting students' cognitive state during learning* [Paper presentation]. Proceedings of the International Conference on Web Intelligence (WI '17) (pp. 442-450). Leipzig, Germany: ACM. https://doi.org/10.1145/3106426.3106453

Lotte, F., Bougrain, L., Cichocki, A., Clerc, M., Congedo, M., Rakotomamonjy, A., & Yger, F. (2018). A review of classification algorithms for EEG-based brain-computer interfaces: A 10-year update. *Journal of Neural Engineering*, *15*(3), 031005. https://doi.org/10.1088/1741-2552/aab2f2

Millán, J. D. R., Rupp, R., Müller-Putz, G. R., Murray-Smith, R., Giugliemma, C., Tangermann, M., Vidaurre, C., Cincotti, F., Kübler, A., Leeb, R., Neuper, C., Müller, K. R., & Mattia, D. (2010). Combining brain-computer interfaces and assistive technologies: State-of-the-art and challenges. *Frontiers in Neuroscience*, *4*, 1–15. https://doi.org/10.3389/fnins.2010.00161

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., Tu-Chan, A., Ganguly, K., & Chang, E. F. (2021). Neuroprosthesis for decoding speech in a paralyzed person with anarthria. *New England Journal of Medicine*, *385*(3), 217–227. https://doi.org/10.1056/NEJMoa2027540

Nguyen, C., Karavas, G., & Artemiadis, P. (2018). Inferring imagined speech using EEG signals: A new approach using Riemannian manifold features. *Journal of Neural Engineering*, *15*(1), 016002. https://doi.org/10.1088/1741-2552/aa8235

Ofner, P., Schwarz, A., Pereira, J., Wyss, D., Wildburger, R., & Müller-Putz, G. R. (2019). Attempted arm and hand movements can be decoded from low-frequency EEG from persons with spinal cord injury. *Scientific Reports*, *9*(1), 7134. https://doi.org/10.1038/s41598-019-43594-9

Park, H. J., Furmaga, H., Cooperrider, J., Gale, J. T., Baker, K. B., & Machado, A. G. (2015). Modulation of cortical motor evoked potential after stroke during electrical stimulation of the lateral cerebellar nucleus. *Brain Stimulation*, *8*(6), 1043–1048. https://doi.org/10.1016/j.brs.2015.06.020

Parvizi, J., & Kastner, S. (2018). Promises and limitations of human intracranial electroencephalography. *Nature Neuroscience*, *21*(4), 474–483. https://doi.org/10.1038/s41593-018-0108-2

Perani, D., Saccuman, M. C., Scifo, P., Anwander, A., Spada, D., Baldoli, C., Poloniato, A., Lohmann, G., & Friederici, A. D. (2011). Neural language networks at birth. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(38), 16056–16061. https://doi.org/10.1073/pnas.1102991108

Prat, C., Yamasaki, B., Kluender, R., & Stocco, A. (2016). Resting-state qEEG predicts rate of second language learning in adults. *Brain and Language*, *157–158*, 44–50. https://doi.org/10.1016/j.bandl.2016.04.007

Rabbani, Q., Milsap, G., & Crone, N. E. (2019). The potential for a speech brain-computer interface using chronic electrocorticography. *Neurotherapeutics: The Journal of the American Society for Experimental Neuro Therapeutics*, *16*(1), 144–165. https://doi.org/10.1007/s13311-018-00692-2

Rahma, R., & Nurhadi, J. (2017). *Can power spectral density (PSD) be used to measure reading concentration* [Paper presentation]. Proceedings of the Tenth Conference on Applied Linguistics and the Second English Language Teaching and Technology Conference in Collaboration with the First International Conference on Language, Literature, Culture, and Education (CONAPLIN and ICOLLITE), Bandung, Indonesia (Vol. 1, pp. 450–453). https://doi.org/10.5220/0007169004500453

Rashid, M., Sulaiman, N., Abdul Majeed, A. P. P., Musa, R. M., Ab. Nasir, A. F., Bari, B. S., & Khatun, S. (2020). Current status, challenges, and possible solutions of EEG-based brain-computer interface: A comprehensive review. *Frontiers in Neurorobotics*, *14*, 25. https://doi.org/10.3389/fnbot.2020.00025

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 27–48). Lawrence Erlbaum Associates.

Rosenfeld, J. V., & Wong, Y. T. (2017). Neurobionics and the brain-computer interface: Current applications and future horizons. *The Medical Journal of Australia*, *206*(8), 363–368. https://doi.org/10.5694/mja16.01011

Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., & Roy, D. (2015). Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(41), 12663–12668. https://doi.org/10.1073/pnas.1419773112

Ryan, D. B., Townsend, G., Gates, N. A., Colwell, K., & Sellers, E. W. (2017). Evaluating brain computer interface performance using color in the P300 checkerboard speller. *Clinical Neurophysiology*, *128*(10), 2050–2057. https://doi.org/10.1016/j.clinph.2017.07.397

Schultz, T., Wand, M., Hueber, T., Krusienski, D. J., Herff, C., & Brumberg, J. S. (2017). Biosignal-based spoken communication: A survey. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(12), 2257–2271. https://doi.org/10.1109/TASLP.2017.2752365

Sengottuvel, K., & Gupta, A. K. (2020). Effects of multilingualism on the brain: A review. *Journal of Indian Speech Language & Hearing Association*, *34*(2), 127–133.

Sereshkeh, A., Trott, R., Bricout, A., & Chau, T. (2017a). EEG classification of covert speech using regularized neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(12), 2292–2300. https://doi.org/10.1109/TASLP.2017.2758164

Sereshkeh, A., Trott, R., Bricout, A., & Chau, T. (2017b). Online EEG classification of covert speech for brain-computer interfacing. *International Journal of Neural Systems*, *27*(08), 1750033. https://doi.org/10.1142/S0129065717500332

Sliwinska, M. W., Vitello, S., & Joseph T. D. (2014). Transcranial magnetic stimulation for investigating causal brain-behavioral relationships and their time course. *JoVE (Journal of Visualized Experiments),* 89, e51735. https://doi.org/10.3791/51735.

Soman, A., Madhavan, C., Sarkar, K., & Ganapathy, S. (2019). An EEG study on the brain representations in language learning. *Biomedical Physics & Engineering Express*, *5*(2), 025041. https://doi.org/10.1088/2057-1976/ab0243

Subasi, A., & Gursoy, M. I. (2010). EEG signal classification using PCA, ICA, LDA and support vector machines. *Expert Systems with Applications*, *37*(12), 8659–8666. https://doi.org/10.1016/j.eswa.2010.06.065

Tremblay, P., & Dick, A. S. (2016). Broca and Wernicke are dead or moving past the classic model of language neurobiology. *Brain and Language*, *162*, 60–71. https://doi.org/10.1016/j.bandl.2016.08.004

Tzimourta, K. D., Christou, V., Tzallas, A. T., Giannakeas, N., Astrakas, L. G., Angelidis, P., Tsalikakis, D., & Tsipouras, M. G. (2021). Machine learning algorithms and statistical approaches for Alzheimer's disease analysis based on resting-state EEG recordings: A systematic review. *International Journal of Neural Systems*, *31*(5), 2130002. https://doi.org/10.1142/S0129065721300023

Vaid, S., Singh, P., & Kaur, C. (2015). EEG signal analysis for BCI interface: A review [Paper presentation]. In *Proceedings of the 2015 Fifth International Conference on Advanced Computing & Communication Technologies (ACCT)*. Haryana, India. IEEE. https://doi.org/10.1109/acct.2015.72

Van Ewijk, L., Siepel, F. J., Visch-Brink, E. G., & Kolk, H. H. (2017). How to differentiate neurogenic and psychogenic stuttering: A multimodal study. *Journal of Fluency Disorders*, *52*, 12–28. https://doi.org/10.1016/j.jfludis.2016.08.003

Werker, J. F., & Hensch, T. K. (2015). Critical periods in speech perception: New directions. *Annual Review of Psychology*, *66*(1), 173–196. https://doi.org/10.1146/annurev-psych-010814-015104

Xiang, H., D. Dediu, L. Roberts, E. V. Oort, D. G. Norris, and P. Hagoort. (2012). The structural connectivity underpinning language aptitude, working memory, and IQ in the perisylvian language network. *Language Learning* 62, 110–130. https://doi.org/10.1111/j.1467-9922.2012.00708.x.