UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft
Faculty for Computer Science, Electrical Engineering and Mathematics
Department of Computer Science
Research Group IT Security

Master's Thesis

Submitted to the IT Security Research Group
in Partial Fulfilment of the Requirements for the Degree of

Master of Science

Brainwave-based User Authentication Models

by
Avinash Kumar Chaurasia

Thesis Supervisor:
Prof. Dr. Patricia Arias Cabarcos

Thesis Examiners:
Prof. Dr. Patricia Arias Cabarcos
Jun.-Prof. Dr. Sebastian Peitz

Paderborn, September 04, 2023

Erklärung

Ich versichere, dass ich die Arbeit ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Quellen angefertigt habe und dass die Arbeit in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen hat und von dieser als Teil einer Prüfungsleistung angenommen worden ist. Alle Ausführungen, die wörtlich oder sinngemäß übernommen worden sind, sind als solche gekennzeichnet.

Paderborn, 04.09.2023

Ort, Datum

Unterschrift

Abstract. Brainwaves present a compelling avenue for secure person authentication because they are inherently unobservable externally and capable of facilitating liveness detection. Harnessing brainwave's unique and individualistic attributes, they have found extensive utility in various authentication applications. Nonetheless, the domain of brainwave authentication research has witnessed an upsurge in diverse experimental setups and the meticulous fine-tuning of parameters to optimize authentication methodologies. The substantial diversity in their methods poses a significant obstacle in assessing and measuring authentic research advancements. To address this multifaceted issue, this thesis introduces a versatile and robust benchmarking framework tailored explicitly for brainwave authentication systems. This framework draws upon the resources of four publicly accessible medical-grade brainwave datasets. It is worth mentioning that our study encompasses a substantial sample size consisting of 195 participants. The number of participants in our study is noteworthy, particularly when compared to the customary approach in brainwave authentication research, which typically involves a participant pool about one-fifth the size of our study. Our extensive assessment encompassed a variety of state-of-the-art authentication algorithms, including Logistic Regression, Linear Discriminant Analysis, Support Vector Machine, Naive Bayes, K-Nearest Neighbours, Random Forest, and advanced deep learning methods like Siamese Neural Networks. Our evaluation approach incorporated both within-session (single-session) and cross-session (multi-session) analysis, covering threat cases like close-set (seen attacker) and open-set (unseen attacker) scenarios to ensure the tool's versatility in different contexts. In within-session evaluation, our framework showcased outstanding performance for several classifiers, particularly Siamese Networks, which achieved an Equal Error Rate of 1.60% in the unseen attacker scenario. Additionally, our benchmarking framework's adaptability is a notable asset, allowing researchers to tailor pre-processing, feature extraction, and authentication parameters to suit their specific requirements.

Contents

1 Introduction

1.1 Motivation

The confidentiality of information is one of the most crucial components of data security, and it is vital that only authorized individuals can access sensitive information [1]. Authentication procedures play an essential role in maintaining information confidentiality by verifying the user's identity requesting access to secure data [2]. User authentication comprises two key stages: enrollment and verification [3]. In the initial phase, the user registers or enrolls in the system, leading to the capture and storage of the user's data in the database. Subsequently, In the second phase, the authenticity of the user's data is checked by matching the user's presented data with his existing data in the database. The system either gives or refuses access to the user based on the degree of resemblance [4]. At present, there are three approaches to user authentication: authentication based on knowledge (e.g., password), authentication based on possession (e.g., token or ID card), and authentication based on biometrics (e.g., fingerprint, iris, facial recognition, or other biometric data) [5].

Knowledge-based authentication is the simplest method, which involves verifying the identity of a user by requesting a password or PIN (personal identity number) known only to the user. Knowledge-based authentication methods, such as the use of passwords, offer several benefits, including user-friendliness and ease of maintenance. Further, passwords can be revoked with ease when compromised. However, passwords suffer from many vulnerabilities, such as complex passwords that are often hard to remember. As a result, users reuse short and easy-to-remember passwords across multiple websites, exposing the passwords to attackers for breach [6]. A study by Das *et al.* [7] analyzed a large dataset comprising several hundred thousand stolen passwords from eleven distinct websites. In addition, they conducted a poll on password reuse, which allowed them to estimate that a substantial proportion of users, ranging from 44% to 51%, employ the same password across several websites. In their study, the authors also developed a cross-site guessing platform to guess approximately 10% of the nonidentical password pairs in fewer than ten attempts and approximately 30% in fewer than 100 attempts. Furthermore, passwords can also be stolen through casual eavesdropping (shoulder surfing) [6], or can be guessed using sophisticated hacking algorithms such as dictionary search attacks in which words and word combinations are hashed and then checked for matches against hashed passwords [8]. The inherent security vulnerabilities associated with password usage compromise their effectiveness in establishing robust authentication systems.

Possession-based authentication requires the user to possess something such as a token ID

to verify their identity. Like passwords, tokens are easy to maintain and can be revoked easily if lost or compromised. Another benefit is its capability to detect compromises, as its absence can be observed, which is not the case with the loss of a password [8]. While tokens may provide certain benefits compared to passwords, they are not an infallible authentication method, given various security vulnerabilities. For instance, tokens are susceptible to theft [9] and duplication, meaning that someone might create a counterfeit device [8].

The third form of authentication is biometric-based, relying on distinctive user biometrics for authentication [8]. Biometrics can be categorized into two groups: physiological and behavioral. Physiological biometrics pertain to the physical attributes of the human body, which inherently differ among individuals [10]. Examples of physiological Biometrics include fingerprints, facial recognition, hand geometry, and iris recognition [11]. On the other hand, behavioral biometrics refers to an individual's behavior, such as gait, voice, or signature. Unlike passwords and tokens, biometrics do not need to be memorized or physically carried everywhere. They are also unique and cannot be imitated easily. Biometrics provide more effective authentication mechanisms than passwords and tokens, but biometric-based authentication still needs to be indomitable. Biometric data, like voice, facial features, iris, retina, and fingerprints, can be captured or photographed [9]. Further, unlike passwords and tokens, biometrics are not easy to replace if they are lost or compromised [8], and the person with specific physical disabilities (e.g., blindness or quadriplegia) cannot use biometric systems, requiring eyes, fingerprints, or gait to authenticate. Each of the authentications, as mentioned earlier methods, has its own merits and weaknesses which need to be addressed. An alternative authentication method is required to overcome existing authentication methods' weaknesses and provide a robust mechanism to verify the identity of users.

There has been a rise in interest in using brain activity for next-generation biometric systems to fill in the gaps left by current biometric techniques or to complement them [12]. The technological advances in the last few years have made it possible to obtain brain signals using Electroencephalography (EEG) and utilize the unique characteristics of EEG signals to verify a person's identity [13]. The following are some of the advantages of brainwaves that give them a giant leap over other biometric traits for authentication:

1. Unlike observable biometric traits like face or gait that can be exploited for identification without consent [14], brain activities remain hidden from external observation and are therefore impervious to any form of surveillance [12].

2. It is also impractical to steal brainwaves due to the susceptibility of a person's brain activity to their stress and mood. Attackers cannot compel the victim to replicate their mental passphrase [14]. For example, suppose a person is frightened or stressed out. In that case, the brainwaves recorded during the authentication phase will differ significantly from the brain data collected during the person's enrollment into the system. Thus, the system would refuse to grant access if an attacker forces the person to provide his brainwaves.

3. Brainwaves can only be produced by living brain tissue [15]. Therefore, brainwaves are a promising candidate for being used as a biometric trait since they can readily handle the main problem of liveness detection in other biometrics [12].

4. Brainwaves are organically a part of the human body, so even those people who are physically disabled can utilize them, unlike with fingerprints or other types of technology, which may not be possible [9].

1.2 Problem Description

As elaborated in the previous section 1.1, authentication systems based on brainwaves offer a compelling alternative to conventional authentication systems. However, simply building a brainwave authentication system under the presumption that the chosen algorithm or evaluation metrics confer optimal security is insufficient. Even the most meticulously designed brainwave authentication system may have hidden flaws that are not immediately discernible. As a result, it is crucial to identify and address the specific research gaps to ensure that the system being developed provides robust and reliable security. The following research gaps in the field of EEG-based person authentication are the focus of the study described in this thesis, which will be designed to improve the system's high levels of security, performance, and stability.

1. **Comparative Performance Evaluation and Reporting**

   An EEG-based person authentication system's effectiveness relies on data preprocessing, feature extraction, and modeling techniques. Numerous machine learning algorithms such as Linear Discriminant Analysis (LDA) [16], Support Vector Machine (SVM) [17], and Naïve Bayes (NB) [18] have been proposed and focused on optimizing the Accuracy (ACC) of the system. However, examining the performance of authentication models based on the ACC metric can be flawed if we have an imbalanced dataset [19]. To comprehensively evaluate an authentication model, other essential metrics like the False Acceptance Rate (FAR), True Positive Rate (TPR), and False Rejection Rate (FRR) must be considered.

   FAR measures the rate at which the system erroneously grants access to an unauthorized person. In contrast, FRR provides insight into situations where the system incorrectly denies access to an authorized individual. The Equal Error Rate (EER), the intersection point between the FAR and FRR, provides a fair and balanced evaluation [12]. Moreover, looking beyond ACC, FAR, and FRR comparisons becomes necessary due to the nuanced trade-offs intrinsic to system implementation that these conventional metrics may obscure [19]. The metrics mentioned are strongly associated with particular settings of classification threshold. Therefore, it is advisable to utilize Receiver-Operating-Characteristic (ROC) curves for visualizing outcomes, as they effectively depict the correlation between FAR and TPR (1-FRR) across various threshold values [12].

   Evaluating existing research on brainwave authentication is intricate due to inconsistent metric reporting, often limited to optimized configurations without ROC curves and variations in samples, algorithms, and experimental conditions. These unreported or unaccounted factors impact performance analysis [20]. Therefore, conducting a study that thoroughly assesses and reports the authentication system's robustness using robust metrics such as FAR, FRR, EER, and ROC curves is essential.

2. **Retraining of Authentication Models**

   A typical brainwave authentication algorithm requires the creation of a unique classifier for each individual. Accordingly, these classifiers are trained to recognize the individual designated as 'authenticated' and reject all other users labeled 'rejected.' Although this strategy was initially successful, it faced significant challenges as new users were added to the system. Each new user obligates extensive retraining of the existing classifiers, a step vital for acclimating these classifiers to the unique characteristics of the new user. This computationally demanding repeated retraining raises significant scaling issues. Additionally, it hinders the system's capacity to effectively adapt to real-world scenarios where user bases frequently change, diminishing its general effectiveness and usefulness.

To address the issue of recurrently retraining authentication algorithms inherent in traditional approaches, a solution must be formulated that harnesses advanced deep learning techniques, such as *Siamese Neural Networks*. These networks can train the model once and then utilize the pre-trained Siamese model to evaluate the legitimacy of newly introduced users to the system, overcoming the frequent retraining of the authentication models.

3. **Triviality on Open-Set Scenarios**

   It is essential to consider all the threat case scenarios when developing any authentication system based on brainwaves. The performance of EEG-based authentication systems can be evaluated using two attack scenarios: close-set and open-set scenarios. The close-set scenario assumes that the attacker is enrolled in the system and, therefore, part of the system. Conversely, the open-set strategies present a more significant difficulty as the system in this particular context recognizes the potential presence of unidentified or unauthorized individuals seeking to obtain access. The open-set scenario provides a more realistic approach since the attacker is not guaranteed to be always known to the system. Moreover, in the context of EEG-based authentication, the presumption that the authentication systems have already encountered the attacker is unrealistic since the authentication systems typically do not have access to the brain signals associated with the attacker[20]. Hence, the authentication systems must be able to identify and reject the known attackers and the attackers utterly unknown to the system. Regrettably, most studies on EEG-based authentication have focused primarily on close-set scenarios, often overlooking the security ramifications of the open-set scenarios. Thus, it is pivotal to devise authentication systems that place equal importance on exploring and addressing open-set scenarios.

4. **Lack of Research on Inter-Session variability**

   Most of the research on brainwave authentication is conducted by utilizing brain signals, usually collected during a single EEG recording session. Researchers would often split the single-session EEG data for training and testing the effectiveness of the authentication system. However, brain signals can be impacted due to the person's surrounding environment or the individual's state of mind, introducing variability in the EEG data across multiple sessions and potentially affecting the system's overall performance. Unfortunately, most researchers working on brainwave authentication systems have not addressed this issue. As a result, an extensive study must be conducted on multi-session EEG data where the robustness of the authentication system should be tested on sessions conducted on different days to investigate if the inter-session variability among users can account for a significant drop in the system's performance.

5. **Reproduciblity of Implementation**

   It is also seen in EEG studies that the parameters of the pre-processing procedures, the toolboxes utilized, and implementation techniques are often hidden or reported in a very abstract manner [21]. This lack of transparency often propels researchers to spend considerable time trying to reproduce the results reported by state-of-the-art (SOA) proposals. As a result, the process of replication and advancement in brainwave authentication is impeded due to the opaque style of reporting followed within the scientific community.

   To cultivate a more cooperative and forward-thinking scientific community, researchers must adopt a stance of transparency when disseminating the intricate aspects of their authentication system's implementation. By transparently revealing methodological specifics,

researchers can enable their peers to understand better, replicate, and validate their findings. Therefore, conducting a comprehensive study on brainwave authentication becomes essential where the specific implementation details are made transparent, providing researchers with valuable resources to evaluate and enhance their methodologies.

6. **Benchmarking Datasets**

   Although many studies are available on brainwave authentication, there is still a glaring shortage of open EEG datasets in the scientific community since most researchers chose to keep the EEG data private. Furthermore, the majority of EEG datasets that have been made available to the public involve a small number of participants (N<=25), including studies such as [22, 23, 24, 25]. These small-size datasets do not provide a complete picture of the real-world performance of brainwave authentication systems, and the results generated by utilizing those datasets could be highly optimistic as they do not capture the entire spectrum of EEG variability across a larger population. As a remedy, a comprehensive study should be undertaken, utilizing EEG datasets encompassing participant numbers exceeding 25. This approach would enhance comparing and evaluating various authentication methods using extensive datasets, promoting more comprehensive and authentic assessments.

In conclusion, because of the extreme differences in the experimental approach employed by various researchers, it is difficult to assess the actual research progress on brainwave authentication. In order to tackle this issue, I address the research question:

How do state-of-the-art (SOA) and deep learning-based brainwave authentication models compare when assessed under identical conditions?

1.3 Solution Overview

This section presents an initial overview of the envisioned proposal's framework. Within the scope of this thesis, we have undertaken two primary endeavors. Firstly, we have crafted an advanced benchmarking tool meticulously designed to encompass a range of automated machine authentication pipelines. This tool facilitates a comprehensive performance assessment of diverse EEG-based authentication models across various open EEG datasets. The benchmarking tool has been carefully designed to meet the research questions outlined in section 1.2.

Simultaneously, we engage in a comprehensive comparative analysis of distinct brainwave authentication algorithms. This analysis encompasses varying evaluation methodologies, threat scenarios, and the diverse parameters influencing authentication algorithm performance. The following points offer a concise overview of our suggested solutions tailored to address the research questions articulated in the previous section.

1. Each dataset in our benchmarking tool incorporates data from a sizeable population (N>=25). The collective number of participants from our selected datasets comes out to be 195, which is approximately a quadruple increase over earlier studies in brainwave authentication. Therefore, based on such a large population, our designed authentication framework allows for better coverage of EEG variability throughout a broader spectrum. Consequently, the results derived from our study, which includes 195 participants, will be more reliable and generalizable as they are less likely to show bias or overly optimistic expectations.

2. Our benchmarking tool employs Siamese Neural Networks alongside some SOA algorithms like Support Vector Machine (SVM), Logistic Regression (LR), Linear Discriminant Analysis (LDA), Random Forest (RF), Naive Bayes (NB) and k-Nearest Neighbours (KNN). Siamese Neural Network is a specific type of neural network with two or more identical sub-networks working in tandem. These concurrent sub-networks are trained with the same hyperparameters to generate the embedding in latent space. Such embedding serves as compact, representative vectors of the input data. This parallel configuration is then utilized to ascertain the similarity among the inputs by comparing their feature vectors [26]. One of the most significant advantages of using Siamese Network is that it mitigates the problem of retaining once a new user is enrolled into the system. Rather than retraining the entire model each time a new user is registered, Siamese Network can generate unique embeddings for the newcomer and compare it to the existing ones, thus reducing computational time significantly.

3. One of the main objectives of this study is to investigate the influence of inter-session variability among individuals arising from EEG data collection across multiple sessions. The tool has been carefully crafted to facilitate authentication assessments in datasets, including single-session and multi-session data. This approach allows us to evaluate the performance of authentication algorithms when applied to EEG data collected from diverse sessions. Consequently, we can better comprehend the ramifications of inter-session variability on human authentication.

4. This thesis delves into a crucial research area that highlights the importance of understanding both close-set (known attackers) and open-set scenarios (unknown attackers) for evaluating the effectiveness and real-world application of brainwave authentication systems. Hence, our study goes beyond focusing solely on close-set scenarios and strongly emphasizes open-set scenarios. As a result, our tool has been diligently designed to facilitate the evaluation of diverse authentication methods across both close-set and open-set threat scenarios.

5. Our tool evaluates the performance of authentication algorithms using a variety of evaluation metrics such as TPR, FAR, FRR, EER, and ROC-Curves, all of which ensure unbiased results—particularly ROC-Curves, less sensitive to imbalanced datasets [19]. We also report FRR corresponding to FAR at 1%, which is essential to balance the system's security and usability. Lower FAR correlates to security's enhanced security measures while FRR pertains to the system's ease of use [20]. Therefore, it is imperative to ascertain whether lowering the FAR threshold to increase security may unintentionally render the system less usable. Utilizing all the evaluation mentioned above metrics in our study provides an effective solution to the first research question outlined in the previous section.

6. Our tool has been specifically designed to meet the needs of researchers active in brainwave authentication. One of the main objectives of our study is to build a framework that should alleviate the time-consuming processes of pre-processing, feature extraction, parameter selection, and classification. The framework's adaptability allows it to integrate with the new data provided by the researchers seamlessly. Our framework significantly reduces the time burden for researchers and offers essential guidance in determining the optimal parameters for their studies.

1.4 Thesis Structure

In this chapter, we have articulated the primary motivations driving this study, outlined the research questions we want to answer, and given a rough outline of the approach we want to take. The next chapter will deal with brainwave authentication's foundations, including the core concepts of brainwave authentication, such as standard EEG devices, data acquisition procedures, data processing methods, and authentication algorithms. Chapter 3 will be focused on the current proposals and state-of-the-art research works compared to the challenges considered in this thesis. In the subsequent chapter 4, we offer an in-depth analysis of the surveyed open datasets and provide a high-level overview of the workflow within our benchmarking suite. The practical implementation and methodologies devised by us to build the framework will be described in detail in chapter 5. The evaluation aspects of this study will be discussed in chapter 6. In this chapter, we do two kinds of evaluation. One aspect to consider is the evaluation of our benchmarking tool itself. This assessment involves conducting tests on our tool using different authentication parameters to determine if our tool effectively addresses the issues outlined in section 1.2. The second phase of the assessment employed our tool to reproduce findings from previous studies on benchmarking brainwave authentication systems. Subsequently, a comparison study was conducted between the replicated results obtained using our tool and the original findings of the studies as mentioned above. In the upcoming chapter 7, the limitations inherent in our study will be discussed. Chapter 8 concludes with a discussion of this study's findings and potential future enhancements to the proposed benchmarking tool.

2 Background

Biometric authentication generally uses unique physical or behavioral traits of humans that are generally collected through sensors, processed, and compared to stored samples for access. Brainwave patterns, however, require specific tasks or stimuli for acquisition, making them distinct from other biometrics [20]. As depicted in Figure 2.1, the EEG-based authentication system is segmented into distinct phases, encompassing data acquisition, subsequent data processing, and the pivotal classification model responsible for user verification.

- **EEG measurement Devices (Section 2.1)**: EEG devices encompass various equipment, from sophisticated medical-grade headsets to user-friendly and portable consumer devices. In addition to headsets, the EEG experiment necessitates using amplifiers that magnify the subtle brain signals, enabling their analysis.
- **Data Acquisition Procedures(Section 2.2)**: The processes or tasks for EEG data acquisition that need to be carried out to induce distinct brainwave activity and capture the related changes in voltage levels. [20].
- **EEG Pre-Processing (Section 2.3)**: In this stage, the objective is to eliminate any interference from the EEG data obtained during the acquisition process, thereby enhancing the clarity and quality of the EEG signals.

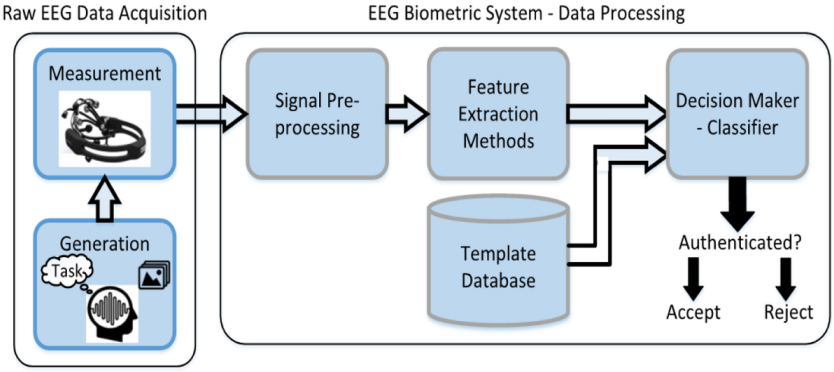


Figure 2.1: An authentication process based on EEG involves a sequence of essential stages, including data acquisition, data processing, and the classifier acting as the decision maker [20].

- **EEG Feature Extraction (Section 2.4)**: Identifies and extracts the distinctive signal characteristics essential for the process of authentication, enabling accurate verification.
- **Authentication Algorithms (Section 2.5)**: Constructs the subject classification models aimed at determining whether the individual is verified or not.

2.1 EEG measurement Devices

To acquire the EEG) signals, an EEG apparatus necessitates the inclusion of sensors capable of establishing conductive contact with the scalp. An amplifier that facilitates real-time filtering and common mode rejection, an analog-to-digital converter (A/D converter), and a personal computer (PC) to store the digitized data [27]. The EEG headset, which records the brain activity, includes sensors arranged according to the 10-20 or 10-10 system, which are international standards for consistent placement of electrodes, ensuring comparability across subjects and studies. The naming of the electrodes within these systems follows a specific convention that represents the brain region underneath (e.g., F for frontal, C for central) and an odd or even number indicating the hemisphere (odd for left, even for right) [28]. Frequently utilized EEG devices include the ActiveTwoSystem designed by Biosemi (Amsterdam, Netherlands)[1] and the g.GAMMAcap developed by G.Tech Medical Engineering GmbH[2]. An example EEG headset from G.Tech is illustrated in Figure 2.2 (a). These medical-grade EEG systems can support up to 256 channels, allowing for comprehensive data collection. This feature offers a benefit as it facilitates a greater extent of spatial coverage on the scalp, resulting in a more comprehensive dataset [27].

Although medical-grade EEG devices are known for their ability to gather data of high quality, the considerable expense associated with these devices and the complexities needed in establishing the EEG connection pose notable obstacles. In response to these constraints, there has been a proliferation of cost-effective and user-accessible EEG devices in recent times, presenting viable substitutes. Instances of such devices include the ENOBIO developed by Neuroelectrics (Barcelona, Spain)[3], the EPOC/EPOC+ wearable neuroheadset designed by Emotiv Systems, Inc. (San Francisco, USA)[4], along with the Muse headband crafted by InteraXon (Ontario, Canada)[5]. An EPOC/EPOC+ wearable EEG headset equipped with 14 sensors is depicted in Figure 2.2 (b). Consumer devices are cheaper than medical-grade EEG headsets and more friendly but have a poor signal-to-noise ratio [27].

2.2 Data Acquisition Procedures

The activity of neurons in the brain is notably impacted by individuals' mental conditions, displaying a marked sensitivity to both external environmental triggers and internal self-control mechanisms. This necessitates the development of tailored data collection approaches for capturing EEG signals effectively [31]. EEG data acquisition often entails implementing meticulously planned EEG experiments, wherein subjects engage in a range of cognitive tasks or maintain a state of rest, with the option of having their eyes either open or closed [27].

Resting-related tasks are the simplest to accomplish. Typically, resting-state tasks involve recording brain activity when participants are calm, relaxed, and not performing cognitive tasks.

[1] http://www.biosemi.com/
[2] https://www.gtec.at/
[3] https://www.neuroelectrics.com/
[4] https://www.emotiv.com/epoc/
[5] https://choosemuse.com/

(a) Medical grade EEG headset from g.tec [29]

(b) EPOC/EPOC+ wearable headset [30]

Figure 2.2: In Figure (a), we can observe the g.GAMMAcap, which is among the frequently utilized medical-grade EEG headsets. Moving on to Figure (b), we see the EMOTIV EPOC+ device equipped with 14 EEG electrodes.

As a result, resting state protocols have been employed in many brainwave authentication studies such as [32, 33]. Paranjape *et al.* [34] suggested an EEG-based authentication system while subjects sit in a relaxed state with closed/open eyes task. The Autoregressive model (AR) was utilized for extracting the discriminant biological features, and a classification ACC of 80% was achieved. Although resting tasks provide a straightforward approach to data collecting, they are susceptible to external influences, making it challenging to guarantee a completely silent environment in a real-world application setting [31].

In contrast to protocols for rest, protocols for cognitive activities are characterized by a higher degree of complexity. One classification of cognitive protocols pertains to mental tasks. Mental tasks involve the subject imagining something specific (e.g., imagining moving their left and right hand or imagining closing or opening a fist), causing the associated EEG signals to appear [27]. In a study by Brigham and Kumar [35], the brain activity of six participants was examined. These individuals were directed to imagine the articulation of two syllables, namely /ba/ and /ku/, while deliberately shifting the rhythm. The participants needed to be provided explicit instructions regarding the desired rhythm. Following a similar approach as observed in the study by Paranjape *et al.* [34], the researchers in the study also employed AR coefficients to extract features from the signals of each electrode. Utilizing the SVM classification model, their results demonstrated an impressive accuracy rate 97.76%. Mental tasks demonstrate suitability for people across various physical limitations and visual impairments, exhibiting a high degree of applicability [31]. Nevertheless, it is worth noting that motor imagery and mental tasks demand specialized training to generate proper brain responses, making them challenging to execute [35].

The other type of cognitive protocol is based on event-related potentials (ERP). ERPs are a particular type of evoked potentials, time-locked to brain variations that appear in reaction to external stimuli [31]. They are generally elicited by exposing subjects to external audio or visual stimuli. ERPs can be influenced by various factors, including the subject's level of knowledge, motivation, and cognitive abilities, making them more likely to manifest unique and

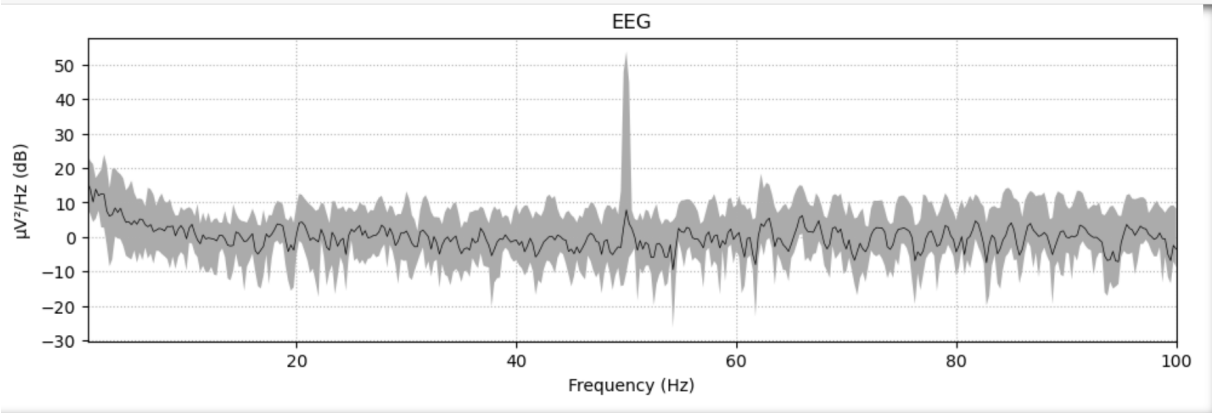


Figure 2.3: EEG signal strength against frequency, 0–100 Hz, showing 50 Hz line noise.

distinctive characteristics beneficial for authentication objectives [36]. Mu *et al.* [37] introduced an innovative approach to ERPs by exposing participants to stimuli in the form of self-photos and non-self-photos. They utilized fuzzy entropy extracted from the EEG signals for personal authentication. They employed Neural Networks (NN) to classify these features, yielding an impressive accuracy rate exceeding 87.3%. Compared to EEG, one notable drawback of ERPs is the increased complexity of the elicitation methods associated with ERPs. EEG can be obtained without needing specific stimulation of the user, but ERPs can only be obtained when the user is subjected to a particular and carefully controlled kind of stimulation [27].

2.3 EEG Pre-Processing

After the EEG data is acquired, it is imperative to eliminate any noise intruding on the EEG signals to obtain clean and precise readings. EEG signals can be corrupted by various artifacts, either of a physiological or non-physiological nature. Physiological artifacts are non-EEG signals introduced by different biological activities such as heartbeat, muscle contractions, or eye movements. In contrast, non-physiological artifacts typically arise from the EEG acquisition system or external environmental factors, including electromagnetic fields from the electronic devices [27]. Following are some of the standard cleaning processes usually employed in EEG.

2.3.1 EEG Filtering

As previously stated, EEG data is susceptible to interference from electrical appliances, leading to noise mostly at frequencies of 50 Hz in Europe (as depicted in Figure 2.3) and 60 Hz in the USA. This phenomenon arises due to the prevailing power line frequencies in the corresponding geographical areas. Therefore, employing a notch filter, specifically a band-stop filter with a small stopband centered at 50 Hz or 60 Hz, is common practice to eliminate line noise effectively [38]. Other filtering methods include low pass filtering, which entails the elimination of higher frequencies; high pass filtering, which eliminates frequencies below a specific threshold while preserving high frequencies [39]; and bandpass filtering, which combines the above filtering techniques. The bandpass filter selectively keeps frequencies within the specified upper and lower bounds while eliminating frequencies that do not fall within this range. Furthermore, the Butterworth filter, known for its maximally flat magnitude in the passband, is widely used in pre-processing EEG data [27].

2.3.2 Artifacts Rejection

One often employed approach for mitigating physiological artifacts is discarding EEG data over a predefined threshold of EEG voltage [40], such as 100μV since physiological artifacts generally exhibit significantly greater magnitudes in comparison to cerebral activity [27]. However, a more sophisticated method of rejecting artifacts is the peak-to-peak rejection method. This approach involves identifying and removing EEG segments that surpass a predetermined voltage range. This range is commonly determined by measuring the peak-to-peak amplitude, which is the difference between the highest positive and lowest negative deflections within a specific time frame [41]. While the peak-to-peak rejection technique efficiently eliminates noisy signals, its application can inadvertently remove substantial valuable EEG data. This outcome is particularly possible when the method is not executed after meticulous assessment, as the amplitude of EEG signals can significantly differ based on experimental configurations and the characteristics of the utilized headsets.

2.3.3 Independent Component Analysis

The Independent Component Analysis (ICA) is an innovative signal processing technique that enables the separation of sources that have been linearly mixed at the sensors. This separation is achieved by assuming simply the statistical independence of the sources [42]. According to Hyvärinen *et al.* [43], the ICA algorithm can be defined by the following equation.

$$S \cdot X = U \tag{2.1}$$

where S is the unmixing matrix, X is the signal of EEG channels where each row corresponds to a sensor channel, each column corresponds to a time point in the recorded signals, and U represents the matrix of the estimated independent source signals. This method enables the segregation of EEG signals from noise and randomly mixed signals, contributing significantly to enhanced signal quality and analysis.

2.4 EEG Feature Extraction

The feature extraction technique is vital in EEG-based authentication, transforming pre-processed EEG signals into concise yet informative representations [44], facilitating accurate subject classification. EEG features can be organized into various domains, encompassing the time domain (such as Autoregressive Coefficients), the frequency domain (like Power Spectral Density), and the time-frequency domain (including Wavelet Transform). The subsequent methods are frequently utilized for feature extraction in EEG-based authentication studies.

2.4.1 Autoregressive Coefficients

Autoregressive (AR) coefficients are a class of time-domain features frequently utilized in EEG-based authentication. The AR model is a form of linear regression that involves regressing the current observation of a time series against one or more previous observations of the same series [45]. The following equation can mathematically represent the AR model [27]:

$$x(n) = -\sum_{i=1}^{p} a_i x(n-i) + e(n). \tag{2.2}$$

Where x(n) represents the current value of a particular channel, a_i denotes the AR coefficients at specific delay i, e(n) represents the error at time n, and p represents the order of the model.

Estimating AR coefficients can be accomplished using methods including Yule-Walker and Burg. The coefficients offer valuable insights into the temporal dynamics of EEG data [46], hence contributing to the characterization of subject-specific patterns and assisting in the authentication process. Hine *et al.* [47] proposed an EEG-based biometric recognition system that employs AR coefficients extracted through the Burg method to capture distinguishing features from a cohort of 50 subjects engaged in the study. In contrast to employing any state-of-the-art (SOA) machine learning algorithm for subject identification, this study adopted a different approach by utilizing the Manhattan distance to measure the similarity of the features. This method used the Manhattan distance metric to compare the enrolled samples with the corresponding test samples from the same subject.

2.4.2 Power Spectral Density

Transforming EEG data into the frequency domain facilitates extracting and discriminating prominent frequency components. EEG signals can be divided into numerous frequency bands, such as δ (1-4 Hz), θ (4-8 Hz), α (8-12 Hz), β (12-30 Hz), and γ (30-45 Hz). These frequency bands correspond to various types of brain activity [27]. The delta wave is a prominent oscillatory activity observed within the 1–4 Hz frequency range during deep or slow wave sleep. It is primarily associated with attention to internal cognitive processes. On the other hand, the theta wave, ranging from 4–8 Hz, is more closely linked to memory retrieval and access. The alpha wave, falling within the frequency band of 8–14 Hz, is predominantly generated in the parietooccipital region during states of relaxation with closed eyes. In contrast, the beta band, spanning 14–30 Hz, is specifically associated with the conscious perception of stimuli. Lastly, exceeding 30 Hz, the gamma band is involved in the transient functional integration of neural activity across different brain regions [48].

The Power Spectral Density (PSD) is employed to represent the power distribution of a signal across different frequency points [31], and it is calculated by various methods such as Fourier Transformation (FT) or Discrete Fourier Transformation (DFT) using Welch's periodogram algorithm. Welch's algorithm has been utilized in many studies to estimate frequency bands' power spectrum. Welch's technique involves segmenting the input signals with overlap and generating the periodogram by squaring the magnitude of the Discrete Fourier Transform (DFT) [27]. This approach reduces the variance from the signal by averaging the overlapping segments of the periodograms [49] and thus producing an unbiased power spectral estimate of the different frequency bands in EEG. A pertinent example of the application of Welch's periodogram algorithm can be found in the work of Hema *et al.* [50]. In their study, the authors harnessed the potential of Welch's method to compute the PSD of EEG Beta waves. Ericsen *et al.* [51] employed Welch's approach of PSD estimation, which is based on the Fast Fourier Transform (FFT) technique, to extract the band power of 4 frequency bands, namely θ, α, β, and γ.

2.4.3 Wavelet Transform

The Wavelet Packet Decomposition (WPD) emerges as a unique evolution of the discrete wavelet transform (DWT), recognized for its augmented filtration technique applied to the discrete temporal data [52]. This amplified filtration approach empowers the WPD to achieve an intricate, multi-level dissection of signals across the time-frequency spectrum [53]. The WPD approach provides a wider range of frequency resolutions compared to the traditional discrete wavelet transform. In contrast to the DWT, which decomposes a signal into its core approximation and detail coefficients components, the WPD technique follows a more detailed approach. The analysis extends beyond the fundamental levels of detail and approximation coefficients, exploring more intricate layers of complexity. The WPD approach exhibits divergence by systematically

unraveling the signal's coefficients, encompassing intricate details and broad patterns. This process leads to constructing a comprehensive and complex wavelet packet tree [45].

2.5 Authentication Algorithms

Following the transformation of EEG signals into distinctive features, the subsequent stage involves subject classification based on these extracted features. This classification method treats each individual as a separate category, and the extracted features are organized into categories corresponding to each person. The classification model is then trained using supervised learning to establish a direct mapping between the features and individual identities. This mapping enables authentication through a one-to-one relationship between the extracted features and the respective individuals [31]. In more recent times, the application of deep learning methods has also gained traction in brainwave authentication research. Below, we present a compilation of some of the frequently employed state-of-the-art (SOA) and deep learning authentication algorithms for EEG-based authentication systems.

2.5.1 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) represents a traditional linear learning technique that aims to identify a linear combination of attributes across diverse categories to characterize or discern them [31]. The primary objective of LDA is to employ hyperplanes to segregate data from distinct classes. This segregation is achieved by projecting the data into a lower-dimensional space. LDA seeks to optimize the separation between classes by maximizing the inter-class distance. This optimization is carried out under the assumption of normal data distribution, ensuring equality of covariance matrices across various categories [27]. LDA stands as a widely employed classifier within the realm of brainwave authentication studies. Rocca *et al.* [54] conducted research in which they gathered EEG data from a sample of 36 participants while they were in a state of rest with their eyes closed. Bump modeling was employed to extract pertinent features from the raw EEG signal, and the classification task was executed using the LDA classifier. The study produces exceptional results, with ACC reaching as high as 99.69%. In the study by Koike-Akino *et al.* [55], EEG signals were collected from 25 subjects in an ERP-focused EEG experiment. To enhance the efficiency of feature extraction and address the high dimensionality of EEG data, the researchers utilized Principal Component Analysis (PCA). Employing LDA for classification, the team achieved a remarkable accuracy rate of 96.7%, underscoring the effectiveness of their approach in accurate subject identification.

2.5.2 Support Vector Machine

The Support Vector Machine (SVM) is a binary classification model that employs a hyperplane to classify two distinct data classes by optimizing the margin, which refers to the distance between the closest training instances from different classes [27]. SVMs are known for their strong generalization skills. What makes SVM particularly powerful is its emphasis on the points lying closest to the margin, called support vectors [56]. Illustrated in Figure 2.4, the support vectors play a pivotal role in positioning the optimal hyperplane [57], enabling SVM to prioritize the most informative instances and mitigating the impact of outliers [58]. This characteristic makes SVM inherently resistant to overfitting [59], ensuring that the model generalizes well to new, unseen data. Furthermore, SVM is not only limited to linear separations; it can also utilize kernel functions to transform the data into a higher-dimensional space [58]. This transformation can help SVM capture complex and nonlinear relationships between features, enabling it to handle intricate decision boundaries that may not be possible in the original feature space.

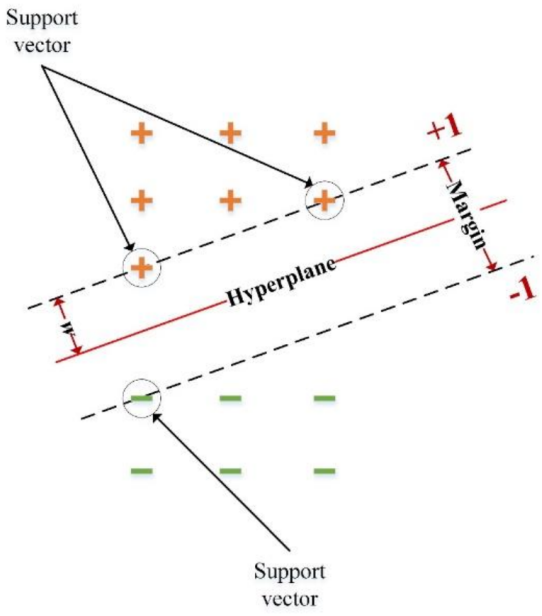


Figure 2.4: An SVM hyperplane visually represents the maximum separation between the support vectors associated with the two classes, i.e., positive and negative [57].

SVMs have been widely utilized in brainwave authentication research because they can process intricate EEG data and effectively deliver precise subject identification. Pham *et al.* [3] employed an SVM algorithm to analyze EEG data from a cohort of 9 participants actively involved in motor imagery activities. The researchers extracted vital features from the EEG signals, precisely AR linear parameters, and PSD components. The results showcased a remarkable performance with an EER spanning from 0% to 3.3%.

2.5.3 Logistic Regression

Logistic Regression (LR) is a statistical approach extensively employed in machine learning to address binary classification challenges. The sigmoid function, often the logistic function, determines the relationship between input attributes and the likelihood of belonging to a particular class [60]. Analyzing the training data, the algorithm selects an optimal decision boundary that effectively segregates the two classes [61]. LR has found its application in EEG-based authentication systems because it can effectively characterize complex relationships within brainwave patterns for subject classification tasks. Piplani *et al.* [62] an EEG-based identity authentication method utilizing two publicly available datasets involving 31 subjects. They employed the XGBoost with LR classifier for classification, achieving a baseline accuracy of 90.8% in their study.

2.5.4 K Nearest Neighbour

The K-Nearest Neighbour (KNN) is a straightforward, non-parametric technique. It implies that it arrives at decisions by considering the majority consensus of the nearest or most akin data points to the given inputs [27]. KNN operates through a two-step process. During the initial stage, it identifies the data points close to the target data point. The determination of closeness is accomplished via the use of distance metrics such as Euclidean or Manhattan distance. In the subsequent step, the algorithm assigns the target data point to a particular class based on the

classes of its neighboring data points [63]. Zúquete *et al.* [64] conducted a study that employed visual stimulation to elicit brain responses from 70 individuals, with the objective of biometric identification. The KNN classifier achieved an average Area Under Curve (AUC) of 0.9817, indicating strong performance in discriminating individuals based on their brain responses.

2.5.5 Gaussian Naive Bayes

The Naive Bayes (NB) theorem is a method rooted in probability theory, specifically Bayes' theorem, which elucidates the likelihood of a specific event occurring given prior knowledge about associated occurrences. Gaussian Naive Bayes (GNB) is a particular variant of the Naive Bayes algorithm that assumes the features follow a Gaussian (normal) distribution [65]. Put differently; it is assumed that the continuous values of attributes belonging to each class follow a normal distribution. This opposes the conventional assumption made by Naive Bayes that features are categorical and adhere to discrete distributions. Valsaraj *et al.* [66] conducted a comprehensive analysis of EEG signals to identify distinctive features associated with both physical movement and imagined upper limb motions. This research endeavor encompassed 10 participants and focused on four distinct upper limb movements. Employing the GNB algorithm for authentication purposes, the study achieved an impressive accuracy rate of 89% for imaginary motions and 85.7% for physical movement tasks.

2.5.6 Random Forest

The Random Forest (RF) [67] algorithm is a popular ensemble learning technique employed in machine learning to address classification and regression problems. It creates multiple decision trees during the training phase and then combines their predictions to make more accurate and robust predictions [68]. One key benefit of RF is its ability to handle high-dimensional feature spaces [69]. EEG data frequently encompasses a substantial number of channels and temporal dimensions, leading to a considerable quantity of features. Random Forest's feature selection mechanism and ensemble approach allow it to effectively manage these complex feature sets, preventing overfitting and enhancing generalization [69]. Another advantage is RF's resilience to noisy data [67]. As previously mentioned in section 2.3, it is essential to acknowledge that EEG signals are prone to several artifacts and sources of noise, which have the potential to impact the accuracy of categorization. RF ensemble approach mitigates this problem by averaging out the impact of noise, improving the overall robustness of the authentication system. RF has been employed in various EEG-based authentication works because of its adaptability to high-dimensional data, capability to handle complex relationships, and robustness against noise. In the study by Chowdhury and Imtiaz [70], EEG data was collected across three consecutive sessions involving 21 subjects. The research showcases that the proposed machine learning model, based on the RF algorithm, demonstrates an authentication accuracy of approximately 83.2%.

2.5.7 Deep Learning

The authentication algorithms mentioned above are state-of-the-art machine learning techniques commonly employed in EEG-based authentication studies. However, a limitation inherent in these methods lies in their dependency on a discriminative feature set resulting from the feature extraction process. Furthermore, many of these machine learning approaches primarily address static data, rendering them less proficient in accurately classifying EEG signals that exhibit temporal variations [31].

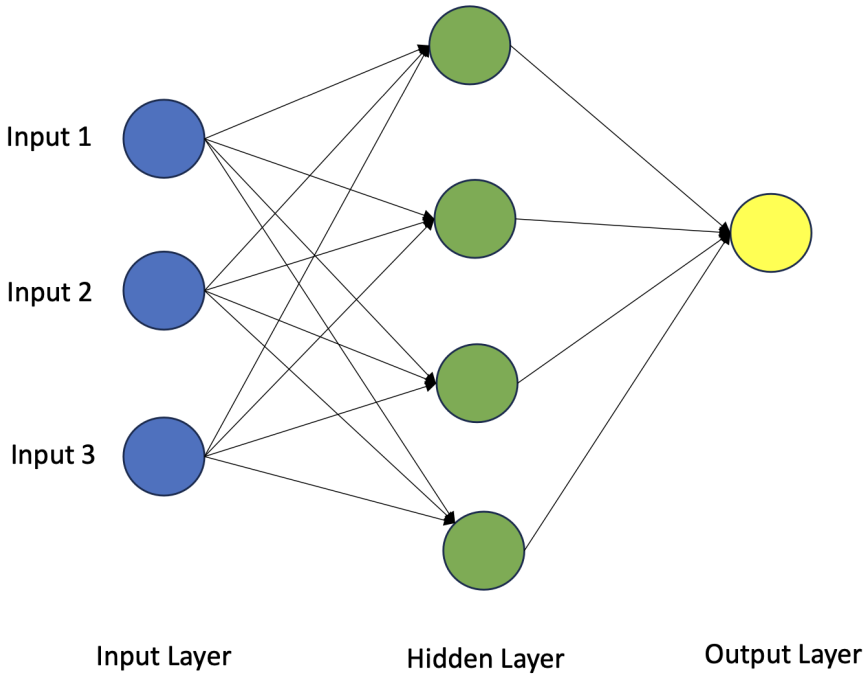


Figure 2.5: A sample configuration of a Neural Network structure encompassing an input layer, a single hidden layer, and an output layer.

Deep learning methods, such as neural networks (NN), have emerged as robust solutions to address this limitation. As depicted in Figure 2.5, the standard approach to developing a neural network often includes using a multi-layer perceptron design. This architecture consists of three main layers: the input, multiple hidden, and output layers. The network utilizes the feedforward mechanism in combination with the backpropagation algorithm to ease the training of data and the creation of weight matrices. Consequently, predictions can be derived using the ascertained weight matrix [27].

Numerous deep learning methodologies, including Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM) networks, have found widespread application in a multitude of EEG-centered authentication studies due to their ability to find intricate brain patterns in the raw EEG signals. As an illustration, in a study by Yu *et al.* [71], the authors utilized a direct input approach, feeding the raw EEG signals of 8 subjects into a CNN architecture. The input dataset encompassed 534 time points gathered from 44 EEG channels. This devised framework yielded an impressive accuracy rate of around 97%, with a remarkably low FAR of 0.06% and a minimal FRR of 3.15%. This approach demonstrates the efficacy of CNN models in processing raw EEG signals for robust authentication purposes.

3 Related Work

Brain signals unique attributes and individualistic patterns have attracted considerable research on constructing brainwave authentication systems. Many researchers have presented various studies on brainwave authentication systems, utilizing different classifiers, EEG acquisition tasks, and distinct features. However, due to the considerable diversity in experimental approaches adopted by various researchers, evaluating the advancements in brainwave authentication research has become complex. Recent studies have emerged to address this challenge by undertaking comprehensive comparative analyses. These analyses aim to provide a clearer understanding of the efficacy and performance of brainwave authentication methodologies. Moreover, researchers have also sought to investigate cross-session variability among individuals, further enhancing the depth and comprehensiveness of the research landscape. Alongside these efforts, some studies have delved into advanced deep learning techniques, such as Siamese Networks, as potential solutions to issues like retraining within SOA algorithms when new users are added to the system. Consequently, this section will explore a range of relevant studies that closely align with the research goals of our study, including research that benchmarks brainwave authentication algorithms, examines inter-session variability among individuals, and investigates Siamese Networks as a potential solution to the retraining issue in SOA algorithms.

3.1 Previous Research on Evaluating and Comparing Brainwave Authentication Methods

In brainwave authentication, benchmarking studies play a vital role in setting new standards and evaluating the effectiveness of newly proposed methods compared to the SOA authentication algorithms. This section overviews some of the benchmarking studies conducted for brainwave authentication. Khalafallah *et al.* [72] conducted a comprehensive analysis of authentication algorithms, including LR, SVM, and LDA. They gathered EEG data from 29 participants wearing Neurosky Mindwave[1] headsets and 10 wearing Emotiv[2] headsets while subjects were resting with their eyes closed for 50 seconds. The study found that LR performed the best among the algorithms tested. With the Mindwave dataset, they achieved a false acceptance error (FA) of 3% and a higher false rejection error (FR) of 48%, resulting in overall accuracy (ACC) of approximately 80%. The Emotiv dataset achieved an FA of 0.3%, an FR of 13.93%, and an ACC of 92.88%. Jayarathne *et al.* [73] took a distinct approach in their benchmarking study

[1] https://store.neurosky.com/pages/mindwave
[2] https://www.emotiv.com/epoc/

of EEG-based authentication systems. They gathered EEG data from 12 participants through an ERP task involving visual simulations of 4-digit numbers on a screen. The study focused on SVM, LDA, and KNN algorithms for analysis. After assessing accuracy across different EEG channel combinations, the study concluded that the most effective classifier was KNN with an ACC of $99.0 \pm 0.8\%$, followed by SVM with $98.03 \pm 0.1\%$ ACC and LDA with $98.01 \pm 0.5\%$ ACC.

Fang *et al.* [74] conducted a study using the publicly available DEAP [75] dataset, which included 32 participants. EEG data collection in the dataset involved participants watching a 40-second music video depicting five emotions: neutral, angry, sad, happy, and pleasant. Feature extraction was performed using PSD and Differential Entropy (DE). The study encompassed a comparative analysis of several algorithms, including KNN, RF, SVM, and a modified algorithm, the Multi-Feature Deep Forest Method (MFDM), an extended version of the RF classifier. The study's findings indicated that the average accuracy for MFDM, RF, SVM, and KNN was approximately 71%, 68%, 52%, and 63%, respectively. The outcomes derived from this study were not particularly promising. A comprehensive comparative study by Arias-Cabarcos *et al.* [12] in 2021 encompassed an extensive dataset of 52 subjects engaged in five distinct ERP tasks. The study involved feature extraction through AR coefficients and power spectrum (PS) analysis of α, β, and γ frequency waves. Among the four SOA algorithms, including SVM, KNN, GNB, and LR, the GNB classifier demonstrated the best performance, achieving an EER of 14.5%, followed by SVM with 40% EER and KNN with 47% EER.

Deep learning methods have recently gained prominence in various comparative analyses concerning brainwave authentication, aiming to assess their effectiveness compared to SOA algorithms. One such study conducted by Huang *et al.* [76] incorporated both SOA algorithms such as NB, LR, and a deep learning approach known as Back Propagation Neural Network (BPNN). EEG data was gathered from 30 participants as they engaged in ERP tasks that included auditory and visual stimuli. The study extracted seven statistical features from the data, such as mean, median, standard deviation, entropy, maximum, minimum, and skewness, to get a comprehensive insight into the data distribution, central tendency, and variation. NB demonstrated the worst performance among the classifiers, registering average ACC, TPR, and False Positive Rate (FPR) of 77.96%, 75.71%, and 19.80%, respectively. LR exhibited superior performance compared to NB, achieving average ACC, TPR, and FPR values of 81.59%, 79.04%, and 15.05%, respectively. Remarkably, the BPNN method showcased exceptional performance, boasting average ACC, TPR, and FPR of 82.69%, 81.96%, and 17.38%, respectively. Meanwhile, Zhang *et al.* [77] conducted an extensive comparative analysis on a broader spectrum of authentication algorithms, encompassing five distinct methods: KNN, Bagging, RF, AdaBoost, and NN. The study utilized two well-established public datasets, namely the Fantasia ECG dataset [78, 79] and the UCI EEG dataset[3]. The study included EEG data of 20 subjects from each dataset. On the UCI dataset, which features EEG data, RF demonstrated the highest accuracy at 86%, trailed by NN and KNN. However, classifiers Bagging and AdaBoost displayed relatively poorer performance, achieving accuracy levels of only 66.7% and 73.9%, respectively.

3.2 Siamese Neural Networks in Brainwave Authentication Studies

As noted in section 1.2, most brainwave authentication studies employed SOA machine learning algorithms to discern between genuine users and imposters. These models often require the learning algorithms to retrain whenever new users are added to the system, which reduces the

[3] https://archive.ics.uci.edu/dataset/121/eeg+database

model's effectiveness and hinders practical application [26]. Some studies proposed a solution to this problem by employing deep learning procedures to learn embeddings of the brain signals and subsequently calculating similarities between them. Following this approach, Bidgoly *et al.* [80] presented a notable study employing the publicly available Physionet dataset [81] for brainwave authentication. The dataset contains EEG recordings from 109 subjects, captured as the subjects performed resting tasks for 5 seconds. The study utilized CNN to generate the brain embeddings during training and verify the authenticity of the new users by comparing their data with the stored samples using similarity metrics like Cosine Similarity, Euclidean Distance, and Manhattan Distance. The best-performing similarity function was Cosine Similarity with EER of just 1.96%, followed by Manhattan and Euclidean with 3.91% and 5.65% EER, respectively. The study provides a more realistic scenario and addresses the critical challenge of identifying new users whose brain data were not introduced during training. However, this approach may not be universally accepted since deep learning methods like CNN often require large amounts of data to optimize parameters during the model's training, an aspect often impractical given the limited size of most brainwave datasets [82].

Maiorana [83] proposed a broader solution to overcome the problem of frequent retraining and to obtain the results with minimal EEG samples by employing the Siamese Neural Network approach. The study aimed to perform EEG-based verification and investigate the effects of intra-class variability across subjects whose brain signals were collected in 5 sessions over 15 months. Two identical CNNs received inputs in the form of the pre-processed brain samples and, then, were trained with the same parameters and weights to produce the brain embeddings. Afterward, the similarity of these embeddings was computed using Euclidean distance. The achieved EER was less than 7% for the 30-second verification probe, a significantly good result considering the cross-session variability in brain data.

Lately, Fallahi *et al.* [26] presented their work on Siamese Networks for brainwave-based recognition in verification and identification mode. The study was conducted employing the EEG recordings from two publicly available EEG datasets such as **BrainInvaders (b12015a)** [84] and **ERP Core** [85]. Unlike Maiorana's [83] methodology, which used contrastive loss function for determining the similar and dissimilar brain embeddings, Fallahi *et al.* opted for a triplet loss function for their approach. As a result, three sub-networks, each with five convolution layers, produce embeddings, which were then evaluated under both close-set (i.e., seen attackers) and open-set (i.e., unseen attacker) scenarios. In verification mode, the calculated EERs for the close-set scenario were notably less than those of open-set scenarios, with dataset b12015a having an EER of a mere 0.14% for seen attackers. Similar trends were seen in identification mode where EER for the dataset b12015a was 0.34%, the lowest among all the datasets.

3.3 Existing studies exploring cross-session variability

Although studies investigating the effects of inter-session variability in brainwave authentication are scarce, researchers have focused on this area. One of the most extensive works on this area was done by Huang *et al.* [86] in 2022, who explored EEG variability across sessions, subjects, and tasks. The study contains EEG data from 106 subjects; 96 out of 106 participated in two sessions on different days. Six paradigms, including resting state, transient state sensory, steady state sensory, cognitive oddball, motor execution, and steady-state sensory with selective attention, were conducted throughout the entire EEG experiment. 12th-order AR, PSD, and Mel Frequency Cepstral Coefficients (MFCCs) were chosen to extract the discriminant features from the brain signals, and the SVM classifier was employed to perform the identification and verification task. Additionally, Huang et al.'s research included both within-session and cross-session evaluations in the context of identification and verification. There was a noticeable

performance decline in the cross-session evaluation compared to the within-session evaluation. In the verification task, the average EER across all paradigms increased twofold, escalating from 0.16 in within-session evaluation to 0.32 in cross-session evaluation. Similarly, in the identification task, the average accuracy fell dramatically from 0.70 in within-session to 0.31 in cross-session evaluation. This study's results show the necessity for more significant research into EEG variability across sessions and subjects.

While the results of Huang et al.'s cross-session evaluation were inferior, Seha and Hatzinakos [87] in 2020 produced impressive results in a similar study area using steady-state Auditory Evoked Potentials (AEPs) for EEG-based recognition. The study involved an EEG experiment on 40 subjects across two sessions on separate days. The study demonstrated exceptional results even when evaluated across cross-session (multiple-sessions). EER of a mere 2-4% was achieved in cross-session evaluation, which is 16 times more effective than that of Huang et al.'s work on cross-session evaluation.

4 Solution Approach

This study aims to develop a benchmarking suite for EEG-based authentication and incorporates various open medical-grade EEG datasets that include a substantial number of participants (n>100). The performance and robustness of the different authentication algorithms will be compared with appropriate metrics to determine which algorithm is most effective and on which dataset. As illustrated in Figure 2.1, the initial step in constructing a brainwave authentication pipeline involves EEG data collection. Therefore, as a starting point, we conducted a thorough survey of numerous EEG datasets, elaborated upon in section 4.1.

4.1 Survey Open Datasets

Creating an efficient and robust EEG benchmarking framework involves collecting high-quality EEG datasets, which is essential since the quality of datasets can significantly influence the overall effectiveness of the framework. The following issues can arise if poor-quality datasets are used for developing the benchmarking framework:

1. **Random Classification**: Noise in the EEG data can obscure the model from identifying the meaningful brain data and random noise. It could lead the model to classify the users based on their brain data randomly.

2. **Erroneous or Biased Results**: The imbalance in the participant's population in the datasets may lead to overestimating the evaluation metrics such as accuracy [19]. Additionally, skewed datasets introduce biases into the system, so the results generated by those authentication systems cannot be trusted.

3. **Increased Pre-Processing Time**: Most of the data cleaning is done during the pre-processing stage, and considering that the bad-quality datasets also have a low signal-to-noise ratio (SNR), the researchers often spend a considerable amount of time handling the noisy data.

4. **Overfitting or Underfitting**: Low-quality datasets can induce issues of overfitting or underfitting during the construction of machine learning models. Overfitting transpires when the model's complexity becomes excessive, causing it to incorporate noise into its learning process. On the contrary, underfitting appears when the model's simplicity is inadequate in capturing the intricate patterns within the data [88]. Both situations can potentially result in incorrect predictions and a decrease in the model's effectiveness.

5. **Limited Reproducibility**: If the datasets are of inadequate quality, the other researchers would not be able to reproduce the results, questioning the reliability of the initial research.

While consumer devices are known for their user-friendly interface and simplicity, it is essential to note that the data produced by these devices generally exhibit a lower signal-to-noise ratio (SNR) than those generated by medical-grade EEG equipment. Considering the potential pitfalls of utilizing low SNR datasets, we focus on high-quality medical-grade EEG datasets for our study. Open datasets vary across the EEG headsets, the number of electrodes (channels), stimuli tasks, EEG paradigms, physical setup, and file format. As a result, researchers have traditionally recorded a new dataset or used one of the few well-known datasets when they have to validate a new approach [21]. However, recording a new medical-grade EEG dataset can be an intricate task as it requires experts' assistance to set up the devices and correctly monitor the participant's brain activity. Therefore, our study primarily focuses on harnessing publicly available high-quality EEG datasets as the first step.

Considering that ERPs have a reasonably good SNR, less susceptibility to background perturbations [89], and can assess instantaneous reactions to short stimuli [31], we propose to focus on the comparison of different algorithms based on ERP paradigms like P300 and N400 which can fill the gaps left by other data acquisition protocols and provides a more robust authentication mechanism. P300 is a positive deflection in voltage that reaches its peak at 300 milliseconds (ms) following exposure to a specific stimulus and is usually triggered using the "oddball" paradigm, in which a subject detects an occasional or rare stimulus in a regular train of standard stimuli [90]—for example, encountering a picture of an animal (a rare stimulus) in a series of images, targeting human celebrities (standard stimuli). On the other hand, N400 is a negative deflection that peaks around 400 ms after the presentation of a stimulus, and N400 responses are associated with stimuli connected to semantic processing, such as language processing [12]. As a result, we decided to exclusively survey and concentrate on the open datasets based on ERP paradigms like P300 and N400 on the internet.

Collecting quality EEG datasets was tedious since most researchers in the EEG domain do not make their datasets public because of privacy and confidentiality issues. Nevertheless, despite these obstacles, our assiduous search yielded more than 40 datasets, procured from websites known for providing repositories for high-quality EEG datasets, such as **OpenBCI** [1], **Zenodo** [2], **MOABB** [3], **Dryad** [4], **OSF** [5] and **Figshare** [6]. Table 4.1 and Table 4.2 list some of the publicly available P300 and N400 datasets that we reviewed during our study, organized chronologically by the year of their release. Four datasets [91, 92, 85, 93] were selected for our research, based on the ERP and the other criteria: 1) an ERP paradigm such as P300 or N400 2) raw data available 3) implementation code available 4) Multi samples per subject available 6) Number of subjects (N>=25). We chose the second condition to apply the standardized pre-processing, feature extraction, and authentication steps across all datasets. This uniform process is essential to evaluate their performance under similar experimental conditions, which is impossible without access to unprocessed raw data. As a result, we discarded datasets from our study, which only provided pre-processed data. Additionally, we did not want to utilize datasets where subjects provide a single sample because a single brain sample cannot capture the EEG variability across different instances of the same subject. Consequently, we applied the condition to include only the datasets with multiple samples per subject. In the subsequent

[1] https://openbci.com/community/publicly-available-eeg-datasets/
[2] https://zenodo.org/
[3] http://moabb.neurotechx.com/docs/datasets.html
[4] https://datadryad.org/stash
[5] https://osf.io/
[6] https://figshare.com/

sections, we provide a concise overview of the datasets included and excluded in our study.

Table 4.1: Publicly available ERP datasets based on P300 (oddball) paradigm

Dataset	Year	#Subjects	EEG Device	#Channels	Sampling Rate	#Sessions	EEG task
BrainInvaders12 [94]	2012	25	NeXus-32	16	128 Hz	1	Visual Stimuli
BrainInvaders13a [95]	2013	24	g.GAMMAcap	16	512 Hz	1	Visual Stimuli
BrainInvaders14a [96]	2014	64	g.Sahara	16	512 Hz	1	Visual Stimuli
BrainInvaders14b [97]	2014	37	g.GAMMAcap	32	512 Hz	1	Visual Stimuli
Gao et al. [98]	2014	30	Neuroscan	12	500 Hz	1	Visual Stimuli
BrainInvaders15a [91]	2015	43	g.GAMMAcap	32	512 Hz	1	Visual Stimuli
BrainInvaders15b [99]	2015	44	g.GAMMAcap	32	512 Hz	1	Visual Stimuli
Mouček et al. [100]	2017	250	BrainVision	3	n.a.	1	Visual Stimuli
Hubner et al. [23]	2017	13	BrainAmp DC, Brain Products	31	1000 Hz	1	Visual Stimuli Auditory Stimuli
Sosulski and Tangermann [101]	2019	13	BrainAmp, EasyCap	31	1000 Hz	1	Visual Stimuli
Lee et al. [102]	2019	54	BrainAmp	62	1000 Hz	2	Visual Stimuli
Simões et al. [22]	2020	15	g.tec	8	250 Hz	7	Visual Stimuli
Goncharenko et al. [103]	2020	60	NVX-52	8	500 Hz	1	Visual Stimuli
Chatroudi et al. [104]	2021	24	g.tec	64	1200 Hz	1	Visual Stimuli
Cattan et al. [105]	2021	21	g.USBamp, g.tec	16	512 Hz	1	Visual Stimuli
ERPCORE: P300 [85]	2021	40	Biosemi	30	1024 Hz	1	Visual Stimuli
Won et al. [106]	2022	55	Biosemi	32	512 Hz	1	Visual Stimuli

4.1.1 Overview of the selected Datasets

This section provides an overview of the datasets incorporated into our study. All the datasets mentioned below were carefully selected following a comprehensive analysis, ensuring they meet all the criteria for dataset selection.

1. **BrainInvaders15a** [91]

   The EEG recordings in this dataset were made while 50 participants (36 males, 14 females) with a mean (standard deviation) age of 23.55 (3.13) were playing the Brain Invaders visual P300 BCI video game. The user interface employs a unique paradigm on a grid of 36 symbols, with one symbol designated as the target and the remaining 35 as non-targets. These symbols are presented in a pseudo-randomized fashion to elicit the P300 response. In Figure 4.1, the interface of Brain Invaders is depicted during the initial level, explicitly capturing the instance when a cluster of six non-Target symbols briefly flashed in white. The red symbol represents the Target. The non-illuminated objects not exhibiting a flashing behavior are depicted in grey. In the study, participants played Brain Invaders for three sessions, each with nine levels and varying flash durations. Nevertheless, there was an absence of a substantial hiatus between each session. Hence, the three-game rounds

Table 4.2: Publicly available ERP datasets based on N400 (Semantic Priming) paradigm

Dataset	Year	#Subjects	EEG Device	#Channels	Sampling Rate	#Sessions	EEG task
Pijnacker et al. [107]	2017	45	actiCap	32	500 Hz	1	Auditory Stimuli
Draschkow et al. [108]	2018	40	BrainAmp, actiChamp	64	1000 Hz	1	Visual Stimuli
Marzecová et al. [109]	2018	18	BrainAmp	59	500 Hz	1	Visual Stimuli
Mantegna et al. [93]	2019	31	BrainAmp, EasyCap	65	512 Hz	1	Auditory Stimuli
ERPCORE: N400 [85]	2021	40	Biosemi	30	1024 Hz	1	Visual Stimuli
Hodapp and Rabovky [110]	2021	33	BrainAmp	64	1000 Hz	1	Visual Stimuli
Rabs et al. [111]	2022	38	BrainVision	26	500 Hz	1	Visual Stimuli
Schoknecht et al. [112]	2022	38	ActiCap, ActiChamp	58	500 Hz	1	Visual Stimuli
Toffolo et al. [25]	2022	24	Biosemi	128	512 Hz	1	Auditory Stimuli
Lindborg et al. [113]	2022	40	BrainVision	64	2046 Hz	1	Visual Stimuli
Stone et al. [114]	2023	64	TMSi Refa	32	512 Hz	1	Visual Stimuli

were regarded as a unified session. Three flash durations (50 ms, 80 ms, and 110 ms) were employed to record EEG data using 32 active wet electrodes.

2. **COGBCI: Flanker** [92]: It was relatively straightforward to acquire single-session datasets; however, finding appropriate multi-session datasets proved much more difficult. Despite encountering some multi-session datasets, only a few satisfy the stringent policies set for our study to be used for benchmarking. Section 4.1 covers the factors that guided us to choose our datasets for the analysis in great detail. After analyzing a handful of multi-session datasets, we narrowed our selection to one particular dataset, which offered three EEG recording sessions, i.e., COG-BCI. The COG-BCI dataset described in this study consists of recordings from 29 participants who completed three separate sessions, each conducted at an interval of 7 days. Each session included four distinct tasks: the Psychomotor Vigilance Task (PVT) [115], the N-Back Task [116], the Multi-Attribute Task Battery (MATB) Task [117], and the Flanker Task [118]. These tasks were specifically designed to elicit various cognitive states. The authors employed a 64-electrode Ag-AgCl ActiCap (Brain Products Gmbh) EEG system with an ActiCHamp (Brain Products Gmbh) amplifier placed following the extended 10-20 system.

   Due to its similarity to ERP paradigms, the Flanker task was selected as the optimal choice for our investigation out of all four tasks. The task induces interference and conflict effects, similar to the N400 paradigm, by presenting stimuli with congruent and incongruent flankers. As our study concentrates on ERP analysis, the flanker task provides a relevant framework for investigating cognitive control and neural responses using ERPs. The Flanker task is a choice reaction task derived from the study of Eriksen and Eriksen (1974) [118] and is designed to induce conflict while making a binary choice. The participants are exposed to stimuli consisting of five arrows positioned at the center of a computer screen. Participants are instructed to respond to the central arrow while disregarding the surrounding (flanker) arrows. These flanker stimuli can aim in the same direction as the

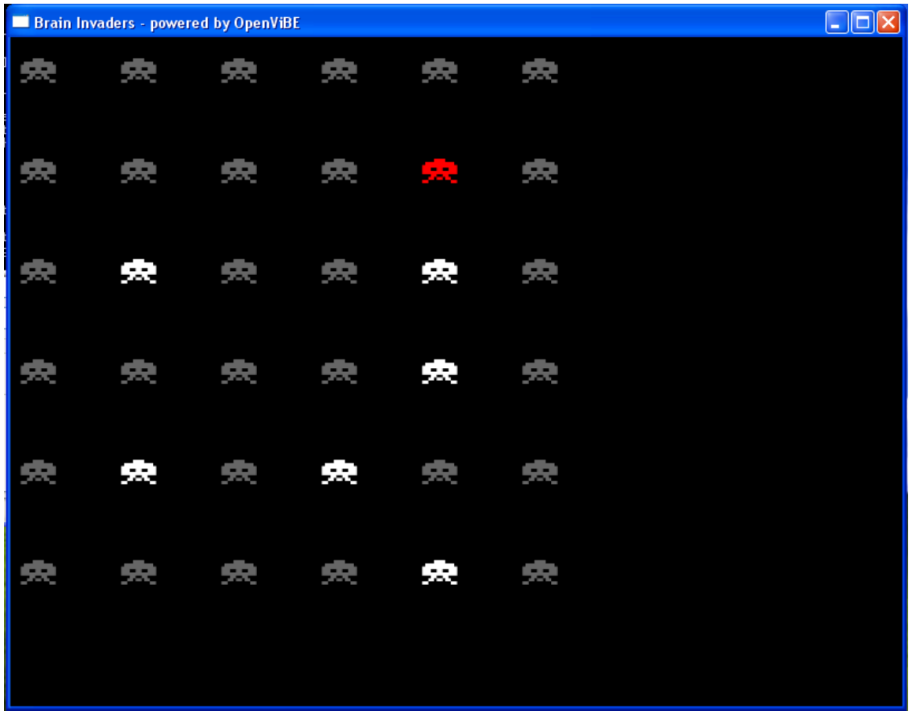


Figure 4.1: Brain Invaders user interface at the game's introductory stage [84].

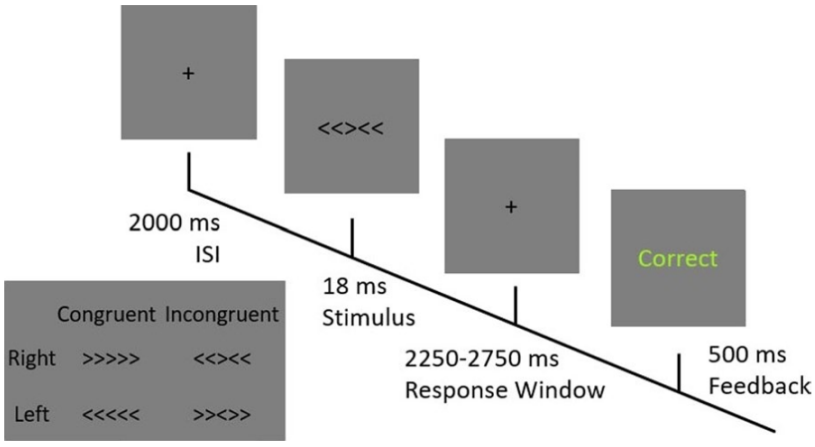


Figure 4.2: Flanker Task: After an Inter stimulus of 2000 ms, one of four possible stimuli (bottom left) is displayed for 18 ms. Participants then have between 2250 and 2750 milliseconds to respond before receiving 500 milliseconds of feedback [92].

Figure 4.3: The experimental setup for the ERPCORE: N400 task involving a specific configuration designed to elicit and measure the N400 component [85].

central target (congruent condition) or in the opposite direction (incongruent condition). Figure 4.2 illustrates the flanker task's experimental procedure. Upon the conclusion of the trial, the participant is provided with feedback regarding the outcome of their performance, explicitly indicating whether their response was correct, incorrect, or a miss. A total of 120 trials are conducted, with each complete run having an approximate duration of 10 minutes.

3. **ERPCORE: N400** [85]: This dataset has been used in various brainwave-based recognition studies such as [20, 26, 119]. It was developed for seven often studied ERP components: N170, MMN, N2pc, N400, P3, lateralized readiness potential (LRP), and ERN. The study included 40 participants, consisting of 25 females and 15 males. The participants were selected from the University of California, Davis community. The mean age of the participants was 21.5 years, with a standard deviation of 2.87. The age range of the participants was between 18 and 30 years. For our study, we focused on the N400 task. A word pair judgment task was employed to elicit the N400 component in this task. Every experimental trial comprised a red prime word that was subsequently followed by a green target word. Participants were required to indicate whether the target word was semantically related or unrelated to the prime word. The experimental setup for ERPCORE: N400 is depicted in Figure 4.3.

4. **Mantegna et al. (mantegna)** [93]: The dataset utilized in this study is derived from EEG investigations, explicitly focusing on the analysis of N400 target word modulations. The researchers of this study examined the potential for disentangling integration and prediction in the modulation of ERPs N400 during language processing. To do this, they used a stimulus assignment to complete sentences with rhyming words in various contexts with varying degrees of word predictability. All individuals who took part in the experiment were native speakers of the Dutch language, as the investigation was carried out in Dutch. In this experimental study, participants were provided with rhyming sentence completions. This experiment was carried out in three distinct stages. The first two stages consist of conducting online experiments with thirty and, respectively, 44 individuals. The third and ultimate stage of the study entails conducting an EEG experiment involving 31 participants. This experiment involves participants listening to 135 rhyming sentences with either congruent or incongruent endings. The primary objective of this experiment is to elicit N400 ERPs. Figure 4.4 illustrates an instance of a sentence pair.

4.1.2 Datasets Excluded from the Final Study

The subsequent points outline and provide descriptions of datasets that could have been considered for our study but were ultimately excluded because they did not meet certain inclusion criteria we had established.

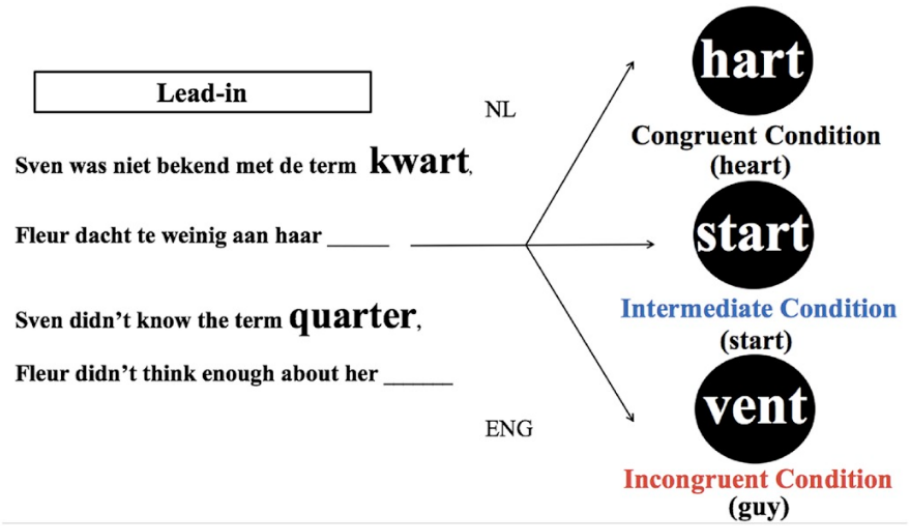


Figure 4.4: Three alternative target words were selected for each sentence pair. In the congruent case, there was overlap in the rhymes, and the target word was easy to guess based on its meaning. In the middle case, there were words that rhymed with the target word, but the target word was not predictable based on its meaning. There was no rhyme overlap in the incongruent case [93].

- **Mouček et al.** [100]: This dataset was made available for public use in 2017. The EEG experiments were conducted in primary and secondary schools across the Czech Republic, involving approximately 250 students (aged 7 to 17). The study aimed to elicit P300 by asking the participant to select a number between 1 and 9. The subject is presented with corresponding visual stimuli while experimenters observe online event-related potential waveforms and attempt to predict the number being considered.

  This dataset has, by far, the most participants, i.e., 250, and also fulfills all the conditions we set for the dataset inclusion in our study. However, the issue resides in the methodology employed during the execution of the experiment. According to our dataset analysis, each subject has a variable number of brain samples. Each subject's EEG experiment was terminated when the experimenter accurately guessed the number being tested. Consequently, the number of brain samples for certain participants is meager because the experimenter was able to correctly predict the number after observing the P300 waveforms of the subject for a short period. Conversely, the experimenter could not accurately guess the correct number for other subjects even after three attempts, resulting in more samples being observed for such subjects. We believed such an unbalanced dataset could be susceptible to bias and overfitting, so we chose not to include it in our study.

- **Hubner et al.** [23]: As shown in Table 4.1, this dataset was generated at a sampling rate of 1000 Hz using the EEG amplifier BrainAmp DC. The EEG experiment involved the visual representation of German sentence *"Franzy jagt im komplett verwahrlosten Taxi quer durch Freiburg"* three times, and the participants were asked to spell it. The pool of participants in this dataset was a meager 13, which led to its exclusion from our study.

- **Sosulski and Tangermann** [101]: The dataset was generated utilizing the P300 (auditory oddball) paradigm, in which participants were instructed to focus their attention

on infrequent high-pitched target tones while disregarding frequent low-pitched non-target tones. Similar to the study by Hubner et al., this dataset contains only 13 subjects, whereas our selection criteria for datasets require a participant cohort of at least 30 subjects.

- **Draschkow et al.** [108]: The purpose of generating this dataset was to elicit N400 effects, and the EEG experiment in this study was carried out on a sample of forty participants. Participants were exposed to semantic inconsistencies, wherein an object exhibited incongruity with the intended meaning of a given scene. The dataset was initially deemed suitable for our study and was included. However, we encountered an issue while working on this dataset. Our framework is designed to scrap EEG data from the internet directly. Unfortunately, during the retrieval process from the data repository, we encountered an error indicating an issue with the file's integrity. Despite implementing numerous technical alternatives, our attempts to resolve this issue have proven unsuccessful. As a result, we were unfortunately obliged to omit this dataset from our study, as it remained inaccessible for subsequent analysis and processing.

- **Hodapp and Rabovky** [111]: This research presented 120 pairs of German sentences to 33 participants. The sentence pairs were intentionally constructed so the ultimate target word in each pair could exhibit either semantic congruence or incongruence. The EEG experiment aimed to induce N400 effects in the participants. Nevertheless, the publicly available data provided by the researcher has already undergone pre-processing. As indicated in section 4.1, the lack of access to raw data poses a challenge to implementing standardized pre-processing, feature extraction, and authentication techniques on the datasets. Consequently, we opted to exclude this dataset from our research analysis.

- **Simões et al.** [22]: The dataset used in this study comprises 15 autistic persons who were subjected to a total of 7 training sessions. During the EEG experiment, stimuli were exhibited in a virtual bedroom setting using the Vizard toolbox. The participants were tasked with identifying specific things hidden among conventional furniture items. The dataset records P300 responses, offering valuable insights into the cognitive processes of individuals with autism. This dataset would have been appropriate for investigating the issue of cross-session variability across subjects in our study. Regrettably, the sample size for participants was restricted to 15 subjects, which limited the inclusion of this dataset in our study.

- **Huang et al.** [86]: The dataset in question has been previously discussed in section 3.3, where it was noted that it offers a highly comprehensive analysis of cross-session evaluation. We decided to incorporate this dataset into our research and test whether or not we could replicate the results. However, it came to our attention that the researchers responsible for this dataset have solely made available the pre-processed data, omitting the raw data. Consequently, we were compelled to exclude this dataset from our research.

Once the datasets have been collected, the next step is to create an outline of the benchmarking, which is covered in the following section.

4.2 Workflow

Once the necessary EEG datasets are acquired, the subsequent step establishes a workflow that presents an abstract view of our envisioned benchmarking tool. This benchmarking framework is organized into five integral components: datasets, paradigm, evaluation, pipeline, and analysis, as illustrated in Figure 5.1. We have drawn inspiration for our benchmarking workflow from the

MOABB (Mother of all BCI benchmarks) [21] work. In their research, Jayaram and Barachant investigated widely used BCI algorithms using 22 publicly available datasets involving over 250 participants. However, their study did not encompass authentication algorithms. Thus, we have modified their approach to construct a benchmarking suite tailored for brainwave authentication systems. The following section provides an overview of all of the modules described depicted in Figure 5.1. We also provide statistical and visualization tools to help visualize the performance of authentication techniques.

- **Datasets**: This module offers abstract access to open datasets. It entails downloading open datasets from the internet and providing effective data management.
- **Paradigm**: The purpose of this module is to conduct pre-processing on the unprocessed EEG data. Datasets exhibit distinct characteristics based on ERP paradigms such as P300 and N400. Nevertheless, both conditions elicit ERP responses after the individual's exposure to unexpected stimuli. Consequently, the datasets for the P300 and N400 paradigms undergo pre-processing using identical parameters.
- **Pipeline**: This module extracts features from data that has been pre-processed. These characteristics are extracted in the time and frequency domains and are discussed in detail in section 5.3.
- **Evaluation**: The authentication algorithms are developed and utilized for training and testing the features extracted within the pipeline module. The performance of authentication modules is assessed through various evaluation schemes, including within-session and cross-session evaluation. In addition, we will evaluate the efficacy of authentication protocols across multiple threat scenarios, including both closed-set and open-set scenarios.
- **Analysis**: After obtaining the performance metrics, this module offers various methods for conducting statistical analysis on the performance of diverse datasets and algorithms. The analysis will be conducted utilizing multiple visualization techniques.

It is essential to acknowledge that the execution of the procedures above necessitates the utilization of the Scikit-Learn [120] pipeline. This pipeline facilitates the execution of various Python pipelines comprising distinct datasets, paradigms, feature extraction methods, and algorithms. The upcoming chapter 5, will provide comprehensive insights into implementing our benchmarking tool.

5 Benchmarking Tool Implementation

It is imperative to establish a standardized pipeline encompassing the entire process, from pre-processing the data to extracting relevant features and validating the algorithm's performance to promote the comparability and reproducibility of brainwave authentication algorithms [21]. By adopting a common pipeline framework, researchers can ensure consistency and facilitate the evaluation of different brainwave authentication algorithms. It also enables the researchers to spend more time on algorithm design and evaluation rather than doing repetitive and error-prone tasks. The standard pipeline will be implemented as a wrapper around the scikit-learn [120] pipeline library, which provides various tools for programming machine-learning models. Moreover, using the scikit-learn library to construct standardized models guarantees credibility, as the pipeline offered by scikit-learn is widely trusted within the machine learning community. Adopting a standardized benchmarking framework will contribute to advancing brainwave authentication techniques, facilitate collaboration, and expedite progress in the field.

5.1 Loading Datasets

The datasets, as discussed in section 4.1.1, offer a comprehensive and varied collection of data points, exhibiting notable variations in ERP paradigms, sample size, and subject sessions, are crucial for our study. The wide range of datasets available presents a compelling prospect for conducting comprehensive analysis and exploration. However, the heterogeneous nature of the datasets offers difficulty in their utilization and data management, mainly when performing various analyses and evaluating new algorithms. A Python interface is developed to overcome these obstacles and enhance the efficiency of accessing the datasets. This interface aims to optimize and improve the process for accessing and managing these datasets. The interface utilizes the MNE Python package's capabilities, a comprehensive and versatile software package specifically developed for various tasks such as data preprocessing, source localization, statistical analysis, and functional connectivity estimation among spatially distributed brain regions [121]. The Python interface employs the MNE package to access and arrange public datasets into a hierarchical structure consisting of subjects, sessions, and discrete recordings within each session [21]. The hierarchical structure of data facilitates efficient data management, enhancing the ability to navigate and retrieve specific data as required.

Once the datasets are loaded locally, the raw EEG data are transformed into a standard MNE data format. Standardizing the unprocessed EEG data into raw MNE data is crucial as it is the foundation for all subsequent steps like pre-processing, feature extraction, and evaluation.

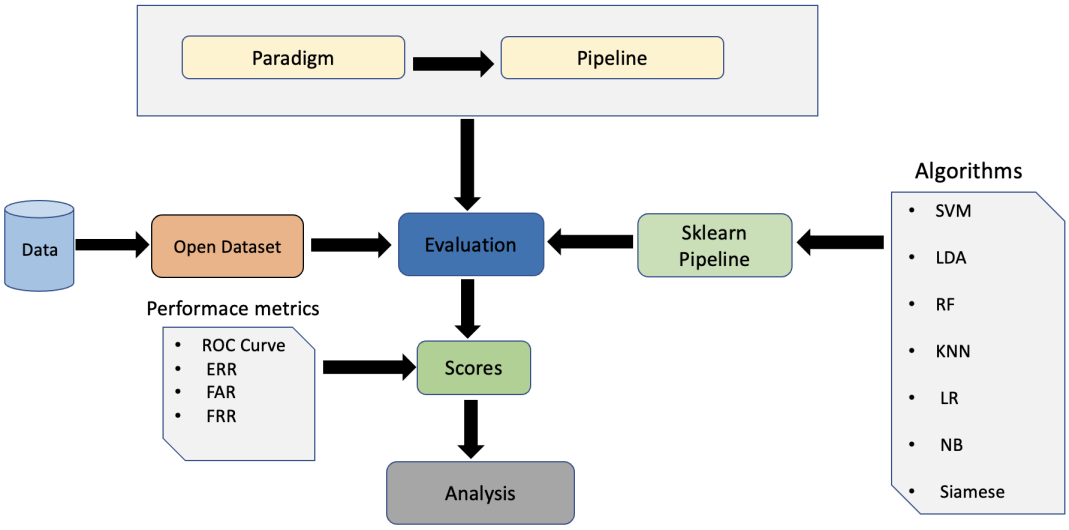


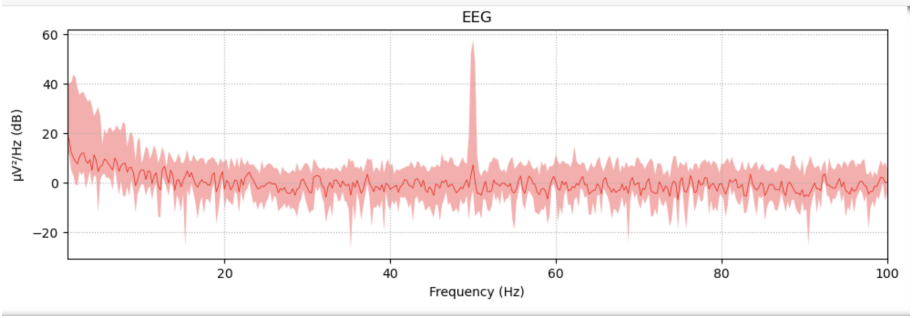
Figure 5.1: Overview of benchmarking suite [21]

While converting unprocessed brain data into standardized MNE data, the following actions were followed to ensure consistency across the datasets and to incorporate all pertinent brain samples into the unprocessed MNE data.

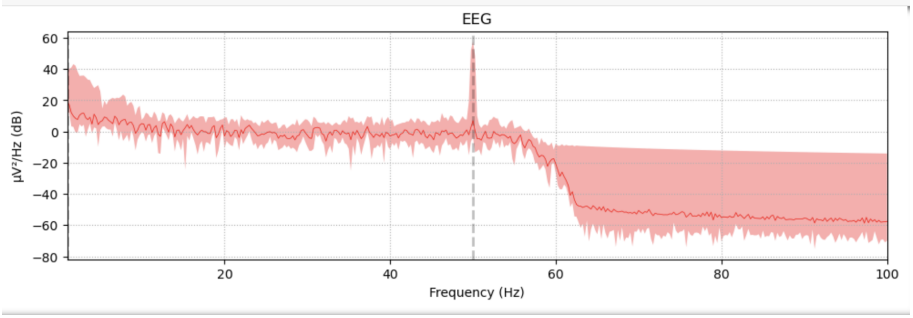
- EEG data can be quantified in micro voltage or on a voltage scale. The choice of measuring scale is contingent upon the specific EEG devices researchers employ. Upon analyzing our chosen datasets, it was observed that ERPCORE: N400, Mantegna, and COGBCI: Flanker exhibited congruity in their measurement scale. However, the BrainInvaders15a dataset was originally measured on a microvoltage scale. To ensure consistent data scalability across all four datasets, we rescaled the EEG data of BrainInvaders15a.

- Following the information presented in section 4.1.1, it has been established that the Mantegna dataset consists of three distinct categories of events, namely congruent, intermediate, and incongruent. According to the research conducted by Mantegna et al., [93], it was observed that both intermediate and incongruent stimuli evoke the N400 effect. Based on the observation mentioned earlier, we opted to merge the intermediate and incongruent stimuli into a unified category, denoted as 'incongruent' within the context of our research. This strategy was implemented to bring attention to individual differences in the EEG induced by these stimuli, specifically the N400 ERPs.

- Researchers commonly use the button press method to record time-locked responses to stimuli to guarantee participants focus on EEG activities and the reliability of recorded brain responses. The conventional approach entails utilizing online processing, wherein researchers selectively retain events that elicit accurate responses while disregarding those that indicate a lack of attention. This methodology effectively excludes brain responses that may be random and do not accurately represent ERPs. The same online processing method was observed in the BrainInvaders15a, Mantegna, and ERPCORE: N400 datasets in our study. However, the dataset provided by the COGBCI: Flanker did not adhere to this particular practice, which necessitated the implementation of offline processing. In this instance, we kept both the congruent and the incongruent time events accompanied by accurate participant feedback, increasing the dataset's utility for ERP research.



(a) Line noise at 50 Hz due to electrical appliances around the EEG devices



(b) Filtered brain signal after applying finite band pass filtering

Figure 5.2: Power Spectral Density of the brain signal before (a) and after (b) applying filtering

5.2 Pre-Processing

After the datasets have been loaded, it is necessary to establish the pre-processing procedures for EEG data. Various methods exist for cleansing artifacts; however, the procedures must remain consistent to ensure the validity of comparisons between algorithms or datasets [21]. We have adhered to established best practices commonly employed in pre-processing methodologies within brainwave authentication studies [27]. The first stage of EEG data cleaning involves the elimination of line noise originating from electronic devices present within the experimental environment during the EEG recording. The application of finite bandpass filtering in the 1 to 50 Hz range is employed for this purpose. The selected range was determined based on eliminating the 50 Hz line noise and filtering out signals originating from flat channels with frequencies below 1 Hz. Figure 5.2 (a) depicts the unprocessed raw signal, which exhibits a significant signal strength at 50 Hz due to line noise. On the other hand, Figure 5.2 (b) illustrates the consequences of implementing a bandpass filter, revealing a noticeable stabilization in the raw signals after the data filtration.

The subsequent procedure involves the extraction of epochs from the raw signals. The data is temporally aligned to a range spanning from -200 to 800 ms relative to the onset of the stimulus. Baseline correction was applied to each epoch by subtracting the mean baseline period, which ranged from -200 to 0 ms. Baseline correction is employed to reduce the drifting effects of DC offsets [26]. Much noise from higher frequencies, such as power lines or very low frequencies from flat channels, is removed during filtering. Nevertheless, the epoch data would still contain significant artifacts caused by eye or muscle movements that need to be isolated. In practice, it is common to employ thresholds approximately equal to 100μV or 150μV to eliminate these artifacts effectively [27]. However, this method also results in the loss of a significant amount of valuable EEG data. As illustrated in Figure 5.3, thresholds of 100μV and 150μV resulted in the exclusion of over 80% of the total EEG data for datasets such as BrainInvaders15a and

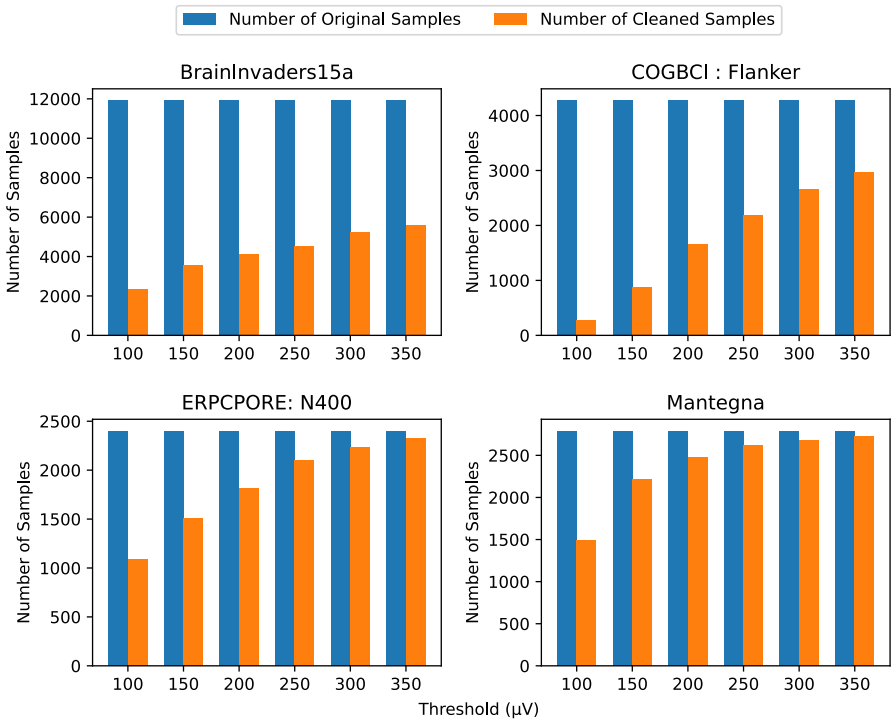


Figure 5.3: Actual number of epochs versus the number of cleaned epochs after conducting epoch rejection with various thresholds. x-axis depicts threshold values in μV whereas y-shows the the sample count

COGBCI: Flanker. Consequently, we sought to identify an alternative approach to effectively eliminate noisy data while minimizing the loss of valuable EEG data.

We implemented a more sophisticated approach to eliminate noisy data by utilizing the *Autoreject* [122] package. This Python package was developed by the original developers of the MNE package, but it has not yet been incorporated into MNE. The Autoreject method addresses the issue of manually determining a threshold by implementing cross-validation on the epochs, allowing for the learning of an optimal rejection threshold specific to each channel. It removes epochs with greater precision and partially repairs them through interpolation techniques. While this method saves a substantial amount of data and corrects noisy trials, we observed that its strategy of performing cross-validation on all user samples could result in data leakage. This prompted us to reevaluate the optimal threshold for rejecting artifacts. We could not employ low threshold values, such as 100μV and 150μV, nor use Autoreject.

Consequently, a decision was made to raise the threshold for artifact rejection to 250μV. A threshold of 250μV does not represent an extreme threshold for rejecting artifacts, as it falls within a moderate range. The selected value is also based on the consideration that setting a threshold higher than 250μV would result in the retention of numerous noisy samples in our pre-processed data. Consequently, the subsequent stages, such as classification, would have yielded random predictions due to including random noisy samples. Hence, the implementation of epoch rejection using a peak-to-peak threshold of 250μV was applied in our study. After performing all the aforementioned pre-processing steps, we averaged the Target(unusual stimuli) and Non-Target(standard stimuli) epochs to check if the ERP signal had been correctly segregated and that the applied pre-processing had successfully minimized other non-task-related brain responses. The visual representation shown in Figure 5.4 depicts the mean evoked po-

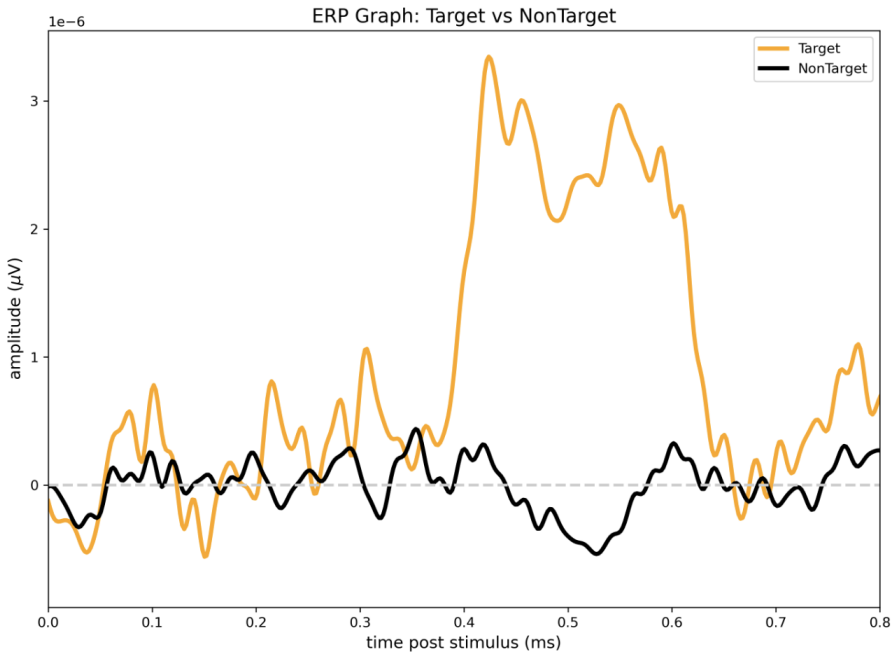


Figure 5.4: The Averaged Evoked Potentials exhibit an increase in amplitude ranging from 250 to 400 milliseconds, which can be attributed to the implementation of the oddball paradigm.

tentials seen in the epochs of dataset BrainInvaders15a. The pre-processing steps described above resulted in a total of 4539, 2193, 2097, and 2618 cleaned epochs for the datasets BrainInvaders15a, COG-BCI: Flanker, ERPCORE:N400, and Mantegna, respectively. These epochs are subsequently employed for feature extraction and the classification process.

5.3 Feature-Extraction

Following the pre-processing of the EEG data, the subsequent stage involves obtaining discriminant characteristics that effectively capture and encode the mental activity of a user, utilizing the refined EEG signal [27]. We surveyed many studies presented for brainwave authentication. We found that the Autoregressive (AR) model and Power spectral Density (PSD) are some of the most widely used methods for extracting features in time and frequency domains [20, 12]. Further, AR's potential to reveal particular inherent characteristics of the EEG signal within a single channel and PSD's ability to extract and distinguish the dominant frequency components [27] make them a promising candidate for our study to extract subject-specific information from the EEG data. Our research's feature extraction procedure, which uses the abovementioned techniques, is outlined below.

- **AR Coefficients**: The AR model is fitted using pre-processed epochs, time series data lasting for 1-second [20]. The coefficients obtained from this procedure are subsequently considered as features. The estimation of AR coefficients can be accomplished by utilizing the Yule-Walker method. The Yule-Walker method is a computational approach that employs a pth-order AR model to analyze a signal subjected to windowing. This is accomplished by minimizing the least square error of forward prediction and directly solving for the AR parameters [46]. Identifying the optimal order for AR modeling is an intricate task since high orders increase the computational cost and very low order does

not represent the signal properly [45]. As a result, We extracted AR features in various orders, such as 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10. However, the most optimum order appears to be 6, and we have set the default order to 6 in our framework. Acknowledging that the AR coefficients are computed individually for each channel within the signal is essential. Therefore, the total number of AR coefficients calculated is proportional to the number of channels utilized for feature extraction. For instance, in the scenario where brain data is analyzed through the utilization of 32 channels, and all of these channels are employed for the feature extraction process, it can be observed that the application of a 6th-order AR model will yield a total of 196 features (32 channels multiplied by an order of 6).

- **PSD**: The PSD of each epoch is computed across different frequency bands, namely low (1-10 Hz), α (10-13 Hz), β (13-30 Hz), and γ (30-50 Hz), utilizing the Welch periodogram algorithm [20]. Welch's periodogram is used to compute the Discrete Fourier Transform (DFT) results [27]. In our study, the PSD for each frequency point in a 1-second epoch is first calculated. Furthermore, in calculating PSD using Welch's algorithms, we utilized four-time windows of equal size on a 1-second ERP epoch, with 50% overlap between each window. Including the time window factor was necessary to separate the genuine frequency modulation of the EEG caused by attention from any artifacts that the attentional modulation of ERPs may have induced [123]. We then computed the average PSD within the specified frequency ranges. This allowed us to determine the average power spectrum of the low, alpha, beta, and gamma frequency bands. Similar to the AR features, the PSD features are likewise computed for each channel.

5.4 Classification

Most brainwave authentication techniques fall under the categories of similarity-based or supervised learning-based recognition systems [27]. In our study, we have employed both learning methods for authentication. Additionally, classification is performed under two evaluation strategies: within-session and cross-session. We undertake a comparative analysis and examination of the suitability of two evaluation schemes and authentication methodologies for various classifiers within two threat case scenarios. The subsequent sections delineate the methods for conducting authentication within the context of similarity or supervised learning techniques.

5.4.1 Supervised based Learning Classification

Authentication is performed by comparing the user's recorded samples with the user's enrolled samples, usually stored during the registration phase, to classify whether the recorded samples match. The fundamental concept entails acquiring knowledge by utilizing a one-vs-all classification methodology employing a binary classification system with two distinct classes. Consequently, a classifier is trained for each subject to be incorporated into the system. Accordingly, a singular classifier is tasked with recognizing an individual issue [20]. Traditional classifiers like LDA, SVM, RF, NB, LR, and KNN are utilized in this study to classify the features we calculated in the feature extraction process. The classification will be performed under the within-session and cross-session evaluation schemes detailed below.

Within-Session Evaluation

Under the within-session evaluation, the training and testing of the features are done utilizing the recorded data from a single session. To avoid overfitting and increasing the reliability of our authentication system, we used RepeatedStratifiedKFold (k=4) to split the single session

data into training and testing. Stratified cross-validation was chosen because it ensures that the features from both classes are represented in the train and test data during each fold. Users with less than four samples were eliminated from the datasets to ensure adequate samples for training and testing [20]. The total number of repetitions conducted was 10, and the results of the evaluation metrics obtained for all folds and runs are averaged and reported. Additionally, we employed feature scaling to prevent overfitting by fitting the StandardScaler [1] on the training set and applying it to both the train and test sets in every iteration. In the dataset, such as COGBCI: Flanker, which has multiple sessions across subjects, the evaluation has been performed across each session. Then, the results from the three sessions have been averaged.

Threat Case Scenarios: The implementation and evaluation of an authentication system across individual sessions are conducted in the context of two attack scenarios: Close-set scenario is a standard one vs. all approach described earlier in this chapter. Under this approach, we trained unique classifiers for each user by marking all of their samples as "authenticated" and all of the samples from all other users as "rejected" [20]. The close-set approach has been extensively utilized in numerous studies. The current process lacks real-world applicability as it operates under the assumption that attackers are already part of the system, making it more straightforward for the model to distinguish genuine users. However, this is only sometimes the case, as the attacker could be an unknown user attempting to imitate a legitimate user. Hence, assessing the authentication system's efficacy in an open-set scenario is imperative. However, implementing an open set is more challenging due to the requirement of training the classifier with known users (enrolled users) and evaluating it with unknown users (attackers who are entirely known to the system). We looked, and unfortunately, no cross-validation technique exists that meets all of our requirements for training and testing in an open-set environment. Thus, we adopted a tailored cross-validation approach to test the robustness of our authentication system in an open-set scenario.

We separated dataset samples into 'authenticated' and 'attackers' groups to implement an open-set scenario. A GroupKFold strategy with a value of K=4 was utilized, wherein 'attackers' SubjectIDs were employed. The data were divided into training and testing, with 75% attackers allocated to training and the remaining 25% assigned to testing in each cross-validation. Each cross-validation iteration constructed a modified training set by randomly selecting 75% of the 'authenticated' samples and the majority (75%) of the 'attackers' samples. The testing set, on the other hand, comprised the remaining 'authenticated' and 'attackers' samples. This GroupKfold method ensured a non-overlapping distribution of 'attackers' participants between training and testing, improving our system's practical validity in real-world settings. For example, ERP-CORE: N400 dataset contains 40 individuals. In this scenario, the epochs of a user are assigned authenticated labels, with 29 being rejected and nine identified as unknown attackers per model [20]. Nine unknown attackers were absent in the training set. The model is trained and tested using a train-test split of 75% and 25%, respectively.

Cross-Session Evaluation

The collection of multi-session EEG recordings poses challenges as it becomes more difficult to ensure that all participants can replicate the experiment accurately after a designated period [86]. It is also the underlying cause of the significant scarcity of datasets. In our study, we have a single multi-session dataset out of all the open datasets, i.e., COGBCI: Flanker, which contains three recorded sessions at an interval of 7 days. Each session consisted of a total duration of 10 minutes. As a result, we performed the cross-session evaluation on this dataset. Under the cross-session evaluation strategy, two sessions containing 20 minutes of EEG recording data are used

[1] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html/

for enrollment or training the classifier. In contrast, the remaining session data is employed for testing or authentication. This approach ensures the substantial data of model training, ensuring the reliability of the resultant classifier. The performance of the model is assessed in a distinct and independent session. A cross-validation strategy is employed to avoid potential bias in session allocation. We used LeaveOneGroupOut [2] cross-validation method to group the sessions into training and testing set. The issue of potential overfitting was addressed by applying feature scaling to the training and testing set. Jayaram and Barachant [21] implemented a comparable cross-session evaluation approach in their development of MOABB (Mother of all BCI Benchmark), a benchmarking framework evaluating the performance of different BCI algorithms on open datasets. Additionally, we excluded users who did not have at least three data sessions to ensure adequate samples in both the training and testing sets. Just like within-session evaluation, classification in cross-session is also conducted using both threat case scenarios.

Threat Case Scenarios: This evaluation scheme includes both closed-set and open-set scenarios. A single classifier is trained to recognize each user in a close-set method. To meet this criterion, samples from a specific user across all three sessions are labeled "authenticated," while samples from all other users are labeled "rejected." The training set consists of data from two out of three sessions and includes authenticated and adversary samples. During testing, the remaining session data is utilized. The close-set method is comparable to the one described for within-session evaluation, except that the model is evaluated using data from multiple sessions and, therefore, more realistic. However, our study goes beyond the close-set scenario and investigates the system's efficacy when exposed to unknown attackers during authentication in a cross-session environment.

The open-set scenario follows the same method of LeaveOneGroupOut cross-validation for grouping session data into training and testing sets. But to accommodate the open-set strategy, we modify the composition of the attackers within these sets. In each round of the cross-validation, a random selection is made, based on the SubjectIDs, to include 75% of the attackers in the training set. The training process excludes the remaining 25% of attackers. In contrast, the testing set comprises solely the attackers omitted during the training phase while excluding the attackers on which the model was trained. By employing this methodology, we effectively establish a scenario wherein the model is evaluated using attackers entirely unknown to the system in a cross-session environment.

5.4.2 Similarity Based Learning

Unlike Supervised learning methods, which entail training a model with other users for decision-making [26], similarity-based techniques identify a person based on the similarity between the brain signals acquired during the enrollment phase and those presented during the verification phase. The similarity between the enrolled and tested samples is calculated using metrics like Euclidean or cosine distance. Siamese Neural Network is a highly effective deep learning (DL) method for implementing similarity metrics for brainwave authentication, employed already in studies such as [26, 87, 83]. This type of learning allows for accurate predictions after training the network with only a few samples [124], overcoming a typical drawback of employing DL approaches for brainwave authentication. It also avoids the usual shortcoming of supervised learning algorithms—retraining—while adding new users to the system, as discussed in subsection 1.3.

In the context of classification, supervised algorithms frequently require extracting discriminant features from raw epochs to enhance the classification process. In contrast, the Siamese

[2] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.LeaveOneGroupOut.html

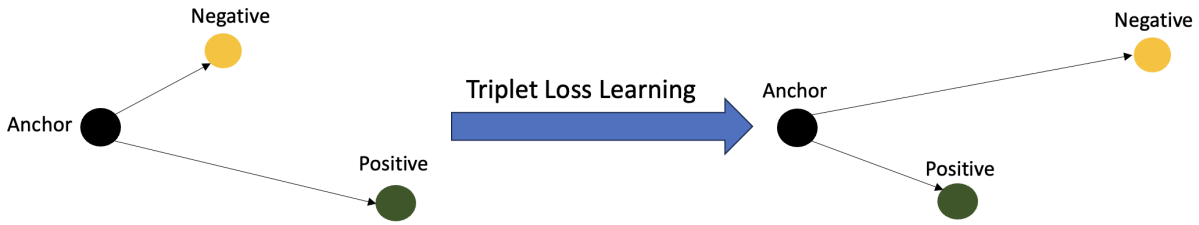


Figure 5.5: The triplet loss function minimizes the Euclidean distance between the anchor and positive embeddings while maximizing the distance between the embeddings of two individuals, precisely the anchor and negative embeddings.

Networks adopt a distinct strategy by circumventing the conventional feature extraction procedure. Instead, it generates feature embeddings directly from the time series data of epochs using the CNN method. The epochs are structured in a two-dimensional array, where the rows correspond to channel indices, and the columns represent discrete-time measurements. SNN can be trained using various loss functions such as Contrastive loss function and Triplet loss [125]. Among these, the *triplet loss* approach is particularly well-suited for biometric recognition [26]. As a result, our study utilizes Siamese Networks with three CNN branches, which are trained using a triplet loss function. As explained by Schroff *et al.* [124] in their study, learning using the triplet loss function involves the provision of three distinct types of inputs, namely an anchor, a positive sample (which shares the same identity as the anchor), and a negative sample (which possesses a different identity than the anchor). Following the completion of this procedure, the embeddings of individuals of the same identity will exhibit minimal distances, while those corresponding to distinct individuals will exhibit significant distances. As a result, once the embeddings are generated, a similarity metric (often Euclidean distance) can be used to verify or identify them. Figure 5.5 illustrates the learning procedure in Siamese Networks using the triplet loss function.

The Siamese architecture proposed by Fallahi et al. [26] was implemented in our study. Therefore, a CNN consisting of five convolution layers was utilized to develop the system. After each convolutional layer, an average pooling layer was applied to reduce the input vectors' dimensionality while preserving each brainwave's unique characteristics. Further, the FaceNet study [124] demonstrated that minimizing triplet loss through the online mining of semi-hard triplets is the most effective method for quick convergence; consequently, this method of triplet selection was also utilized in our study. Furthermore, user authentication is performed using both within-session and cross-session evaluation strategies outlined below.

Within-Session Evaluation

The within-session evaluation in Siamese Neural Network is designed to work well in both the seen attackers (close-set) and unseen attackers (open-set) scenarios. Both scenarios are implemented in a similar methodology, except the first involves comparing the identification sample with all the enrollment samples during testing. In contrast, in the open-set method, the subject's sample being tested is compared to an enrollment database that does not include the subject's specific brain sample [26]. The comparison is conducted through the computation of Euclidean distance. Below is a short overview of the evaluation strategy for both threat case cases.

Threat Case Scenarios: In close-set, the user's samples were divided into training and testing sets using stratified cross-validation with k=4. As a result, we omitted users from the datasets

with less than four samples. The Siamese model is trained on all the users of the dataset. For example, if the dataset has EEG data of 40 subjects, all 40 subjects get enrolled during the training process. Training and testing data is scaled using the standard scaler normalization method. The model learns to generate the brain embeddings, and during verification, the brain embeddings of each user are compared against the enrollment data of all subjects.

Implementing the open-set approach involves utilizing the GroupKFold cross-validation strategy, with a value of k set to 4. During each round of cross-validation, the grouping is done based on SubjectID, resulting in a non-overlapping training set consisting of 30 subjects and a testing set of 10 users if the total number of users in the dataset is 40. This approach tests the model's recognition capability against unseen attackers.

Cross-Session Evaluation

The methodology utilized for implementing cross-session evaluation in Siamese Neural Networks is comparable to the approach employed for the cross-session assessment in supervised learning-based classification tasks. Therefore, LeaveOneGroupOut cross-validation was used for grouping the sessions into training and testing sets in each round of cross-validation.

Threat Case Scenarios: A close-set scenario is attained by employing the previously discussed LeaveOneGroupOut cross-validation technique. In a close set, samples from each subject's independent session are compared to their respective enrollment records. In this instance, the enrollment database comprises the brain samples of all subjects collected during the two sessions. In the open-set scenario, the session data is partitioned into training and testing sets using the LeaveOneGroupOut method, where subjects from the enrollment database that are being verified are excluded. In this case, the enrollment database has two sessions of data, and the evaluation is based on the remaining session data.

5.4.3 Automated Benchmarking

The benchmarking framework is developed with a primary focus on ensuring user-friendliness. Our objective was to enable anyone to effectively utilize this framework, even without a comprehensive understanding of the complex technical intricacies underlying the Python programming language. Consequently, a user-friendly benchmarking script was developed, efficiently analyzing a configuration file written in a clear and concise YAML manner. This configuration file is a control panel for defining various parameters and settings. It automates all the complex tasks involved in data extraction, pre-processing, feature extraction, and classification, as illustrated in Figure 5.1. This streamlined approach eliminates the need for users to delve into intricate programming complexities. Appendix A showcases illustrative examples of such configuration files, underscoring the simplicity and accessibility of our framework's implementation.

Examples of configuration files featuring benchmarking pipelines tailored for within-session and cross-session evaluations on a single dataset are showcased in sections A.1 and A.2, respectively. The examples mentioned above effectively illustrate the flexibility and versatility of our methodology. These examples show that the pipelines can be optimized using default dataset values. Furthermore, they can be seamlessly configured to accommodate various parameter variations, spanning dataset specifics, pre-processing techniques, and algorithm selections. We will explain the significance of each parameter applied on the datasets such as *epochs interval, epochs rejection* in chapter 6.2. In chapter 6.2, we will delve into an in-depth exploration of the significance underlying each parameter employed on the dataset, including *interval,* and *epochs rejection*. This comprehensive analysis will shed light on the crucial role these parameters play in shaping the outcomes of our study.

6 Evaluation and Results

Our benchmarking tool underwent a dual-phase evaluation process. The first phase assessed the tool's functionality, focusing on its capacity to effectively address the challenges outlined in section 1.2. Subsequently, in section 6.2, we present an in-depth analysis of the evaluation and outcomes of the benchmarking tool, showcasing its performance across a range of parameters. In the second phase, the tool was employed to replicate some of the notable benchmarking works in brainwave authentication. In section 6.3, we delve into the results obtained from this replication effort, comparing them with the original outcomes presented in those brainwave authentication studies.

6.1 Evaluation Metrics

It is essential to compare the performance of the algorithms with appropriate metrics because it is seen that a lot of studies present the outcomes of their research on flawed metrics like accuracy. The accuracy of those studies is shown as high as 99%. However, it is worth noting that the sample distribution of the training and testing set is usually imbalanced since most researchers build a single classifier for individual subject. Accordingly, that single user is labeled "authenticated," the remaining users are marked "rejected" for training and testing the authentication model. As a result, the model is trained more on the negative samples. This makes it easy for the model to identify rejected users. Therefore, the high accuracy value represents a biased assessment of the model's performance because of the skewness in the training data. Hence, we choose not to focus on standard metrics like accuracy in our study. Instead, we employed performance metrics like EER, ROC-Curve as the evaluation metrics for our study. In addition, we will report FRR at 1%FAR to evaluate our authentication systems usability with enhanced security measures, given that a low FAR threshold is associated with increased security [20].

6.2 Evaluation and Outcomes of the Benchmarking Tool

Within this section, an extensive array of experiments was meticulously conducted to embark on an in-depth analysis of the performance exhibited by both SOA algorithms and advanced deep learning techniques across the four open datasets, leveraging the capabilities of our benchmarking tool. Our evaluation encompasses within-session and cross-session assessments on all datasets, spanning known attackers (close-set scenario) and unknown attackers (open-set scenario). By comparing the outcomes derived from our benchmarking tool's within-session and

cross-session evaluations, we unveil the influence of EEG variability on the authentication process. Additionally, we explore how various factors, including dataset size, epoch duration, and the utilization of AR and PSD features, contribute to the diverse performance exhibited by authentication algorithms. Consequently, we rigorously test our tool with varying AR, PSD, epoch lengths, and dataset sample sizes to comprehensively capture these nuances.

6.2.1 Experiment 1: Within-Session Evaluation across datasets

In the context of our study, each dataset featured at least one distinct session, prompting us to embark on an extensive evaluation within the confines of individual sessions across all four datasets. This pivotal phase of experimentation involved subjecting our tool to meticulous scrutiny, guided by a predetermined set of parameters below.

Datasets: BrainInvaders15a, ERPCORE: N400, Mantegna2019, COG-BCI Flanker

Utilized Parameters:

- Epoch Interval: 1 second
- Epochs Rejection threshold: 250μV
- Features: AR (order=6), PSD
- Classifiers: LDA, SVM, KNN, RF, NB, LR, Siamese
- Evaluation Type: Within-Session Evaluation
- Threat Case: Close-Set, Open-Set

Our tool underwent comprehensive testing for this experiment across various EEG datasets, specifically BrainInvaders15a, ERPCORE: N400, Mantegna2019, and COG-BCI Flanker. The data processing parameters included epoch intervals set at 1 second and an epoch rejection threshold of 250μV epochs. AR coefficients of order six and PSD were employed for feature extraction. Several classification algorithms were employed, including LDA, SVM, KNN, RF, NB, LR, and Siamese Neural Network. The evaluation used a within-session approach, focusing on close-set and open-set threat cases.

The evaluation outcomes encompass various facets, including identifying the optimal classifier for different datasets under distinct threat scenarios, namely open-set and close-set conditions. Additionally, a performance comparison across datasets in both threat scenarios is conducted. Moreover, a detailed scrutiny of classifiers' performance in close-set and open-set scenarios is carried out, coupled with a comparative analysis between SOA and deep learning methodologies. Finally, the tool is employed to assess the practical applicability of the tested algorithms.

Best Performing Classifier: The outcomes of all of the classifiers applied to the four datasets in terms of the average EER, as determined by the within-session evaluation under the close-set(seen) and open-set(unseen) attacker scenarios, are depicted in the Figure 6.1 and Table 6.1. RF classifier consistently produces the most favorable authentication results, with EER ranging between 1.3% to 4.3%. The Siamese network is the second-best classifier in terms of performance. Siamese could have been the most effective classifier because it achieves an EER of just 1% for the BrainInvaders15a, ERPCORE: N400, and Mantegna2019 datasets in close-set, which is even better than RF. However, the performance of the Siamese model exhibits degradation in an open-set strategy, with the EER reaching a significant increase of up to 14.30% for the COG-BCI Flanker dataset. The RF algorithm likewise experiences a decline

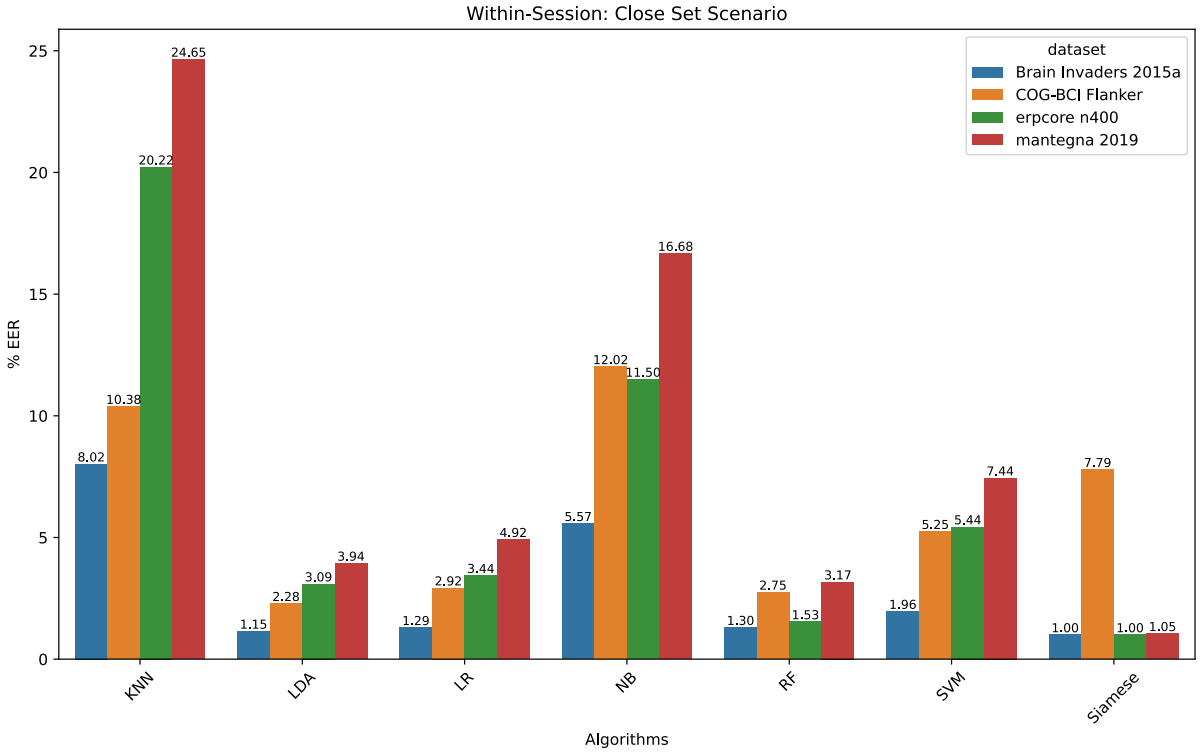
in performance when applied to open-set scenarios. However, the observed increase in the EER is comparatively lower in RF compared to the Siamese algorithm. KNN and NB are the worst performing classifiers with an EER of more than 10% in both close-set and open-set scenarios for three datasets such as ERPCORE: N400, COG-BCI Flanker and Mantegna2019. The subsequent analysis examines the performance of datasets, threat case scenarios, and the learning methodologies utilized by the authentication algorithms. Further, we explore the usability of our authentication system by comparing the performance of classifiers in terms of the calculated FRR at 1% of FAR.

Comparison between datasets: BrainInvaders15a performs better than the other datasets, as seen by the data presented in Figure 6.1. The achieved EER for BrainInvaders15a demonstrates a notable decrease across all classifiers, except the LDA classifier, when used in the open-set scenario. In this case, the EER of BrainInvaders15a is higher than that of datasets like COG-BCI Flanker and ERPCORE: N400. It is worth mentioning that BrainInvaders15a successfully attained an EER of less than 2% for many classifiers, including LDA, LR, RF, SVM and Siamese in close-set. The superior performance of BrainInvaders15a can also be ascribed to the dataset's larger sample size compared to the other datasets. As mentioned in section 5.2, BrainInvaders15a has 4539 samples as compared to 2193 samples of COG-BCI flanker, 2097 samples of ERPCORE: N400, and 2618 samples of Mantegna2019. The higher number of brain samples allows the increased availability of data, which allows for more robust training of the machine learning model [20]. In section 6.2.7, we will thoroughly examine the influence of varying brain sample sizes on the performance of classifiers. Furthermore, based on the analysis of the EER in Figure 6.1 and FRR at a FAR of 1%, as presented in Table 6.1, it can be observed that the ERPCORE: N400 dataset exhibits the second highest level of performance. This is followed by the COG-BCI flanker dataset and the Mantegna2019 dataset.

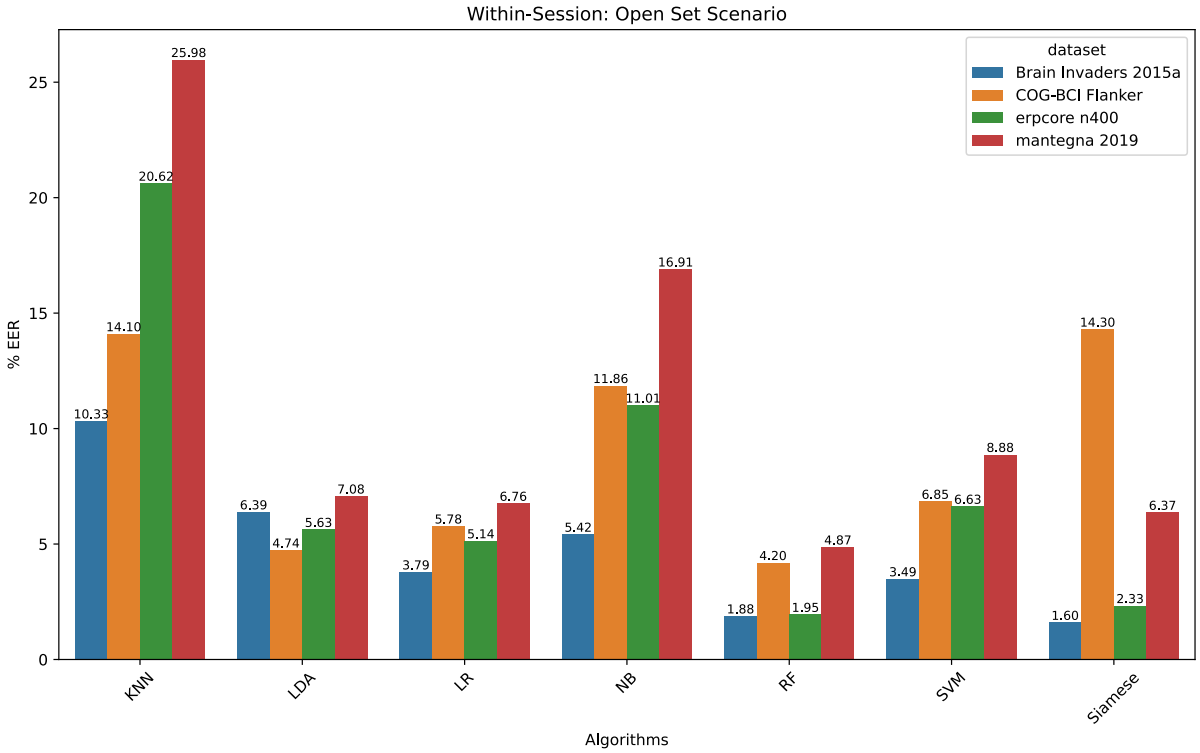
Table 6.1: Average FRR at 1% of FAR for the four datasets in a within-session evaluation scheme, comparing classifiers and threat case scenarios. The values in the table are shown in percentages.

Dataset	Scenario	LDA	SVM	LR	RF	KNN	NB	Siamese
BrainInvadeers15a	**Close-Set**	1.04	3.61	0.98	0.89	23.46	40.54	**0.05**
BrainInvaders15a	**Open-Set**	43.50	9.97	20.68	2.54	33.96	36.23	**2.07**
ERPCORE:N400	**Close-Set**	13.72	16.50	9.19	1.84	50.76	70.55	**0.21**
ERPCORE:N400	**Open-Set**	41.25	21.90	32.02	**4.56**	55.34	62.84	6.09
Mantegna2019	**Close-Set**	20.52	25.90	15.85	6.89	60.99	86.13	**0.75**
Mantegna2019	**Open-Set**	46.51	33.43	37.38	**11.34**	66.49	81.92	27.04
COG:BCI Flanker	**Close-Set**	14.05	19.19	14.05	**13.33**	32.88	59.56	45.07
COG:BCI Flanker	**Open-Set**	28.64	19.46	28.80	**13.87**	43.25	47.17	53.60

Comparison between close-set and open-set scenarios: It was hypothesized that



(a) Mean EER across all subjects in Close-Set



(b) Mean EER across all subjects in Open-set

Figure 6.1: Comparative Analysis of the four data sets performance using various classifiers and attack scenarios based on mean EER across subjects.

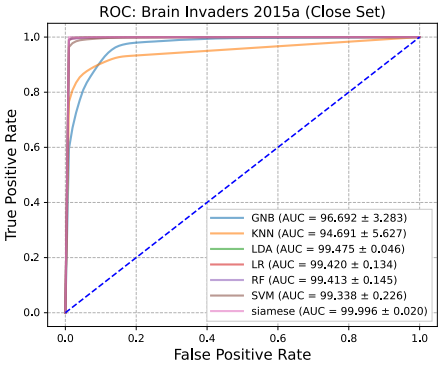
the performance of the authentication system would deteriorate when subjected to evaluation in an open-set situation. The findings from our analysis substantiated our initial concerns. As shown in Figure 6.1, there is an observed increase in EER ranging from 0.2-5.2% for most of the classifiers. A similar trend can be seen from Table 6.1 where FRR at % FAR exhibits an increase ranging from 0.6 to 43.2%. Although almost all the classifiers experience performance degradation when comparing their results in close-set and open-set settings, the most significantly impacted classifiers are LDA and LR. A notable performance decline is observed for classifier LDA where EER for dataset BrainInvaders15a increased from a mere 1.13% in closed-set to 6.3% in open-set, a significant 6-fold increase.

The outcomes of the close-set and open-set may be influenced by the differing sizes of the spaces occupied by the potential attackers in each scenario [20] as outlined in section 5.4.1, the evaluation strategy of close-set requires training the authentication model with N-1 attackers where N represents the total number of users. In contrast, the classifiers in an open-set scenario learn from approximately three-quarters of the N-1 attackers. Consequently, in close-set settings, the attacker spaces are more significant than in open-set, thereby enabling more effective training of machine learning models in close-set. Additionally, in close-set environments, the system is designed to optimize its ability to discern a particular group of pre-identified users(enrolled users). This approach produces favorable results due to the limited variability in the training set. However, in the context of open-set scenarios, the model is required to effectively address the existence of unseen users, which introduces the additional complexity of accurately identifying authorized and unauthorized users.

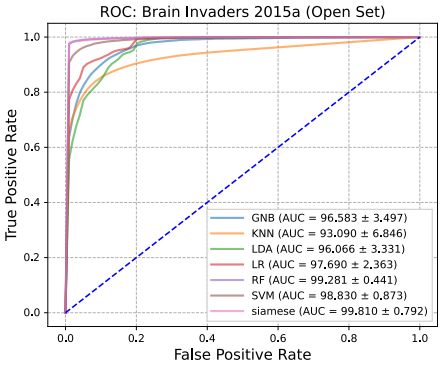
Comparison between traditional and deep learning methods: SOA machine learning algorithms such as LDA, SVM, LR, RF, KNN, and NB have always been widely utilized in EEG-based authentication systems. These algorithms provide good results when the number of classes is known and fixed. However, the performance of these algorithms tends to decline when tested with smaller data samples. They also necessitate extracting discriminant features from the raw EEG data. To address these problems, researchers started focusing on deep learning methods such as Siamese Networks, which learn directly from the time series EEG data, removing the overhead of the feature extraction process. Moreover, they do not require retraining while adding new users to the system. As a result, Siamese networks have achieved remarkable success in biometrics-based authentication studies such as face recognition [126, 124, 127] and Brainwave authentication [26, 83].

The results obtained in our study have also demonstrated Siamese networks as one of the best-performing algorithms among all the authentication algorithms. The ROC curves of the four datasets are presented in Figure 6.2, showcasing the operational capabilities of the traditional and deep learning authentication models in closed-set and open-set settings. The AUC under ROC-Curve represents a single value representing the system's ability to differentiate between genuine users and imposters [20]. A higher AUC implies an improved performance as it indicates the system has a higher TPR value and a lower FPR. The Siamese Networks have a higher AUC score than SOA classifiers in close-set and open-set scenarios across most datasets. However, it is worth noting that in the COG-BCI flanker dataset, classifiers like LDA, LR, SVM, and RF outperform the Siamese Networks regarding the AUC score for both threat cases. The findings depicted in Figure 6.1 EER Plots regarding the average EER align with the earlier observation. The results suggest that Siamese networks outperform other classifiers, consistently achieving the lowest EERs across three of the four datasets, specifically BrainInvaders15a, ERPCORE: N400, and Mantegna2019 in the close-set scenario.

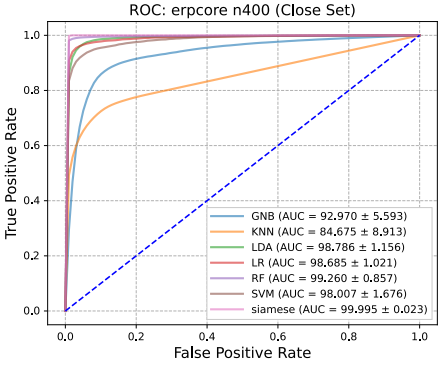
Usability: FAR is a crucial metric when assessing the overall security of the authentication system because it represents how many times the system allows an unauthorized user to authenticate. Therefore, the FAR threshold for most authentication systems is generally set low. The



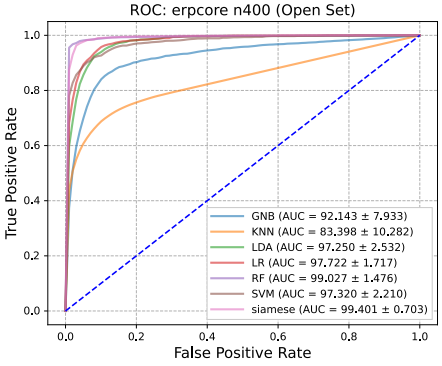
(a) ROC: BrainInvaders15a Close-Set



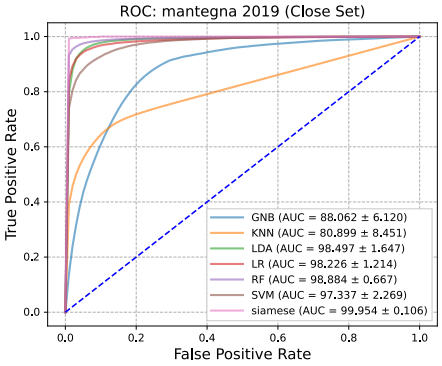
(e) ROC: BrainInvaders15a Open-Set



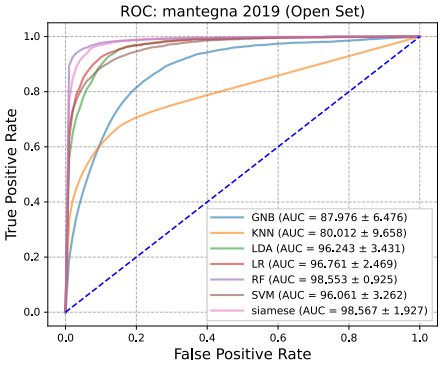
(b) ROC: ERPCORE N400 Close-Set



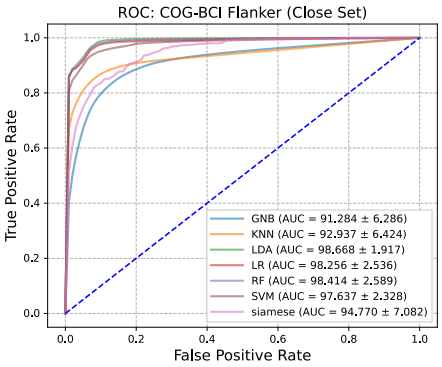
(f) ROC: ERPCORE N400 Open-Set



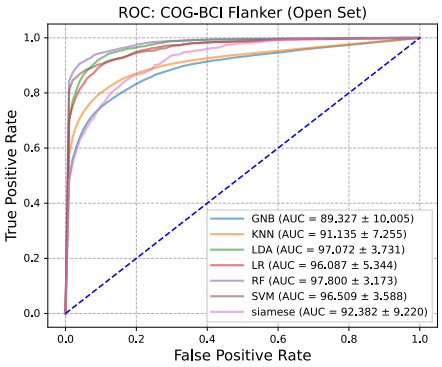
(c) ROC: Mantegna 2019 Close-Set



(g) ROC: Mantegna 2019 Open-Set



(d) ROC: COGBCI Flanker Close-Set



(h) ROC: COGBCI Flanker Open-Set

Figure 6.2: Comparative analysis of ROC-Curves for all 4 datasets in within-session evaluation

FAR's relevance spans across a spectrum, with a lower threshold of 1% for applications with lower security requirements and an even more stringent point of 0.00001% for applications necessitating the highest levels of security [128]. On the other hand, FRR quantifies the effectiveness of the authentication system in terms of its usability. A low FRR indicates that authentic users are not experiencing rejections. It is essential to strike a balance between the two metrics as each increases at the expense of the other. Consequently, we calculated FRR and FAR at 1% for each dataset in close-set and open-set. By assessing the system's performance at this particular threshold of FAR and FRR, we can gain insights into the system's efficacy in real-world situations. For example, a higher value of FRR at 1% FAR implies that genuine users are being denied access more frequently, adversely affecting the overall user experience. Conversely, a low FRR and FAR of 1% indicate that the system effectively identifies and accepts genuine users while upholding an acceptable level of security. Therefore, an efficient authentication system should always have a low FRR at 1% FAR.

The findings in Table 6.1 indicate that the most optimal setup is achieved when employing Siamese Networks on the dataset BrainInvaders15a for authentication. In the close-set scenario, Siamese Networks achieved FRR of just 0.05% at a threshold of 1% FAR. The authentication system exhibits notable usability and robustness even in open-set scenarios, as evidenced by an FRR of just 2.07% at 1% FAR. This suggests that the system remains effective even when faced with previously unseen attackers. RF has demonstrated its effectiveness as the second-highest-performing classifier in multiple instances. Specifically, it achieved the best FRR at a FAR of 1% in four different scenarios: ERPCORE: N400 (open-set), Mantegna2019 (open-set), and COG-BCI Flanker (close and open-set).

6.2.2 Experiment 2: Cross-Session Evaluation across Multi-Session Datasets

During this phase of experimentation, our tool underwent cross-evaluation using multi-session datasets, employing the predefined parameters outlined below.

Dataset: COG-BCI Flanker

Utilized Parameters:

- Epoch Interval: 1 second
- Epochs Rejection threshold: 250μV
- Features: AR (order=6), PSD
- Classifiers: LDA, SVM, KNN, RF, NB, LR, Siamese
- Evaluation Type: Cross-Session Evaluation
- Threat Case: Close-Set, Open-Set

Our tool is exclusively assessed using the COG-BCI dataset due to its unique provision of multiple EEG data sessions. The pre-processing procedures, parameters for feature extraction, the array of employed classifiers, and the considered threat cases mirror those of the initial Experiment 1 detailed in section 6.2.1. The sole distinction lies in the evaluation mode, which shifts to a cross-session assessment.

Results of Experiment 2: Figures 6.3 and 6.4 illustrate the outcomes of all classifiers applied to COG-BCI Flanker in terms of the mean EER and ROC-Curve, as confined by the cross-session evaluation. The performance of all the classifiers in both the close-set and open-set

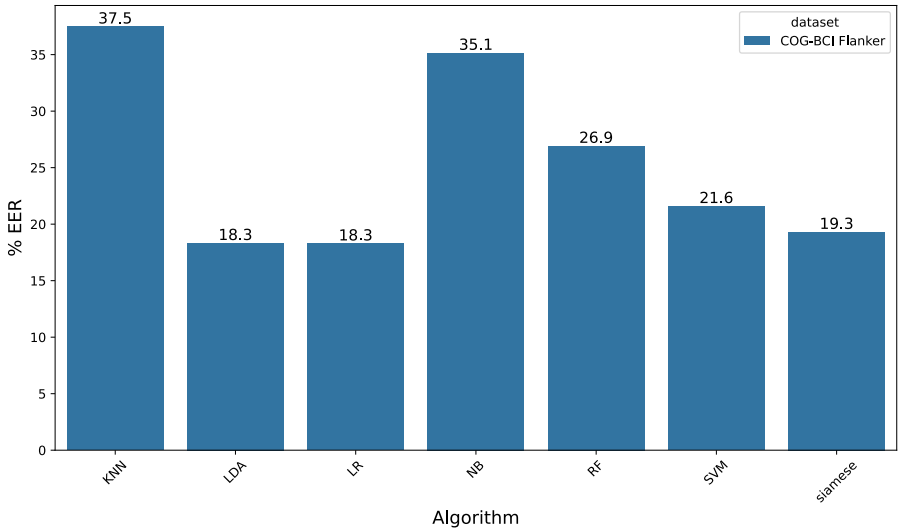


Figure 6.3: Average EER for cross-session evaluation on COG-BCI Flanker dataset, comparing performance across different authentication algorithms.

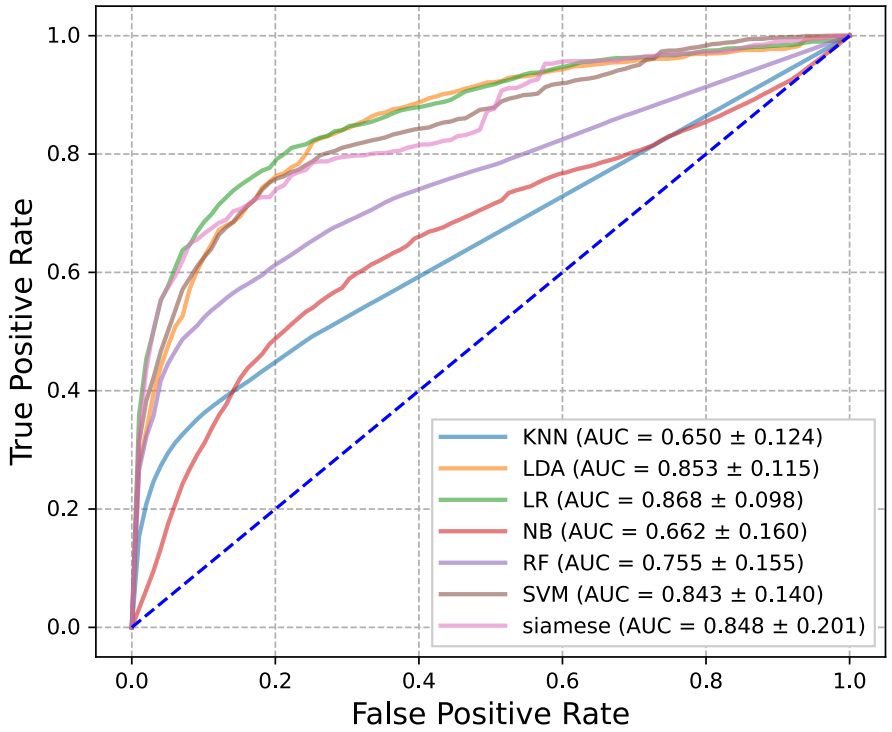


Figure 6.4: Performance comparison of the dataset COG-BCI Flanker in cross-session evaluation scheme. ROC Curves are depicted for different classifiers in close-set attacker scenario.

scenarios was observed to be comparable in cross-session. Therefore, the findings shown in both figures pertain to the close-set scenario. The attained EER for the SOA classifiers, namely LR and LDA, is identical. Both classifiers have also achieved the least EER among all the classifiers. However, LR demonstrates superior performance compared to LDA when evaluating its performance based on the AUC metric. This can be observed in 6.4, where the ROC curves indicate that LR has attained a slightly higher AUC value. The Siamese Network demonstrates the third highest classification performance, achieving an EER of 19.3%. Although RF exhibited outstanding performance in within-session evaluation, surpassing all classical classifiers and Siamese Networks, its performance in cross-session evaluation is unsatisfactory. RF is the fourth best-performing classifier, exhibiting a notable EER of 26.9%. Additionally, the results of the cross-session evaluation support the conclusions drawn from the within-session assessment, indicating that the NB and KNN classifiers demonstrate the highest EER values compared to other classifiers. This finding suggests the need for additional exploration into these two classifiers' limitations and possible enhancements.

The results from the cross-session examination of the COG-BCI Flanker dataset have yielded insights into notable discoveries. The optimal outcomes were not attained during the cross-session. In addition to the EEG variability and unpredictability resulting from the time duration between the enrollment and authentication process, several other factors may have influenced the outcomes in our cross-session settings. These factors include electrode resetting across sessions, variations in human brain states, and template ageing [87]. Moreover, the performance of SOA algorithms has been impacting more prominently in cross-session evaluation than in Siamese Networks. The results of our cross-session setting align with Arnau-González *et al.* [129] work, which utilized three publicly available datasets to investigate user identification in both single-session and multi-session scenarios. Similar to our study, Arnau-González et al. also performed feature extraction by computing PSD across θ, δ, α, β, and γ bands and utilized classifiers such as SVM, KNN, Multilayer Perceptron (MLP), and AdaBoost for building the identification models. The researchers opted to use accuracy as the performance metric in their investigation. The classifiers exhibited much-improved performance in the single session setup, with accuracy rates over 90% for various classifiers across all datasets. Nevertheless, the system's performance showed a notable decline during the evaluation conducted under a cross-session scenario. The accuracy reached in cross-session evaluation was 79%, a substantial decrease compared to the best accuracy of 99% gained in single-session evaluation. The consistent findings between our cross-session study and the work of Arnau-González et al. emphasize the need to consider temporal factors when creating authentication models for practical deployment.

6.2.3 Experiment 3: Comparative Evaluation of Within-Session and Cross-Session Approaches

One of the main objectives of this thesis is to comprehensively study the impact of EEG variability across single and multi-session settings. We utilized the capabilities of our benchmarking tool to conduct thorough evaluations of various authentication models in both within-session and cross-session scenarios, as detailed in the preceding sections. The results are presented in sections 6.2.1 and 6.2.2 respectively. These evaluations have yielded significant insights into the performance of our classifiers in different conditions and have illuminated the difficulties associated with temporal variations in EEG signals.

Our tool is employed in this experiment to evaluate the COG-BCI Flanker dataset, both within-session and cross-session. Subsequently, the outcomes of these two evaluation schemes are compared to assess the influence of single-session and multi-session setups on the performance of the authentication algorithms.

Datasets: COG-BCI Flanker

Utilized Parameters:

- Epoch Interval: 1 second
- Epochs Rejection threshold: 250μV
- Features: AR (order=6), PSD
- Classifiers: LDA, SVM, KNN, RF, NB, LR, Siamese
- Evaluation Type: Within-Session Evaluation, Cross-Session Evaluation
- Threat Case: Close-Set

Table 6.2: Average Performance of classifiers on the COG-BCI Flanker [92] dataset, comparing Within-Session and Cross-Session Evaluation

Metric	LDA	SVM	LR	RF	KNN	NB	Siamese
Within-Session							
%EER	2.28 ± 2.55	5.25 ± 3.50	2.92 ± 3.28	2.75 ± 3.28	10.38 ± 6.92	12.02 ± 7.24	7.79 ± 6.54
FRR at 1% FAR	14.46	19.19	14.05	13.33	32.88	59.56	45.07
Cross-Session							
%EER	18.32 ± 11.47	21.58 ± 13.18	18.28 ± 11.47	26.91 ± 13.67	37.47 ± 10.34	35.15 ± 13.80	19.30 ± 17.49
FRR at 1% FAR	72.37	68.74	64.15	73.43	84.66	96.83	67.53

Results of Experiment 3: According to Table 6.2, the results of the multi-session (cross-session) evaluation are significantly poorer than the single-session (within-session) evaluation for dataset COG-BCI Flanker. A significant decrease in the performance of RF can be observed, which was identified as the overall most efficient classifier across all datasets for within-session evaluation, as discussed in section 6.2.1. The cross-session EER experiences a substantial increase of 878.5% (from 2.75% to 26.91%), and FRR at 1% FAR raises to 450.8% (from 13.33% to 73.43%). LR and LDA have comparable EER and FRR at 1% FAR in within-session and cross-session schemes. These SOA algorithms experience performance degradation as EER increases from 2.28% to 18.32% for LDA and 2.92% to 18.28% for LR. Furthermore, the Siamese Networks likewise exhibit an increased trend in EER. However, the rise in the EER was only 147.7%(from 7.79% to 19.30%), indicating that the observed decline in performance was less pronounced in Siamese Networks than in SOA classifiers.

The Siamese Networks, a deep learning technique, exhibited a higher level of resilience in both within-session and cross-session evaluations, which is a favorable finding. Although the Siamese Networks also showed an elevation in EER, the magnitude of this increase was noticeably less significant compared to SOA classifiers. As mentioned above, the observation implies that Siamese Networks can learn underlying feature representations and capture similarities among EEG samples, hence exhibiting enhanced resilience to temporal variations in EEG signals. The efficacy of Siamese Networks in addressing cross-session evaluations underscores the potential of deep learning techniques in mitigating some constraints encountered by SOA classifiers, thereby presenting encouraging prospects for further investigation in EEG-based authentication systems.

6.2.4 Experiment 4: Evaluation of Time Domain Features

Feature extraction plays a crucial role in developing a resilient EEG-based authentication system. In our study, we extracted features in the time domain by estimating the AR coefficients. These features were passed as input to the SOA classifiers for training and testing. As a result, we evaluate our tool on various orders of AR coefficients. These feature sets included AR coefficients (order=1,2,3,4,5,6,7,8,9,10).

Datasets: BrainInvaders15a, ERPCORE: N400, Mantegna2019, COG-BCI Flanker

Utilized Parameters:

- Epoch Interval: 1 second
- Epochs Rejection threshold: 250μV
- Features: AR (order=1,2,3,4,5,6,7,8,9,10)
- Classifiers: LDA, SVM, KNN, RF, NB, LR
- Evaluation Type: Within-Session Evaluation
- Threat Case: Close-Set

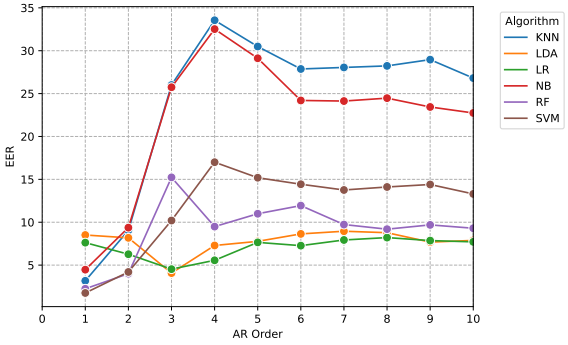
The aforementioned parameters were applied to subject our tool to a rigorous evaluation across a diverse set of datasets, namely BrainInvaders15a, ERPCORE: N400, Mantegna2019, and COG-BCI Flanker. The preprocessing procedures involved segmenting the EEG data into epochs of 1-second intervals and applying a threshold of 250 V for the rejection of epochs with artifacts. For feature extraction, exclusively AR coefficients with orders ranging from 1 to 10 were utilized. The evaluation exclusively employed SOA classification algorithms, including LDA, SVM, KNN, RF, NB, and LR. The assessment was conducted using a Within-Session approach, with particular emphasis on the Close-Set scenario. The outcomes of this experiment was discussed below.

Results of Experiment 4: Figure 6.5 portrays the performance of traditional classifiers on the four datasets, showcasing the impact of varied AR orders. The optimal performance is evident for the BrainInvaders15a, ERPCORE: N400, and Mantegna2019 datasets when employing the lowest AR order, 1. Remarkably, the analyzed datasets demonstrate an increase in EER as the AR order increases. Notably, the COG-BCI Flanker dataset exhibits an EER increase from orders 1 to 3, followed by a consistent decline from orders 3 to 10. However, the data presented underscores a noticeable improvement in classifier efficiency at an AR order of 6. The LDA classifier emerges with the lowest EER across varying AR orders among the classifiers tested.

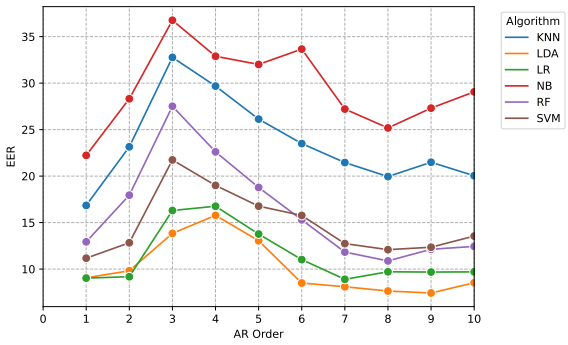
6.2.5 Experiment 5: Evaluation of Frequency Domain Features

Within our study, we harnessed frequency domain characteristics by employing Power PSD as a feature extraction technique. Through this specific experiment, we aim to assess how the utilization of extracted PSD features influences the performance of the authentication algorithms across the complete spectrum of the four selected datasets. The evaluation setup in this experiment remains consistent with that of experiment 4 6.2.4, except for the distinct modification that solely PSD features were extracted and employed for the training and testing of classifiers.

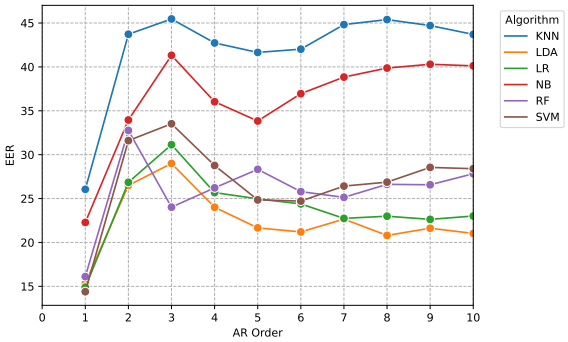
Datasets: BrainInvaders15a, ERPCORE: N400, Mantegna2019, COG-BCI Flanker



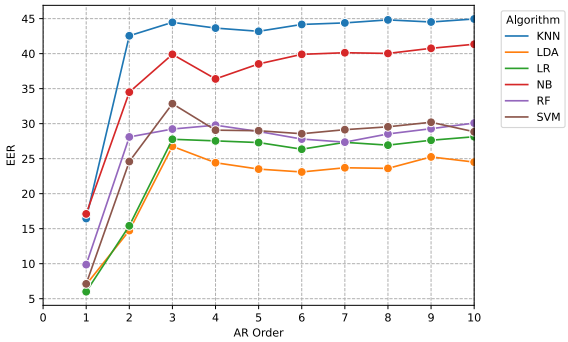
(a) Effect of AR features with orders (1 to 10) on performance of the dataset BrainInvaders15a



(b) Effect of AR features with orders (1 to 10) on performance of the dataset COGBCI Flanker



(c) Effect of AR features with orders (1 to 10) on performance of the dataset ERPCORE: N400



(d) Effect of AR features with orders (1 to 10) on performance of the dataset Mantegna2019

Figure 6.5: Impact of Auto Regressive(AR) Features on the performance of the datasets. Figures (a), (b), (c) and (d) depicts the change in the EER of the traditional classifiers for the datasets BrainInvaders51a, COG-BCI Flanker, ERPCORE: N400 and Mantegna2019 respectively.

Utilized Parameters:

- Epoch Interval: 1 second
- Epochs Rejection threshold: 250μV
- Features: PSD
- Classifiers: LDA, SVM, KNN, RF, NB, LR
- Evaluation Type: Within-Session Evaluation
- Threat Case: Close-Set

Results of Experiment 5: Illustrated in Figure 6.6, the classifiers' performance notably improved when solely PSD features were employed, surpassing the evaluation conducted using only AR features. The data in Figure 6.6 underscores that the EER remains consistently below 16% across all classifiers and datasets. Particularly remarkable is the robust performance of the RF classifier, achieving an EER of less than 5% across each dataset. However, it is worth noting that NB exhibited a consistent decline in performance across all datasets when utilizing PSD features.

Interestingly, the results from Experiment 4, as depicted in Figure 6.5, revealed that utilizing solely AR features led to significantly elevated EER values. Specific classifiers like KNN exhibited EER values as high as 45% on datasets such as ERPCORE: N400 and Mantegna2019. In contrast, the EER obtained using only PSD features did not exceed 16%. This comparison accentuates a distinct enhancement when employing features extracted from the frequency domain. This discrepancy suggests that frequency domain features are adept at capturing unique EEG patterns among individuals, underscoring their greater efficacy than time domain features.

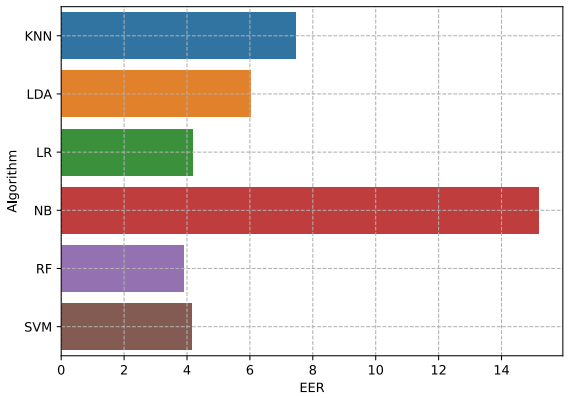
6.2.6 Experiment 6: Evaluation of the combination of Time and Frequency Domain Features

We observed in Experiment 5 and 6 about the distinctive performance trajectories of authentication systems by solely employing AR and PSD features, respectively. Building upon this investigative foundation, we delve into Experiment 7, where we comprehensively explore the impact of integrating both AR and PSD features within our authentication process.

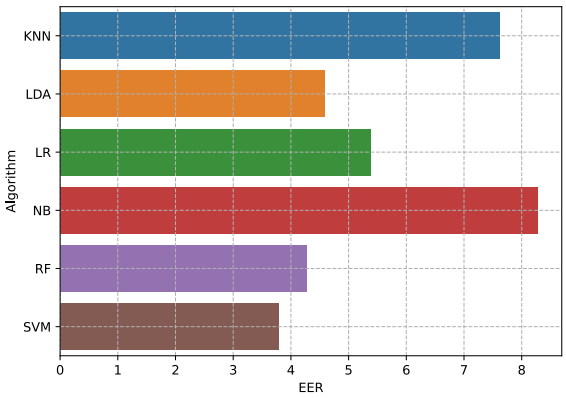
Datasets: BrainInvaders15a, ERPCORE: N400, Mantegna2019, COG-BCI Flanker

Utilized Parameters:

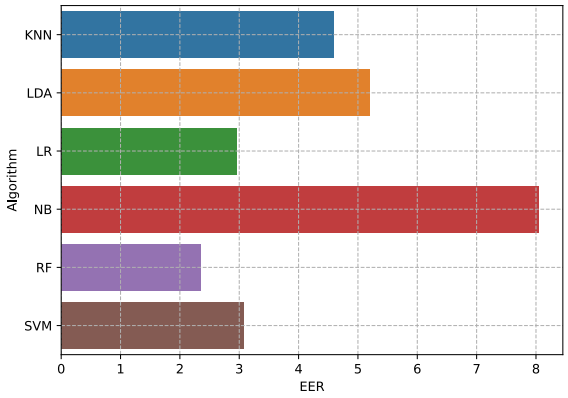
- Epoch Interval: 1 second
- Epochs Rejection threshold: 250μV
- Features: AR (order=1,2,3,4,5,6,7,8,9,10), PSD
- Classifiers: LDA, SVM, KNN, RF, NB, LR
- Evaluation Type: Within-Session Evaluation
- Threat Case: Close-Set



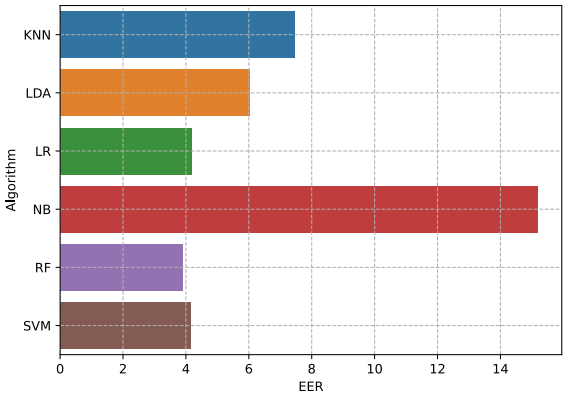
(a) Effect of PSD features on the performance of the dataset BrainInvaders15a



(b) Effect of PSD features on the performance of the dataset COG-BCI Flanker

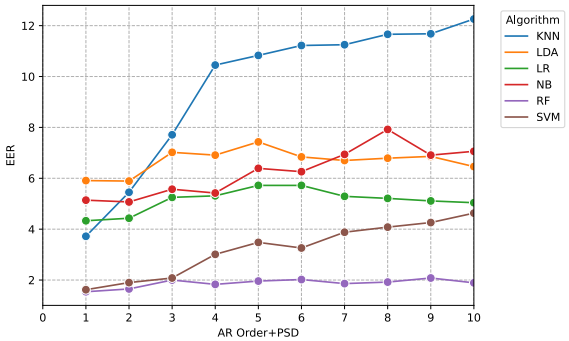


(c) Effect of PSD features on the performance of the dataset ERPCORE: N400

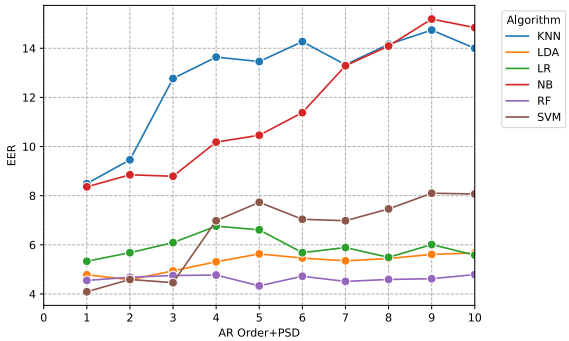


(d) Effect of PSD features on the performance of the dataset Mantegna2019

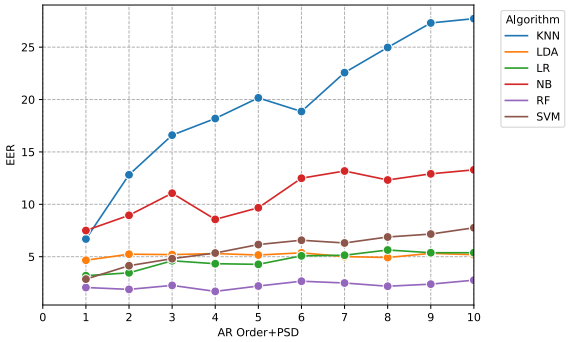
Figure 6.6: Impact of Power Spectral Density (PSD) Features on the performance of the datasets. Figures (a), (b), (c), and (d) depict the change in the EER of the traditional classifiers for the datasets BrainInvaders51a, COG-BCI Flanker, ERPCORE: N400 and Mantegna2019, respectively.



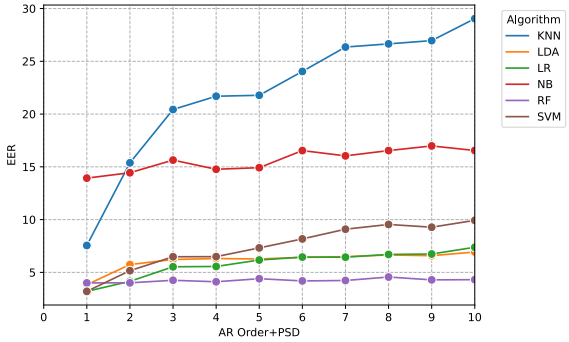
(a) Effect of the combination of AR (from order 1 to 10) and PSD features on the performance of the dataset BrainInvaders15a



(b) Effect of the combination of AR (from order 1 to 10) and PSD features on the performance of the dataset COG-BCI Flanker



(c) Effect of the combination of AR (from order 1 to 10) and PSD features on the performance of the dataset ERPCORE: N400



(d) Effect of the combination of AR (from order 1 to 10) and PSD features on the performance of the dataset Mategna2019

Figure 6.7: The influence of combining PSD and AR features with orders ranging from 1 to 10 is assessed in terms of the datasets' performance. The corresponding changes in the EER of traditional classifiers for the BrainInvaders51a, COG-BCI Flanker, ERPCORE: N400, and Mantegna2019 datasets are illustrated in Figures (a), (b), (c), and (d) respectively.

Results of Experiment 6: While the performance of the classifiers improved using PSD features, the best performance across all datasets is achieved using a combination of AR and PSD features, as depicted in Figure 6.7. This observation highlights the significant benefits of incorporating both separate and complementary features of EEG signal representation. The integration of temporal dynamics collected by AR features and the frequency-specific information provided by PSD features enhances the robustness and comprehensiveness of the authentication systems. This integration offers classifiers with a broader and more varied range of characteristics, which is crucial for understanding the intricacies present in EEG signals among various subjects, sessions, and tasks.

6.2.7 Experiment 7: Evaluation of the Tool with Varied Dataset Sizes

The sample size of the dataset passed to the model for learning plays a crucial role in impacting the overall performance of the authentication system. As mentioned in section 5.2, artifact rejection during the pre-processing stage involves the peak-to-peak rejection approach. The observation was made that varying rejection thresholds impact the quantity of the dataset that triggers an alert. Consequently, this experiment will evaluate the influence of various rejection

criteria on the efficacy of the best-performing SOA classifier, which is RF. The evaluation will be performed on the four datasets using the within-session evaluation scheme within the context of the close-set scenario.

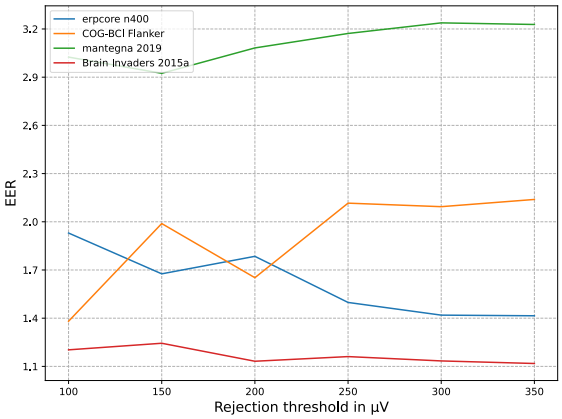
Datasets: BrainInvaders15a, ERPCORE: N400, Mantegna2019, COG-BCI Flanker

Utilized Parameters:

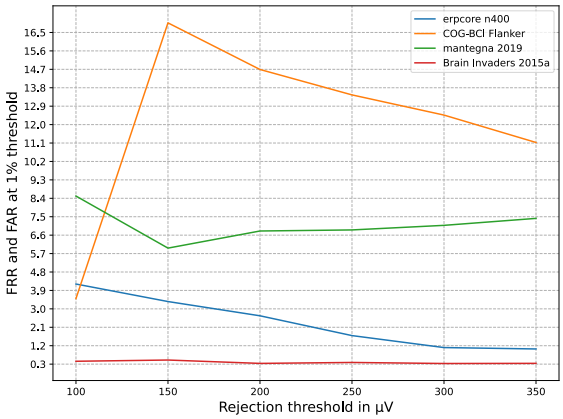
- Epoch Interval: 1 second
- Epochs Rejection threshold: 150μV, 200μV, 250μV, 300μV, 350μV
- Features: AR, PSD
- Classifiers: RF
- Evaluation Type: Within-Session Evaluation
- Threat Case: Close-Set

Results of Experiment 7: Figure 6.8 (a) and (b) presents the obtained EER and FRR at 1% FAR with different rejection thresholds. The results indicate a drop in the EER for the ERPCORE: N400 dataset as the rejection threshold increased from 100μV to 150μV. There was a modest increase in EER at 200μV, followed by a continuous decrease in the EER from 200μV to 350μV thresholds. Similarly, a constant reduction in FRR at a FAR of 1% was observed as the rejection threshold increased from 100μV to 350μV. This suggests a positive correlation between the number of samples and the classifier's performance on ERPCORE: N400, indicating that the classifier's performance improves as the number of samples increases. Nevertheless, this assumption is not universally applicable to all datasets. In the case of the Mantegna2019 dataset, we noticed a notable increase in the EER as the rejection threshold was raised from 150μV to 350μV. However, a slight improvement was observed at the 150μV threshold, where the EER decreased by 3.66% (from 3.02% to 2.92%) and FRR at 1% FAR drops by 29.84% (from 8.51% to 5.97%). This implies that the overall performance of the Mantegna2019 dataset deteriorated as the sample size increased. As the thresholds governing the rejection of epochs are raised, we observe a consistent pattern of improvement and decline in the performance of the RF classifier across the ERPCORE: N400 and Mantegna2019 datasets. However, the COG-BCI Flanker dataset exhibits a distinct pattern in the version of the EER metric, displaying a continuous fluctuation as the thresholds are incrementally increased. Furthermore, it is noteworthy that the dataset BrainInvaders15a demonstrates a minimal shift in EER and FRR at 1% FAR despite variations in the thresholds for epochs rejection. This observation underlines the dataset's robustness to changes in the rejection threshold, suggesting a consistent classifier performance under different rejection conditions.

Based on the findings mentioned above, it is crucial to recognize that implementing a predetermined threshold for rejection may introduce limitations to the suitability of our methodology, given the size of EEG datasets can vary among different types of headsets and experimental conditions. To enhance the flexibility of our framework, we have devised a design that allows researchers to specify their threshold for rejecting epochs. This approach allows for increased customization and adaptation to the unique characteristics of different experimental setups and datasets, thereby enhancing the applicability and robustness of the system in diverse EEG authentication scenarios.

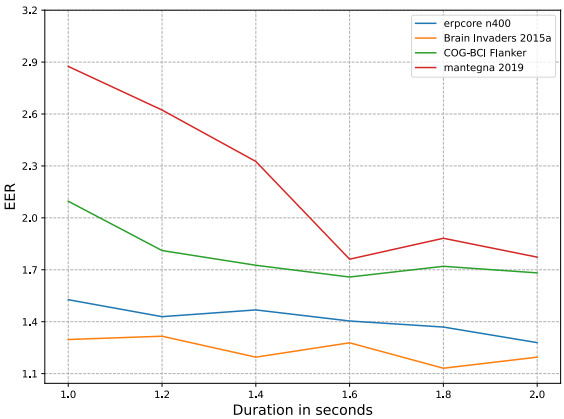


(a) Effect of rejection thresholds on EER with RF classification

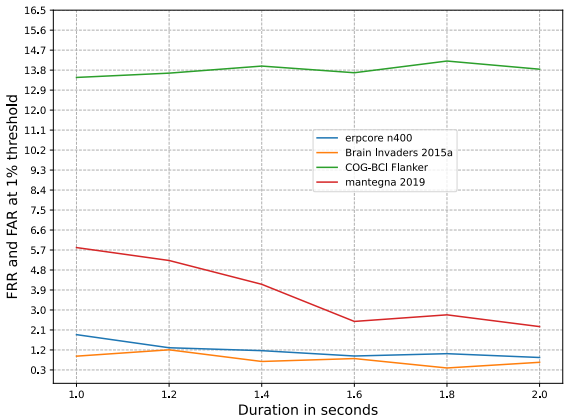


(b) Effect of rejection thresholds on FRR at 1% FAR with RF classification

Figure 6.8: Impact of applying epochs rejection on the performance of the four datasets. Figures (a) and (b) show the EER and FRR at 1% FAR for classifier RF.



(a) Effect of epochs duration on EER with RF classification



(b) Effect of epochs duration on FRR at 1% FAR with RF classification

Figure 6.9: Impact of epoch duration on classification scores and epoch rejection on the four datasets. Figures (a) and (b) shows the EER and FRR at 1% FAR for classifier RF.

6.2.8 Experiment 8: Evaluating of the Tool with Varied Epoch Duration

In this experiment, we evaluate our tool to analyze the effect of different epoch durations on the performance of our authentication system using an RF classifier and the same pre-processing and feature extraction pipeline. The duration of the epochs was meticulously arranged, encompassing a range of 1.0 seconds, 1.2 seconds, 1.4 seconds, 1.6 seconds, 1.8 seconds, and 2.0 seconds. Each epoch was preceded by a 200-millisecond interval before the ERP event. The selection of these specific time intervals enables us to thoroughly investigate the impact of various temporal windows surrounding the ERP occurrence on the system's classification performance. Through examining several epochs, our objective is to get vital knowledge regarding the ideal duration that effectively enhances the accuracy and resilience of the authentication system in real-world scenarios.

Datasets: BrainInvaders15a, ERPCORE: N400, Mantegna2019, COG-BCI Flanker

Utilized Parameters:

- Epoch Interval: 1 seconds, 1.2 seconds, 1.4 seconds, 1.6 seconds, 1.8 seconds, 2 seconds
- Epochs Rejection threshold: 250μV
- Features: AR, PSD
- Classifiers: RF
- Evaluation Type: Within-Session Evaluation
- Threat Case: Close-Set

Results of Experiment 8: As shown in Figure 6.9 (a) and (b), the duration of epoch affects the performance of the classifier RF. The figure illustrates a discernible trend wherein an extension in the epoch duration from 1 second to 2 seconds correlates with a substantial reduction in EER and FRR at 1% FAR, particularly evident in the case of Mantegna2019 dataset. Notably, for the Mantegna2019 dataset, the EER experiences a noteworthy drop of 38.74% (from 2.87% to 1.76%) and the FRR at % also witnesses a significant decline of 57.24% (from 5.81% to 2.48%) as the epochs duration increases from 1 to 1.6 seconds. The dataset BrainInvaders15a displayed consistent fluctuations in its EER as the epochs' duration was extended. Notably, the EER exhibited oscillations of both increments and decrements, occurring within intervals as short as 0.2 seconds in epochs length. A marginal alteration in performance is noted for the ERPCORE: N400 and COG-BCI Flanker datasets, as evidenced by the nearly consistent EER and FRR values at a 1% FAR over increasing epochs duration.

The variability in datasets performance resulting from varied epoch durations highlights the importance of adaptability within our system. Acknowledging the heterogeneous characteristics of EEG data obtained from various types of headsets and experimental configurations, we have developed our framework to allow the user to customize epoch durations. The option to adjust the duration of epochs will enable researchers to customize the time intervals to align with the unique attributes of their data, hence augmenting the flexibility and resilience of our authentication system.

6.3 Evaluation of the Authentication Approaches Utilizing the Tool

We formulated our benchmarking framework by aligning with established conventions in data pre-processing, feature extraction, and classification, aiming to encapsulate the most respected and influential techniques in brainwave authentication. For instance, our decision to apply raw EEG data filtering and artifact removal through peak-to-peak thresholds during data cleaning is grounded in recognized best practices (as detailed in section 2.3). Furthermore, our choice to utilize AR coefficients and PSD for feature extraction, addressing both temporal and frequency domains, mirrors prevalent and validated approaches employed in brainwave authentication, as highlighted in section 2.4. We aimed to construct a benchmarking tool that embodies current state-of-the-art practices and techniques by integrating these well-established methodologies. In subsequent testing phases, we subjected our benchmarking tool to comprehensive evaluations, employing a selection of the most commonly utilized machine learning algorithms in brainwave authentication studies, as outlined in section 2.5. These algorithms included LDA, SVM, RF, KNN, NB, and LR. Furthermore, we incorporated advanced deep learning techniques such as Siamese Neural Networks, ensuring that our framework provides comprehensiveness and adaptability to embrace advanced methodologies.

We have structured our data processing, feature extraction, and classification approach by drawing inspiration from recent benchmarking studies in brainwave authentication. This selected framework was influenced by the methodologies employed in two prominent studies. The first of these studies, conducted by Arias-Cabarcos et al. in 2023 [20], involved benchmarking brainwave authentication models utilizing EEG data collected from 56 participants through the Emotiv+ consumer device. This data collection process encompassed participants engaging in five distinct ERP tasks, which generated ERP effects through P300 and N400 paradigms. The data preprocessing in this study entailed bandpass filtering within the range of 1 to 50 Hz, baseline correction, and epoch rejection set at 150 μV. The epochs were extracted from raw EEG signals with a duration of 1 second, encompassing 100 milliseconds before and 900 milliseconds after the event. Discriminant features were then extracted from both the time and frequency domains, accomplished by estimating AR coefficients and computing the Power Spectrum (PS) of low, α, β, and γ waves. In terms of authentication, the study employed six state-of-the-art algorithms—LR, SVM, LDA, KNN, GNB, and RF—while considering both close-set and open-set attack scenarios. Performance comparison was carried out through metrics such as EER, ROC curves, and FRR and FAR at 1%. The study achieved favorable outcomes, including an EER of 7.2% utilizing the SVM classifier. These achievements are particularly noteworthy given that the EEG data originated from a consumer device, which typically exhibits a lower signal-to-noise ratio than medical-grade EEG devices. Furthermore, within this study, the researchers extended their evaluation to include an assessment of their authentication methodology using high-quality open brainwave data obtained from a medical-grade headset, introducing a comparative analysis of discrepancies. For this purpose, they employed the ERP CORE (Compendium of Open Resources and Experiments) dataset [85]. This is the same dataset that we have also operated in our study. Notably, they replicated the identical preprocessing, feature extraction, and classification processes on the ERP CORE dataset as those applied to their consumer device data. Significantly, the performance of the classifiers experienced a marked enhancement in the ERP CORE dataset. This observation underscores the pivotal role of data quality in constructing robust brainwave authentication systems. The RF classifier demonstrated exceptional performance on the ERP CORE dataset, with an EER of only 1.04%. Additionally, it achieved an FRR of 1% and a FAR of 1.1%. This underscores the critical significance of utilizing high-quality data when developing brainwave authentication systems for enhanced performance.

Another study that significantly influenced the development of our benchmarking framework was undertaken by Fallahi *et al.* [26]. As previously discussed in section 3.2, this study has significantly influenced our approach to implementing the Siamese Neural Network for our research. Notably, Fallahi *et al.* employed high-quality medical-grade brainwave datasets, including ERP CORE dataset [85] and BrainInvaders15a [91]. The focus of their study revolved around the use of biometric identification and verification techniques, explicitly examining the measurement of similarity between registered brain samples and those obtained during the verification process. The research presented a notable solution to the common challenge of retraining state-of-the-art algorithms when new users are added to the system.

6.3.1 TestBed: Replication of other authentications works

Our benchmarking tool was harnessed to replicate and verify the outcomes of the previously mentioned study by Arias-Cabarcos et al. [20] . Their study also delved into dataset ERPCORE: N400, mirroring our own research approach. Employing our tool, we meticulously evaluated the performance across this dataset, leveraging the following parameter configurations.

Datasets: ERPCORE: N400

Utilized Parameters:

- Epoch Interval: 1 second
- Epochs Rejection threshold: 250μV
- Features: AR, PSD
- Classifiers: LDA, SVM, KNN, RF, NB, LR
- Evaluation Type: Within-Session Evaluation
- Threat Case: Close-Set, Open-Set

The aforementioned parameter settings closely mirror those employed in the two referenced studies, with one exception: we incorporated an "Epochs rejection at 250μV" preprocessing step. In alignment with both studies, our benchmarking encompassed the application of AR and PSD for feature extraction and an assortment of classifiers, including LDA, SVM, KNN, RF, NB, LR. We selected "Within-Session Evaluation" for the parameter evaluation type, given the single-session nature of the dataset ERPCORE: N400. Additionally, we set parameters for both "Close-Set" and "Open-Set" threat-case scenarios, mirroring the focused exploration of seen and unseen attacker scenarios in the study by Arias-Cabarcos *et al.* [20].

Results: As depicted in Table 6.3, our tool's results align with the findings of Arias-Cabarcos et al. [20] study. It is noteworthy that RF stands out as the most effective classifier among all the SOA algorithms, mirroring the observations from Arias-Cabarcos *et al.* [20]. For instance, in Arias-Cabarcos et al.'s study, the EER achieved by the RF classifier was 1.04% for the close-set scenario and 1.9% for the open-set scenario. Our tool reached an EER of 1.53% for the close-set scenario and 1.95% for the open-set scenario. Our achieved EER for the RF classifier closely resembles the results of Arias-Cabarcos et al.'s study. Similarly, for the KNN classifier, our achieved EER of 20.22% is slightly lower than the 20.9% EER attained in Arias-Cabarcos' work under close-set settings, indicating minimal disparity between the two results. Additionally, we notice a slight variance in the achieved EER for classifiers like LR, SVM, and LDA between our

Table 6.3: The table displays the mean %EER and FRR at 1% FAR for the dataset ERP-CORE: N400 [85] using within-session evaluation scheme, comparing different classifiers and threat case scenarios. The values in the table are shown in percentages.

Metric	Dataset	LDA	SVM	LR	RF	KNN	NB
Close-Set							
EER	**ERPCORE: N400**	3.09 ± 2.80	5.44 ± 3.79	3.44 ± 2.36	1.53 ± 0.87	20.22 ± 9.41	11.50 ± 5.76
FRR at 1% FAR	**ERPCORE: N400**	13.72	16.50	9.19	1.84	50.76	70.55
Open-Set							
EER	**ERPCORE: N400**	5.63 ± 4.40	6.63 ± 4.22	5.14 ± 3.73	1.95 ± 1.56	20.62 ± 10.36	11.01 ± 7.31
FRR at 1% FAR	**ERPCORE: N400**	41.25	21.90	32.02	4.56	55.34	62.84

study and Arias-Cabarcos' research. For instance, our obtained EER for the LDA classifier in the open-set scenario is 5.63%, marking a 24.3% enhancement compared to the corresponding result of Arias-Cabarcos et al. The outcomes of classifiers LR, LDA, and SVM in our study slightly outperform those achieved in Arias-Cabarcos et al.'s study for the same classifiers. The minor discrepancies can be attributed to several factors. In our research, the available samples for the ERPCORE: N400 dataset after the pre-processing step were 2097, slightly fewer than the 2,268 samples obtained in Arias-Cabarcos et al.'s work. Moreover, our approach to extracting PSD features from raw epochs involved dividing the 1-second ERP epoch into four equally sized time windows with a 50% overlap between each window. This segmentation aimed to distinguish the genuine frequency modulation of the EEG, driven by attention, from potential artifacts induced by the attentional modulation of ERPs [123]. Conversely, precisely implementing PSD features in Arias-Cabarcos et al.'s work needs explicit details. Therefore, the minor variations could also be attributed to differences in the calculation of the PSD for each epoch between our study and that of Arias-Cabarcos et al.'s study.

When evaluating the classifiers' effectiveness using the FRR at 1% FAR metric, a striking correlation emerges through a comparison between our study and the research conducted by Arias-Cabarcos et al. [20]. In a close-set setting, the SVM classifier in our study achieved an FRR at 1% FAR of 16.50%. At the same time, they attained 16.9% for the same SVM classifier in an identical threat scenario, revealing only a marginal variance of 0.40%. Moreover, our investigation unveils a noticeable disparity in performance between open-set (unknown attackers) and close-set (known attackers) scenarios across most classifiers, particularly LDA and LR. As depicted in Table 6.3, classifier LDA experiences a substantial increase of 200.65% (from 13.72% to 41.25%) in FRR at 1% FAR when transitioning from a close-set to an open-set scenario. A similar decline in classifier performance is evident in Arias-Cabarcos et al.'s work, where the FRR at 1% FAR for classifier LDA escalates by 257.37% (from 12.2% to 43.6%) when comparing close-set and open-set scenarios, aligning with our observations.

7 Limitations

Our study introduces a highly adaptable benchmarking framework capable of conducting comprehensive performance comparisons across diverse datasets and employing SOA and advanced deep learning techniques. Intending to enhance the practicality of our tool, we have meticulously designed it to accommodate authentication scenarios involving previously unseen attackers, rendering our benchmarking approach more aligned with real-world situations. However, it is essential to acknowledge that our tool does possess certain limitations, which we will delve into in the upcoming sections, thereby providing a comprehensive perspective on its capabilities and constraints.

7.1 Exclusive Emphasis on ERP Datasets

Our research has focused on gathering predominantly ERP datasets, aligning our benchmarking tool's design with ERP paradigms like P300 and N400. The choice to give priority to ERP datasets, as discussed in section 4.1.1, is based on their favorable Signal-to-Noise Ratio (SNR) and less susceptibility to background disturbances [89]. Although the criteria above provide a strong justification for our decision, it is essential to acknowledge the possibility of further expanding the tool's capabilities by including EEG data obtained from other tasks, such as resting and motor imagery datasets. Such an expansion could effectively broaden the utility of our tool, catering to a broader spectrum of researchers engaged in brainwave authentication, who often explore diverse EEG datasets beyond the limitations of ERP paradigms.

7.2 Constrained Examination of Multi-Session Datasets

Our study incorporated three single-session datasets and a sole multi-session dataset. In section 6.2.1, we conducted comparative analyses, facilitating comprehensive performance evaluations across various datasets. This comparative assessment of EEG-based authentication systems through different datasets offers valuable insights into the system's adaptability, robustness, and generalizability. By scrutinizing the behavior of authentication algorithms across diverse datasets, we can discern performance consistency patterns, identify algorithms' strengths and weaknesses, and assess their reliability under varying conditions. Nevertheless, it is essential to recognize a constraint in our study, which revolves around the fact that we only had access to a single multi-session dataset. While we were able to conduct extensive evaluations with the single-session datasets, the scope of our research was somewhat restricted in terms of benchmarking

brainwave authentication systems against multi-session datasets. Employing solely one multi-session dataset inherently narrows the spectrum of our exploration and the depth of our findings. A broader range of multi-session datasets could have significantly enriched the scope of our research and further extended the applicability of our benchmarking tool. Therefore, more multi-session datasets should be incorporated into the benchmarking tool.

7.3 Sub optimal Siamese Network Training in Cross-Session Evaluation

The results of our cross-session evaluation demonstrated less favorable outcomes than the performance observed in the dataset authentication conducted in a single session. The expectation was that utilizing Siamese Networks, capable of capturing intrinsic brain patterns through deep learning, would yield improved outcomes. This phenomenon can be attributed to the inherent capability of deep learning approaches to effectively adapt to diverse multi-session EEG data, in contrast to traditional algorithms that operate on predetermined features. Indeed, there exist empirical investigations that provide support for this notion. For example, the study by Maiorana [83] reported an EER below 7% using Siamese Networks for subject verification in a cross-session setup. Similarly, Seha and Hatzinakos in 2020 [87] achieved remarkable outcomes, attaining EER levels between 2-4% for cross-session evaluation through deep learning methods. However, our chosen approach for training and testing multi-session data could be the reason behind the suboptimal performance in our cross-session evaluation. Our strategy involves employing EEG data from two sessions for training and utilizing the remaining session data for testing. While this training and testing strategy aligns with certain studies like [21, 130, 16], it has contributed to the less satisfactory outcomes in our specific case.

The underlying principle of Siamese Networks is around the recognition of people through the analysis of similarities between registered brain samples and those presented during the verification procedure. Using a training strategy incorporating two separate sessions representing various temporal points inadvertently introduces unpredictability into the training process. The EEG data obtained during these two sessions may display significant variations due to different circumstances, including cognitive states, environmental stimuli, and the person's physiological parameters. As a result, the acquired embeddings from these sessions will probably exhibit variations, which might undermine the network's capacity to develop persistent patterns of similarity. When these variable embeddings from the two training sessions are compared against those from the remaining session during testing, disparities in EEG data patterns can lead to dissimilarities in the resultant embeddings. This inconsistency in embeddings makes the task of similarity comparison challenging and can subsequently contribute to the suboptimal performance we have observed.

Moreover, it is worth noting that employing data from two sessions for training, mainly when these sessions are seven days apart, introduces a layer of assumption that may not reflect real-world scenarios. Assuming that individuals would be prompted to register twice into an authentication system within a short timeframe is often not aligned with practical usage scenarios.

8 Conclusion and Future Works

8.1 Conclusion

In this study, we have successfully developed a robust benchmarking framework tailored for EEG-based authentication systems, utilizing four publicly available ERP datasets. Our comprehensive evaluation included a range of state-of-the-art algorithms, encompassing LR, LDA, SVM, NB, KNN, RF, and deep learning techniques like Siamese Neural Networks. Our evaluation strategy incorporated both within-session and cross-session assessments, focusing on close-set and open-set scenarios to ensure the tool's applicability in diverse contexts. Remarkably, during within-session evaluation, our results demonstrated exceptional performance for several classifiers, including RF, Siamese, SVM, LDA, and LR. We achieved an impressively low EER of just 1.60% when evaluated with Siamese Networks in the unseen attacker scenario. However, we also observed challenges during cross-session validation, where EER values increased for most classifiers. This highlights the importance of further research and development to enhance cross-session performance.

Moreover, our benchmarking framework's flexibility stands out as a valuable asset. It empowers researchers to customize pre-processing, feature extraction, and authentication parameters according to their needs. The straightforward process, facilitated by a user-friendly YAML configuration file and automated benchmarking scripts, ensures ease of use and minimizes the complexity of programming. As we conclude this thesis, it is evident that our benchmarking tool holds promise for advancing the field of EEG-based authentication. Its open availability will provide researchers with a valuable resource to explore and assess its capabilities, furthering our collective understanding of brainwave authentication. Additionally, as we continue to refine and enhance this tool, we envision a future where it contributes significantly to the development of secure and efficient authentication systems based on EEG data.

8.2 Future Works

The potential of our benchmarking efforts can be significantly broadened by incorporating a broader array of multi-session datasets. Looking ahead, expanding the dataset collection within our tool is essential to encompass those offering multiple sessions. These datasets inherently create a more authentic environment, closely reflecting real-world scenarios. Therefore, implementing benchmarking exercises using these datasets would enhance the understanding of inter-session variability and provide researchers who utilize our tool with a more comprehensive

and holistic viewpoint. Using this strategic method exhibits the potential to augment the resilience and durability of our benchmarking framework, hence fostering profound research and innovation in EEG-based authentication.

Furthermore, as we have previously addressed the limitations of our cross-session evaluation approach in section 7.3, we propose future enhancements to refine this strategy. We recommend a revised cross-session evaluation methodology in the context of multi-session datasets. This involves training the Siamese Networks utilizing the EEG data from one of the sessions using the Triplet Loss technique. Subsequently, for verification, EEG data from the remaining sessions are employed. To illustrate, consider the COG-BCI dataset featuring three EEG sessions per subject. We advocate training the Siamese model with data from session one to generate embeddings from the raw EEG data. The trained Siamese model can generate these embeddings from sessions two and three. Ultimately, the similarity of the embeddings from session two and session three can be individually compared with the enrolled brain embeddings using the Euclidean distance. The proposed methodology allows for the individual evaluation of a subject's data from each session compared to the enrolled samples, resulting in authentication through the determination of similarity. Adopting this suggested alteration can improve the effectiveness of cross-session evaluation, hence offering a more precise and reliable assessment of EEG-based authentication over numerous sessions for each individual.

Finally, it is essential to emphasize that our benchmarking tool will be available to the public. This release will allow academics and developers in brainwave authentication to investigate its practicality and effectiveness directly. As the availability of this tool increases, its acceptance and use by different stakeholders will contribute to a complete assessment of its efficacy across numerous situations. Moreover, it is essential to acknowledge that, like any tool, ours does have inherent limitations. However, these limitations can be addressed and refined through the collaborative efforts of researchers who employ the tool. This iterative improvement process will undoubtedly enhance the overall applicability and value of this benchmarking tool, ultimately benefiting the entire brainwave authentication field.

Bibliography

[1] Sk Al Mamun, Md Ashiq Mahmood, and Md Ashiqul Amin. Ensuring security of encrypted information by hybrid aes and rsa algorithm with third-party confirmation. In *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 337–343. IEEE, 2021.

[2] Isuru Jayarathne, Michael Cohen, and Senaka Amarakeerthi. Brainid: Development of an eeg-based biometric authentication system. In *2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pages 1–6. IEEE, 2016.

[3] Tien Pham, Wanli Ma, Dat Tran, Phuoc Nguyen, and Dinh Phung. Eeg-based user authentication in multilevel security systems. In *Advanced Data Mining and Applications: 9th International Conference, ADMA 2013, Hangzhou, China, December 14-16, 2013, Proceedings, Part II 9*, pages 513–523. Springer, 2013.

[4] Krishna Dharavath, Fazal A Talukdar, and Rabul H Laskar. Study on biometric authentication systems, challenges and future trends: A review. In *2013 IEEE international conference on computational intelligence and computing research*, pages 1–7. IEEE, 2013.

[5] Paul A Grassi, Michael E Garcia, and James L Fenton. Draft nist special publication 800-63-3 digital identity guidelines. *National Institute of Standards and Technology, Los Altos, CA*, 2017.

[6] Arash Habibi Lashkari, Samaneh Farmand, Dr Zakaria, Omar Bin, Dr Saleh, et al. Shoulder surfing attack in graphical password authentication. *arXiv preprint arXiv:0912.0951*, 2009.

[7] Anupam Das, Joseph Bonneau, Matthew Caesar, Nikita Borisov, and XiaoFeng Wang. The tangled web of password reuse. In *NDSS*, 2014.

[8] Lawrence O'Gorman. Comparing passwords, tokens, and biometrics for user authentication. *Proceedings of the IEEE*, 91(12):2021–2040, 2003.

[9] Tien Pham, Wanli Ma, Dat Tran, Phuoc Nguyen, and Dinh Phung. A study on the feasibility of using eeg signals for authentication purpose. In *International Conference on Neural Information Processing*, pages 562–569. Springer, 2013.

[10] Rupinder Saini and Narinder Rana. Comparison of various biometric methods. *International Journal of Advances in Science and Technology*, 2(1):24–30, 2014.

[11] Debnath Bhattacharyya, Rahul Ranjan, Farkhod Alisherov, Minkyu Choi, et al. Biometric authentication: A review. *International Journal of u-and e-Service, Science and Technology*, 2(3):13–28, 2009.

[12] Patricia Arias-Cabarcos, Thilo Habrich, Karen Becker, Christian Becker, and Thorsten Strufe. Inexpensive brainwave authentication: new techniques and insights on user acceptance. In *Proceedings of the 30th {USENIX} Security Symposium ({USENIX} Security 21)*, pages 55–72, 2021.

[13] Daria La Rocca, Patrizio Campisi, Balazs Vegso, Peter Cserti, György Kozmann, Fabio Babiloni, and F De Vico Fallani. Human brain distinctiveness based on eeg spectral coherence connectivity. *IEEE transactions on Biomedical Engineering*, 61(9):2406–2412, 2014.

[14] Sebastien Marcel and Jose Del R. Millan. Person authentication using brainwaves (eeg) and maximum a posteriori model adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(4):743–752, 2007.

[15] Maria V Ruiz Blondet, Sarah Laszlo, and Zhanpeng Jin. Assessment of permanence of non-volitional eeg brainwaves as a biometric. In *IEEE International Conference on Identity, Security and Behavior Analysis (ISBA 2015)*, pages 1–6. IEEE, 2015.

[16] Sherif Nagib Abbas Seha and Dimitrios Hatzinakos. A new approach for eeg-based biometric authentication using auditory stimulation. In *2019 International Conference on Biometrics (ICB)*, pages 1–6. IEEE, 2019.

[17] Mohammed J Abdulaal, Alexander J Casson, and Patrick Gaydecki. Performance of nested vs. non-nested svm cross-validation methods in visual bci: Validation study. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 1680–1684. IEEE, 2018.

[18] Javad Sohankar, Koosha Sadeghi, Ayan Banerjee, and Sandeep KS Gupta. E-bias: A pervasive eeg-based identification and authentication system. In *Proceedings of the 11th ACM Symposium on QoS and Security for Wireless and Mobile Networks*, pages 165–172, 2015.

[19] Shridatt Sugrim, Can Liu, Meghan McLean, and Janne Lindqvist. Robust performance metrics for authentication systems. In *Network and Distributed Systems Security (NDSS) Symposium 2019*, 2019.

[20] Patricia Arias-Cabarcos, Matin Fallahi, Thilo Habrich, Karen Schulze, Christian Becker, and Thorsten Strufe. Performance and usability evaluation of brainwave authentication techniques with consumer devices. *ACM Transactions on Privacy and Security*, 2023.

[21] Vinay Jayaram and Alexandre Barachant. Moabb: trustworthy algorithm benchmarking for bcis. *Journal of neural engineering*, 15(6):066011, 2018.

[22] Marco Simões, Davide Borra, Eduardo Santamaría-Vázquez, GBT-UPM, Mayra Bittencourt-Villalpando, Dominik Krzemiński, Aleksandar Miladinović, Neural_Engineering_Group, Thomas Schmid, Haifeng Zhao, et al. Bciaut-p300: A multi-session and multi-subject benchmark dataset on autism for p300-based brain-computer-interfaces. *Frontiers in Neuroscience*, 14:568104, 2020.

[23] David Hübner, Thibault Verhoeven, Konstantin Schmid, Klaus-Robert Müller, Michael Tangermann, and Pieter-Jan Kindermans. Learning from label proportions in brain-computer interfaces: Online unsupervised learning with guarantees. *PloS one*, 12(4):e0175856, 2017.

[24] Christoph Guger, Shahab Daban, Eric Sellers, Clemens Holzner, Gunther Krausz, Roberta Carabalona, Furio Gramatica, and Guenter Edlinger. How many people are able to control a p300-based brain–computer interface (bci)? *Neuroscience letters*, 462(1):94–98, 2009.

[25] Kathryn K Toffolo, Edward G Freedman, and John J Foxe. Evoking the n400 event-related potential (erp) component using a publicly available novel set of sentences with semantically incongruent or congruent eggplants (endings). *Neuroscience*, 501:143–158, 2022.

[26] Matin Fallahi, Thorsten Strufe, and Patricia Arias-Cabarcos. Brainnet: Improving brainwave-based biometric recognition with siamese networks. In *2023 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 53–60. IEEE, 2023.

[27] Qiong Gui, Maria V Ruiz-Blondet, Sarah Laszlo, and Zhanpeng Jin. A survey on brain biometrics. *ACM Computing Surveys (CSUR)*, 51(6):1–38, 2019.

[28] Priyanka A Abhang, Bharti W Gawali, and Suresh C Mehrotra. *Introduction to EEG-and speech-based emotion recognition*. Academic Press, 2016.

[29] g.tec medical engineering GmbH | Brain-Computer Interface Neurotechnology — gtec.at. https://www.gtec.at/. [Accessed 21-08-2023].

[30] EPOC+ - 14 Channel EEG — emotiv.com. https://www.emotiv.com/epoc/. [Accessed 22-08-2023].

[31] Shuai Zhang, Lei Sun, Xiuqing Mao, Cuiyun Hu, Peiyuan Liu, et al. Review on eeg-based authentication technology. *Computational Intelligence and Neuroscience*, 2021, 2021.

[32] Isao Nakanishi, Sadanao Baba, and Chisei Miyamoto. Eeg based biometric authentication using new spectral features. In *2009 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, pages 651–654. IEEE, 2009.

[33] Kavitha P Thomas and A Prasad Vinod. Eeg-based biometric authentication using gamma band power during rest state. *Circuits, Systems, and Signal Processing*, 37:277–289, 2018.

[34] RB Paranjape, J Mahovsky, L Benedicenti, and Z Koles. The electroencephalogram as a biometric. In *Canadian Conference on Electrical and Computer Engineering 2001. Conference Proceedings (Cat. No. 01TH8555)*, volume 2, pages 1363–1366. IEEE, 2001.

[35] Katharine Brigham and BVK Vijaya Kumar. Subject identification from electroencephalogram (eeg) signals during imagined speech. In *2010 Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 1–8. IEEE, 2010.

[36] DHR Blackwood and Walter J Muir. Cognitive brain potentials and their application. *The British Journal of Psychiatry*, 157(S9):96–101, 1990.

[37] Zhendong Mu, Jianfeng Hu, and Jianliang Min. Eeg-based person authentication using a fuzzy entropy-related approach with two electrodes. *Entropy*, 18(12):432, 2016.

[38] Arnaud Delorme. Eeg is better left alone. *Scientific reports*, 13(1):2372, 2023.

[39] Lucian Jose Gonçales, Kleinner Farias, Lucas Kupssinskü, and Matheus Segalotto. The effects of applying filters on eeg signals for classifying developers' code comprehension. *Journal of applied research and technology*, 19(6):584–602, 2021.

[40] Tzyy-Ping Jung, Scott Makeig, Marissa Westerfield, Jeanne Townsend, Eric Courchesne, and Terrence J Sejnowski. Removal of eye activity artifacts from visual event-related potentials in normal and clinical subjects. *Clinical neurophysiology*, 111(10):1745–1758, 2000.

[41] Steven J Luck. *An introduction to the event-related potential technique*. MIT press, 2014.

[42] Ricardo Vigário, Jaakko Sarela, Veikko Jousmiki, Matti Hamalainen, and Erkki Oja. Independent component approach to the analysis of eeg and meg recordings. *IEEE transactions on biomedical engineering*, 47(5):589–593, 2000.

[43] Aapo Hyvärinen. The fixed-point algorithm and maximum likelihood estimation for independent component analysis. *Neural Processing Letters*, 10:1–5, 1999.

[44] Wan Amirah W Azlan and Yin Fen Low. Feature extraction of electroencephalogram (eeg) signal-a review. In *2014 IEEE conference on biomedical engineering and sciences (IECBES)*, pages 801–806. IEEE, 2014.

[45] Yong Zhang, Bo Liu, Xiaomin Ji, and Dan Huang. Classification of eeg signals based on autoregressive model and wavelet packet decomposition. *Neural Processing Letters*, 45:365–378, 2017.

[46] James Pardey, Stephen Roberts, and Lionel Tarassenko. A review of parametric modelling techniques for eeg analysis. *Medical engineering & physics*, 18(1):2–11, 1996.

[47] Gabriel Emile Hine, Emanuele Maiorana, and Patrizio Campisi. Resting-state eeg: A study on its non-stationarity for biometric applications. In *2017 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5. IEEE, 2017.

[48] Patrizio Campisi and Daria La Rocca. Brain waves for automatic biometric-based user recognition. *IEEE transactions on information forensics and security*, 9(5):782–800, 2014.

[49] Zhiguo Zhang. Spectral and time-frequency analysis. In *EEG Signal Processing and feature extraction*, pages 89–116. Springer, 2019.

[50] Chengalvarayan Radhakrishnamurthy Hema, MP Paulraj, and Harkirenjit Kaur. Brain signatures: A modality for biometric authentication. In *2008 International conference on electronic design*, pages 1–4. IEEE, 2008.

[51] Kavitha P Thomas, AP Vinod, et al. Eeg-based biometrie authentication using self-referential visual stimuli. In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 3048–3053. IEEE, 2017.

[52] Gary G Yen and K-C Lin. Wavelet packet feature extraction for vibration monitoring. *IEEE transactions on industrial electronics*, 47(3):650–667, 2000.

[53] Dingyin Hu, Wei Li, and Xi Chen. Feature extraction of motor imagery eeg signals based on wavelet packet decomposition. In *The 2011 IEEE/ICME international conference on complex medical engineering*, pages 694–697. IEEE, 2011.

[54] Daria La Rocca, Patrizio Campisi, and Jordi Solé-Casals. Eeg based user recognition using bump modelling. In *2013 international conference of the BIOSIG special interest group (BIOSIG)*, pages 1–12. IEEE, 2013.

[55] Toshiaki Koike-Akino, Ruhi Mahajan, Tim K Marks, Ye Wang, Shinji Watanabe, Oncel Tuzel, and Philip Orlik. High-accuracy user identification using eeg biometrics. In *2016 38th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 854–858. IEEE, 2016.

[56] Hichem Sahbi. Totally deep support vector machines. *arXiv preprint arXiv:1912.05864*, 2019.

[57] Danial Jahed Armaghani, Panagiotis G Asteris, Behnam Askarian, Mahdi Hasanipanah, Reza Tarinejad, and Van Van Huynh. Examining hybrid and single svm models with different kernels to predict rock brittleness. *Sustainability*, 12(6):2229, 2020.

[58] Derek A Pisner and David M Schnyer. Support vector machine. In *Machine learning*, pages 101–121. Elsevier, 2020.

[59] Shu-yin Xia, Zhong-yang Xiong, Yue-guo Luo, and Li-mei Dong. A method to improve support vector machine based on distance to hyperplane. *Optik*, 126(20):2405–2410, 2015.

[60] Gi-Chul Yang. Next-generation personal authentication scheme based on eeg signal and deep learning. *Journal of Information Processing Systems*, 16(5), 2020.

[61] Emily C Zabor, Chandana A Reddy, Rahul D Tendulkar, and Sujata Patil. Logistic regression in clinical studies. *International Journal of Radiation Oncology* Biology* Physics*, 112(2):271–277, 2022.

[62] Tanya Piplani, Nick Merill, and John Chuang. Faking it, making it: fooling and improving brain-based authentication with generative adversarial networks. In *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–7. IEEE, 2018.

[63] Rajdeep Ghosh, Souvik Phadikar, Nabamita Deb, Nidul Sinha, Pranesh Das, and Ebrahim Ghaderpour. Automatic eyeblink and muscular artifact detection and removal from eeg signals using k-nearest neighbor classifier and long short-term memory networks. *IEEE Sensors Journal*, 23(5):5422–5436, 2023.

[64] André Zúquete, Bruno Quintela, and Joao Paulo Silva Cunha. Biometric authentication using brain responses to visual stimuli. In *International conference on bio-inspired systems and signal processing*, volume 2, pages 103–112. SCITEPRESS, 2010.

[65] Ali Haghpanah Jahromi and Mohammad Taheri. A non-parametric mixture of gaussian naive bayes classifiers based on local independent features. In *2017 Artificial intelligence and signal processing conference (AISP)*, pages 209–212. IEEE, 2017.

[66] Akshay Valsaraj, Ithihas Madala, Nikhil Garg, Mohit Patil, and Veeky Baths. Motor imagery based multimodal biometric user authentication system using eeg. In *2020 International Conference on Cyberworlds (CW)*, pages 272–279. IEEE, 2020.

[67] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[68] Ranran Zhang, Xiaoyan Xiao, Zhi Liu, Wei Jiang, Jianwen Li, Yankun Cao, Jianmin Ren, Dongmei Jiang, and Lizhen Cui. A new motor imagery eeg classification method fb-trcsp+ rf based on csp and random forest. *IEEE Access*, 6:44944–44950, 2018.

[69] Basavarajaiah DM, Bhamidipati Narasimha Murthy, Basavarajaiah DM, and Bhamidipati Narasimha Murthy. Random forest and concept of decision tree model. *Design of Experiments and Advanced Statistical Techniques in Clinical Research*, pages 133–156, 2020.

[70] AM Mahmud Chowdhury and Masudul H Imtiaz. A machine learning approach for person authentication from eeg signals. In *2023 IEEE 32nd Microelectronics Design & Test Symposium (MDTS)*, pages 1–5. IEEE, 2023.

[71] Ting Yu, Chun-Shu Wei, Kuan-Jung Chiang, Masaki Nakanishi, and Tzyy-Ping Jung. Eeg-based user authentication using a convolutional neural network. In *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)*, pages 1011–1014. IEEE, 2019.

[72] Ayman Khalafallah, Aly Ibrahim, Bahieeldeen Shehab, Hisham Raslan, Omar Eltobgy, and Shady Elbaroudy. A pragmatic authentication system using electroencephalography signals. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 901–905. IEEE, 2018.

[73] Isuru Jayarathne, Michael Cohen, and Senaka Amarakeerthi. Person identification from eeg using various machine learning techniques with inter-hemispheric amplitude ratio. *PloS one*, 15(9):e0238872, 2020.

[74] Yinfeng Fang, Haiyang Yang, Xuguang Zhang, Han Liu, and Bo Tao. Multi-feature input deep forest for eeg-based emotion recognition. *Frontiers in neurorobotics*, 14:617531, 2021.

[75] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, 3(1):18–31, 2011.

[76] Haiping Huang, Linkang Hu, Fu Xiao, Anming Du, Ning Ye, and Fan He. An eeg-based identity authentication system with audiovisual paradigm in iot. *Sensors*, 19(7):1664, 2019.

[77] Qingxue Zhang, Dian Zhou, and Xuan Zeng. Machine learning-empowered biometric methods for biomedicine applications. *AIMS Medical Science*, 4(3):274–290, 2017.

[78] Nikhil Iyengar, CK Peng, Raymond Morin, Ary L Goldberger, and Lewis A Lipsitz. Age-related alterations in the fractal scaling of cardiac interbeat interval dynamics. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, 271(4):R1078–R1084, 1996.

[79] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

[80] Amir Jalaly Bidgoly, Hamed Jalaly Bidgoly, and Zeynab Arezoumand. Towards a universal and privacy preserving eeg-based authentication system. *Scientific Reports*, 12(1):2531, 2022.

[81] Gerwin Schalk, Dennis J McFarland, Thilo Hinterberger, Niels Birbaumer, and Jonathan R Wolpaw. Bci2000: a general-purpose brain-computer interface (bci) system. *IEEE Transactions on biomedical engineering*, 51(6):1034–1043, 2004.

[82] Fabien Lotte, Laurent Bougrain, Andrzej Cichocki, Maureen Clerc, Marco Congedo, Alain Rakotomamonjy, and Florian Yger. A review of classification algorithms for eeg-based brain–computer interfaces: a 10 year update. *Journal of neural engineering*, 15(3):031005, 2018.

[83] Emanuele Maiorana. Eeg-based biometric verification using siamese cnns. In *New Trends in Image Analysis and Processing–ICIAP 2019: ICIAP International Workshops, BioFor, PatReCH, e-BADLE, DeepRetail, and Industrial Session, Trento, Italy, September 9–10, 2019, Revised Selected Papers 20*, pages 3–11. Springer, 2019.

[84] Louis Korczowski, Martine Cederhout, Anton Andreev, Grégoire Cattan, Pedro Luiz Coelho Rodrigues, Violette Gautheret, and Marco Congedo. *Brain Invaders calibration-less P300-based BCI with modulation of flash duration Dataset (bi2015a)*. PhD thesis, GIPSA-lab, 2019.

[85] Emily S Kappenman, Jaclyn L Farrens, Wendy Zhang, Andrew X Stewart, and Steven J Luck. Erp core: An open resource for human event-related potential research. *NeuroImage*, 225:117465, 2021.

[86] Gan Huang, Zhenxing Hu, Weize Chen, Shaorong Zhang, Zhen Liang, Linling Li, Li Zhang, and Zhiguo Zhang. M3cv: A multi-subject, multi-session, and multi-task database for eeg-based biometrics challenge. *NeuroImage*, 264:119666, 2022.

[87] Sherif Nagib Abbas Seha and Dimitrios Hatzinakos. Eeg-based human recognition using steady-state aeps and subject-unique spatial filters. *IEEE Transactions on Information Forensics and Security*, 15:3901–3910, 2020.

[88] Pádraig Cunningham and Sarah Jane Delany. Underestimation bias and underfitting in machine learning. In *Trustworthy AI-Integrating Learning, Optimization and Reasoning: First International Workshop, TAILOR 2020, Virtual Event, September 4–5, 2020, Revised Selected Papers 1*, pages 20–31. Springer, 2021.

[89] Blair C Armstrong, Maria V Ruiz-Blondet, Negin Khalifian, Kenneth J Kurtz, Zhanpeng Jin, and Sarah Laszlo. Brainprint: Assessing the uniqueness, collectability, and permanence of a novel method for erp biometrics. *Neurocomputing*, 166:59–67, 2015.

[90] Terence W Picton et al. The p300 wave of the human event-related potential. *Journal of clinical neurophysiology*, 9:456–456, 1992.

[91] Louis Korczowski, Martine Cederhout, Anton Andreev, Grégoire Cattan, Pedro Luiz Coelho Rodrigues, Violette Gautheret, and Marco Congedo. *Brain Invaders calibration-less P300-based BCI with modulation of flash duration Dataset (bi2015a)*. PhD thesis, GIPSA-lab, 2019.

[92] Marcel F Hinss, Emilie S Jahanpour, Bertille Somon, Lou Pluchon, Frédéric Dehais, and Raphaëlle N Roy. Open multi-session and multi-task eeg cognitive dataset for passive brain-computer interface applications. *Scientific Data*, 10(1):85, 2023.

[93] Francesco Mantegna, Florian Hintz, Markus Ostarek, Phillip M Alday, and Falk Huettig. Distinguishing integration and prediction accounts of erp n400 modulations in language processing through experimental design. *Neuropsychologia*, 134:107199, 2019.

[94] Gijsbrecht Van Veen, Alexandre Barachant, Anton Andreev, Grégoire Cattan, Pedro Coelho Rodrigues, and Marco Congedo. Building brain invaders: Eeg data of an experimental validation. *arXiv preprint arXiv:1905.05182*, 2019.

[95] Erwan Vaineau, Alexandre Barachant, Anton Andreev, Pedro C Rodrigues, Grégoire Cattan, and Marco Congedo. Brain invaders adaptive versus non-adaptive p300 brain-computer interface dataset. *arXiv preprint arXiv:1904.09111*, 2019.

[96] Louis Korczowski, Ekaterina Ostaschenko, Anton Andreev, Grégoire Cattan, Pedro Luiz Coelho Rodrigues, Violette Gautheret, and Marco Congedo. *Brain Invaders calibration-less P300-based BCI using dry EEG electrodes Dataset (bi2014a)*. PhD thesis, GIPSA-lab, 2019.

[97] Louis Korczowski, Ekaterina Ostaschenko, Anton Andreev, Grégoire Cattan, Pedro Luiz Coelho Rodrigues, Violette Gautheret, and Marco Congedo. *Brain Invaders Solo versus Collaboration: Multi-User P300-based Brain-Computer Interface Dataset (bi2014b)*. PhD thesis, GIPSA-lab, 2019.

[98] Junfeng Gao, Hongjun Tian, Yong Yang, Xiaolin Yu, Chenhong Li, and Nini Rao. A novel algorithm to enhance p300 in single trials: Application to lie detection using f-score and svm. *Plos one*, 9(11):e109700, 2014.

[99] Louis Korczowski, Martine Cederhout, Anton Andreev, Grégoire Cattan, Pedro Luiz Coelho Rodrigues, Violette Gautheret, and Marco Congedo. *Brain Invaders Cooperative versus Competitive: Multi-User P300-based Brain-Computer Interface Dataset (bi2015b)*. PhD thesis, GIPSA-lab, 2019.

[100] R Mouček, L Vařeka, T Prokop, J Štěbeták, and P Brha. Event-related potential data from a guess the number brain-computer interface experiment on school children. *Scientific data*, 4(1):1–11, 2017.

[101] Jan Sosulski and Michael Tangermann. Spatial filters for auditory evoked potentials transfer between different experimental conditions. In *GBCIC*, 2019.

[102] Min-Ho Lee, O-Yeon Kwon, Yong-Jeong Kim, Hong-Kyung Kim, Young-Eun Lee, John Williamson, Siamac Fazli, and Seong-Whan Lee. Eeg dataset and openbmi toolbox for three bci paradigms: An investigation into bci illiteracy. *GigaScience*, 8(5):giz002, 2019.

[103] Vladislav Goncharenko, Rafael Grigoryan, and Alina Samokhina. Raccoons vs demons: Multiclass labeled p300 dataset. *arXiv preprint arXiv:2005.02251*, 2020.

[104] Amirmahmoud Houshmand Chatroudi, Reza Rostami, Ali Motie Nasrabadi, and Yuko Yotsumoto. Effect of inhibition indexed by auditory p300 on transmission of visual sensory information. *Plos one*, 16(2):e0247416, 2021.

[105] Grégoire Hugues Cattan, Anton Andreev, Cesar Mendoza, and Marco Congedo. A comparison of mobile vr display running on an ordinary smartphone with standard pc display for p300-bci stimulus presentation. *IEEE Transactions on Games*, 13(1):68–77, 2021.

[106] Kyungho Won, Moonyoung Kwon, Minkyu Ahn, and Sung Chan Jun. Eeg dataset for rsvp and p300 speller brain-computer interfaces. *Scientific Data*, 9(1):388, 2022.

[107] Judith Pijnacker, Nina Davids, Marjolijn van Weerdenburg, Ludo Verhoeven, Harry Knoors, and Petra van Alphen. Semantic processing of sentences in preschoolers with

specific language impairment: Evidence from the n400 effect. *Journal of Speech, Language, and Hearing Research*, 60(3):627–639, 2017.

[108] Dejan Draschkow, Edvard Heikel, Melissa L-H Vo, Christian J Fiebach, and Jona Sassenhagen. No evidence from mvpa for different processes underlying the n300 and n400 incongruity effects in object-scene processing. *Neuropsychologia*, 120:9–17, 2018.

[109] Anna Marzecová, Antonio Schettino, Andreas Widmann, Iria SanMiguel, Sonja A Kotz, and Erich Schröger. Attentional gain is modulated by probabilistic feature expectations in a spatial cueing task: Erp evidence. *Scientific Reports*, 8(1):54, 2018.

[110] Alice Hodapp and Milena Rabovsky. The n400 erp component reflects an error-based implicit learning signal during language comprehension. *European Journal of Neuroscience*, 54(9):7125–7140, 2021.

[111] Elisabeth Rabs, Francesca Delogu, Heiner Drenhaus, and Matthew W Crocker. Situational expectancy or association? the influence of event knowledge on the n400. *Language, Cognition and Neuroscience*, 37(6):766–784, 2022.

[112] Pia Schoknecht, Dietmar Roehm, Matthias Schlesewsky, and Ina Bornkessel-Schlesewsky. The interaction of predictive processing and similarity-based retrieval interference: an erp study. *Language, Cognition and Neuroscience*, 37(7):883–901, 2022.

[113] Alma Lindborg, Lea Musiolek, Dirk Ostwald, and Milena Rabovsky. Semantic surprise predicts the n400 brain potential. *Neuroimage: Reports*, 3(1):100161, 2023.

[114] Kate Stone, Bruno Nicenboim, Shravan Vasishth, and Frank Rösler. Understanding the effects of constraint and predictability in erp. *Neurobiology of Language*, 4(2):221–256, 2023.

[115] David F Dinges and John W Powell. Microcomputer analyses of performance on a portable, simple visual rt task during sustained operations. *Behavior research methods, instruments, & computers*, 17(6):652–655, 1985.

[116] Wayne K Kirchner. Age differences in short-term retention of rapidly changing information. *Journal of experimental psychology*, 55(4):352, 1958.

[117] Yamira Santiago-Espada, Robert R Myer, Kara A Latorella, and James R Comstock Jr. The multi-attribute task battery ii (matb-ii) software for human performance and workload research: A user's guide. Technical report, 2011.

[118] Barbara A Eriksen and Charles W Eriksen. Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & psychophysics*, 16(1):143–149, 1974.

[119] Yan Ma, Yiou Tang, Yang Zeng, Tao Ding, and Yifu Liu. An n400 identification method based on the combination of soft-dtw and transformer. *Frontiers in Computational Neuroscience*, 17:1120566, 2023.

[120] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[121] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, et al. Meg and eeg data analysis with mne-python. *Frontiers in neuroscience*, page 267, 2013.

[122] Mainak Jas, Denis Engemann, Federico Raimondo, Yousra Bekhti, and Alexandre Gramfort. Automated rejection and repair of bad trials in meg/eeg. In *2016 international workshop on pattern recognition in neuroimaging (PRNI)*, pages 1–4. IEEE, 2016.

[123] CM Cómez, M Vazquez, E Vaquero, D Lopez-Mendoza, and M^{a}J Cardoso. Frequency analysis of the eeg during spatial selective attention. *International Journal of Neuroscience*, 95(1-2):17–32, 1998.

[124] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[125] Benyamin Ghojogh, Milad Sikaroudi, Sobhan Shafiei, Hamid R Tizhoosh, Fakhri Karray, and Mark Crowley. Fisher discriminant triplet and contrastive losses for training siamese networks. In *2020 international joint conference on neural networks (IJCNN)*, pages 1–7. IEEE, 2020.

[126] Haoran Wu, Zhiyong Xu, Jianlin Zhang, Wei Yan, and Xiao Ma. Face recognition based on convolution siamese networks. In *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–5. IEEE, 2017.

[127] Mohsen Heidari and Kazim Fouladi-Ghaleh. Using siamese networks with transfer learning for face recognition on small-samples datasets. In *2020 International Conference on Machine Vision and Image Processing (MVIP)*, pages 1–4. IEEE, 2020.

[128] Cavoukian Ann and Stoianov Alex. Biometric encryption: A positive-sum technology that achieves strong authentication, security, and privacy. *Privacy by Design Book available at www. ipc. on. ca. Accessed on December*, 6:2009, 2007.

[129] Pablo Arnau-González, Miguel Arevalillo-Herráez, Stamos Katsigiannis, and Naeem Ramzan. On the influence of affect in eeg-based subject identification. *IEEE Transactions on Affective Computing*, 12(2):391–401, 2018.

[130] Sherif Nagib Abbas Seha and Dimitrios Hatzinakos. Longitudinal assessment of eeg biometrics under auditory stimulation: a deep learning approach. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pages 1386–1390. IEEE, 2021.

A

Appendix

The benchmarking tool is available in our GitLab repository at the following URL: `https://github.com/Avichaurasia/Brain-Models`. The tool has been built using Python and incorporates various statistical and machine-learning Python packages. Hence, it is imperative to establish a Python environment as a first step.

Setting Up the Python Environment

To successfully run this Python project, creating and configuring a suitable Python environment is essential. The following steps need to be followed to set up the environment:

1. **Python Installation**: Initially, verifying the presence of Python in our operating system is crucial. If Python is not pre-installed, it can be acquired straight from the official Python website. Subsequently, we adhere to the installation instructions outlined in the website, customized to suit our operating system. If Anaconda has been successfully installed, it is noteworthy to mention that the Anaconda installation often includes the Python programming language. If the action above is taken, it is possible to go to the subsequent stage.

2. **Anaconda Installation (if needed)**: Suppose Anaconda is not currently installed, and it is desired to utilize it to manage Python environments. In that case, it is possible to get the software by downloading it from the official website, which can be accessed at Anaconda. Anaconda installation may be accomplished according to the instructions tailored to each operating system. Anaconda provides a user-friendly method for creating and administrating virtual environments through the use of Conda. This specific characteristic has significant value in data science and scientific computing initiatives.

3. **Virtual Environment Creation**: It is advisable to establish a virtual environment to segregate the dependencies of this specific project from other Python packages installed on our system. Virtual environments help maintain clean and distinct Python environments for individual projects. To create a virtual environment, follow these steps:

    - **Navigate to Project Directory**: To begin, we access the terminal or command prompt and proceed to the project's root directory by utilizing the *cd* command. For example: cd /path/to/project

- **Environment Configuration File**: Check if the project includes an environment configuration file. This file is typically named *environment.yml* or *requirements.txt* and lists the required Python packages and their versions.

- **Create Virtual Environment**:Subsequently, the requisite command is executed to generate a virtual environment by using the configuration file. An example of a command that may be used for Conda environments is the use of a *environment.yml* file:

  For example: conda env create -f environment.yml

  we can also utilize *requirement.txt* to create the virtual environment by using the pip command:

  python -m venv venv

  source venv/bin/activate

  pip install -r requirements.txt

4. **Activate the Virtual Environment**: After the virtual environment has been established, proceed to activate it. Activation is a crucial process that guarantees using an isolated environment and its corresponding dependencies in our project. To activate the conda environment, it is necessary to utilize the proper command according to the operating system in use:

   For example: conda activate master_thesis (for MacOS/Linux)

Edit Configuration File: Upon activating the Conda environment, navigate to the designated project directory. A file named *single_dataset.yml* can be located within the "configuration_files" folder. The single_dataset.yml file is adjusted based on the exemplified configurations in the following sections.

Execute the Automation Script: Launch the automated script single_dataset_benchmark.py in Python. This script streamlines all the tasks related to data preprocessing, feature extraction, and classification for a single dataset. It conducts benchmarking assessments across multiple classifiers for the specified dataset.

A.1 Appendix: YAML Configuration for Within-Session Evaluation

A.1.1 Configuration with Default parameters for Dataset and Pre-processing Pipeline

Listing A.1: Benchmarking pipeline using the dataset's default parameters and auto-regressive features with SVM classification

```
name: "BrainInvaders2015a"

dataset:
  - name: BrainInvaders2015a
    from: deeb.datasets
```

```
pipelines:
  "AR+SVM":
    - name: AutoRegressive
      from: deeb.pipelines

    - name: SVC
      from: sklearn.svm
      parameters:
        kernel: 'rbf'
        class_weight: "balanced"
        probability: True
```

Run the Python file single_dataset_benchmark.py from the terminal with the command "python single_dataset_benchmark.py".

A.1.2 Pipeline Incorporating Dataset Parameters and Auto-Regressive Order

Listing A.2: Benchmarking pipeline using dataset's parameters and Auto Regressive order with SVM classification

```
name: "BrainInvaders2015a"

dataset:
  - name: BrainInvaders2015a
    from: deeb.datasets
    parameters:
      subjects: 10
      interval: [-0.1, 0.9]
      rejection_threshold: 200

pipelines:
  "AR+SVM":
    - name: AutoRegressive
      from: deeb.pipelines
      parameters:
        order: 5

    - name: SVC
      from: sklearn.svm
      parameters:
        kernel: 'rbf'
        class_weight: "balanced"
        probability: True
```

A.1.3 Pipeline Utilizing Both Auto-Regressive (AR) and Power Spectral Density (PSD) Features

Listing A.3: Benchamrking pipeline for dataset BrainInvaders15a with AR and PSD features with classifier SVM

```
name: "BrainInvaders2015a"

dataset:
  - name: BrainInvaders2015a
    from: deeb.datasets
    parameters:
      subjects: 10
      interval: [-0.1, 0.9]
      rejection_threshold: 200

pipelines:
  "AR+PSD+SVM":
    - name: AutoRegressive
      from: deeb.pipelines
      parameters:
        order: 5

    - name: PowerSpectralDensity
      from: deeb.pipelines

    - name: SVC
      from: sklearn.svm
      parameters:
        kernel: 'rbf'
        class_weight: "balanced"
        probability: True
```

A.1.4 Pipeline Incorporating Siamese Neural Network

Listing A.4: Benchamrking pipeline for dataset BrainInvaders15a with Siamese Networks

```
name: "BrainInvaders2015a"

dataset:
  - name: BrainInvaders2015a
    from: deeb.datasets
    parameters:
      subjects: 10
      interval: [-0.1, 0.9]
      rejection_threshold: 200

pipelines:
  "Siamese":
    - name : Siamese
      from: deeb.pipelines
      parameters:
```

```
        EPOCHS: 10
        batch_size: 256
        verbose: 1
        workers: 1
```

A.1.5 Pipeline Combining Traditional Algorithms and Siamese Neural Network

Listing A.5: Benchamrking pipeline for dataset BrainInvaders15a with traditional and deep learning methods

```
name: "BrainInvaders2015a"

dataset:
  - name: BrainInvaders2015a
    from: deeb.datasets
    parameters:
      subjects: 10
      interval: [-0.1, 0.9]
      rejection_threshold: 200

pipelines:s
  "AR+PSD+SVM":
    - name: AutoRegressive
      from: deeb.pipelines
      parameters:
        order: 6

    - name: PowerSpectralDensity
      from: deeb.pipelines

     "Siamese":
    - name : Siamese
      from: deeb.pipelines
      parameters:
        EPOCHS: 10
        batch_size: 256
        verbose: 1
        workers: 1

    - name: SVC
      from: sklearn.svm
      parameters:
        kernel: 'rbf'
        class_weight: "balanced"
        probability: True
```

A.2 Appendix: YAML Configuration for Cross-Session Evaluation

A.2.1 Pipeline Combining Traditional Algorithms and Siamese Neural Network for Cross-Session Evaluation

Listing A.6: Benchamrking pipeline for multi-session dataset COGBCI: FLANKER with traditional and deep learning methods

```
name: "COGBCIFLANKER"

dataset:
  - name: COGBCIFLANKER
    from: deeb.datasets
    parameters:
      subjects: 10
      interval: [-0.1, 0.9]
      rejection_threshold: 200


"Siamese":
    - name : Siamese
      from: deeb.pipelines
      parameters:
        EPOCHS: 10
        batch_size: 256
        verbose: 1
        workers: 1

pipelines:

  "AR+PSD+SVM":
    - name: AutoRegressive
      from: deeb.pipelines
      parameters:
        order: 6

    - name: PowerSpectralDensity
      from: deeb.pipelines

    - name: SVC
      from: sklearn.svm
      parameters:
        kernel: 'rbf'
        class_weight: "balanced"
        probability: True
```