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

Design and Evaluation of Brainwave Authentication Systems

by
Avinash Kumar Chaurasia

Thesis Supervisor:
Prof. Dr. Patricia Arias Cabarcos

Paderborn, August 14, 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.

Ort, Datum	Unterschrift

Abstract.

Contents

1 Introduction **1**
- 1.1 Motivation . . . 1
- 1.2 Problem Description . . . 3
- 1.3 Solution Overview . . . 5
- 1.4 Thesis Structure . . . 6

2 Background **7**
- 2.1 Background on EEG . . . 7
- 2.2 EEG Instruments and Data Acquisition . . . 7
- 2.3 Data Acquisition Procedures . . . 9
- 2.4 Common EEG Artifacts . . . 9
  - 2.4.1 EEG Features . . . 9
  - 2.4.2 Authentication Algorithms . . . 10
  - 2.4.3 Common Performance Metrics . . . 10

3 Related Work **11**
- 3.1 Existing studies exploring cross-session variability . . . 11
- 3.2 Siamese Neural Networks in Brainwave Authentication Studies . . . 12
- 3.3 Benchmarking works on Brianwave Authentication . . . 12

4 Solution Approach **15**
- 4.1 Survey Open Datasets . . . 15
  - 4.1.1 Potential Datasets Excluded from the Final Study . . . 17
  - 4.1.2 Overview of the selected Datasets . . . 19
- 4.2 Workflow . . . 22

5 Implementation **25**
- 5.1 Loading Datasets . . . 25
- 5.2 Pre-Processing . . . 26
- 5.3 Feature-Extraction . . . 28
- 5.4 Classification . . . 30
  - 5.4.1 Supervised based Learning Classification . . . 30
  - 5.4.2 Similarity Based Learning . . . 32
  - 5.4.3 Automated Benchmarking . . . 34

6 Evaluation and Results **35**
6.1 Evaluation Metrics . . . 35
6.2 Results . . . 35
6.2.1 Within-Session Evaluation Results . . . 36
6.2.2 Cross-Session Evaluation Results . . . 40
6.2.3 Within-Session Vs Cross-Session . . . 42
6.2.4 Impact of Feature Extraction on Performance . . . 43
6.2.5 Effect on Performance due to Epochs Rejection . . . 47
6.2.6 Effect on Performance due to Epoch Duration . . . 48

7 Discussion **51**
7.1 Comparison with existing works . . . 51

8 Conclusion and Future Works **53**
8.1 Conclusion . . . 53
8.2 Limitations . . . 53
8.3 Future Works . . . 53

Bibliography **54**

A Appendix **63**
A.1 Appendix: YAML Configuration for Within-Session Evaluation . . . 63
A.1.1 Configuration with Default parameters for Dataset and Pre-processing Pipeline . . . 63
A.1.2 Pipeline Incorporating Dataset Parameters and Auto-Regressive Order . . . 63
A.1.3 Pipeline Utilizing Both Auto-Regressive (AR) and Power Spectral Density (PSD) Features . . . 64
A.1.4 Pipeline Incorporating Siamese Neural Network . . . 65
A.1.5 Pipeline Combining Traditional Algorithms and Siamese Neural Network 65
A.2 Appendix: YAML Configuration for Cross-Session Evaluation . . . 66
A.2.1 Pipeline Combining Traditional Algorithms and Siamese Neural Network for Cross-Session Evaluation . . . 66

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. Authentication procedures play a crucial role in maintaining information confidentiality by verifying the identity of the user requesting access to secure data [1]. And when it comes to user authentication, the primary goal of user authentication is to deny or confirm an identity claimed by a particular person. Authentication happens in two phases: enrolment and verification. During the first phase, the user has to enroll or register in the system, and therefore user's data is captured and stored in the database. 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, and based on the similarity of the data, the system grants or rejects the user. At the moment, there are three ways to authenticate a user: through something they know (e.g., password), something they own (e.g., a token or an ID card), and something they are (e.g., fingerprint, eyes, face or other biometric data [2].

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 has some advantages, such as passwords being easy to use and maintain. Further, passwords can be revoked easily when compromised. However, passwords suffer from many vulnerabilities, such as complex passwords that are often hard to remember. As a result, users tend to use short and easy-to-remember passwords and reuse them across multiple websites, which exposes the passwords to attackers to breach [3]. A study by Das *et al.* [4] examined several hundred thousand leaked passwords from eleven websites and conducted a password reuse survey based on which it was estimated that 44 to 51% of users reuse the same password across multiple websites. In their study, the authors also developed a cross-site guessing platform that could 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) [3], 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 [5]. 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. A second advantage is that it provides compromise detection since the absence of it is observable (loss of a password, however, does not offer this advantage) [5]. While tokens hold some preeminence over passwords in certain respects, they still need to be a foolproof authentication method because they have several security flaws. For instance, tokens are susceptible to theft [6] and duplication, meaning that someone might create a counterfeit device [5].

Finally, more advanced than the former two authentication methods, biometric-based authentication utilizes user-unique biometrics for verification. Biometrics can be divided into two classes: Physiological and behavioral. Physiological biometrics are related to the physical features of the human body and therefore differ from person to person. Fingerprints, face recognition, hand geometry, and iris recognition are some examples of physiological biometrics [7]. 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. Clearly, biometrics provide more effective authentication mechanisms than passwords and tokens, but biometric-based authentication still needs to be indomitable. It is possible to record or photograph biometric information, such as a voice, face, iris, retina, and fingerprint [6]. Further, unlike passwords and tokens, biometrics are not easy to replace if they are lost or compromised [5], 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 [8]. 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 [9]. The following are some of the advantages of brainwaves that give them a giant leap over other biometric traits for authentication:

1. Brain activities cannot be seen from the outside and are therefore impervious to any form of surveillance [8], unlike other biometric traits, for instance, face or gait, which are observable from the outside and can be exploited to identify users without their consent [10].

2. It is also impractical to steal brainwaves because a person's brain activity is susceptible to their stress and mood, and an aggressor cannot make the victim repeat their mental passphrase [10]. 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 [11]. 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 [8].

4. Brainwaves are organically a part of the human body, so even those who are physically disabled can utilize them, unlike with fingerprints or other types of technology, which may not be possible [6].

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 in order to make sure 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 feature extraction, data pre-processing, and modeling techniques. Numerous machine learning algorithms such as Linear Discriminant Analysis (LDA) [12], Support Vector Machine (SVM) [13], and Naïve Bayes (NB) [14] 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 [15]. Other standard metrics to assess an authentication system's performance include False Acceptance Rate (FAR) and False Rejection Rate (FRR). FAR is defined as the proportion of times the system mistakenly accepts the unauthorized person. On the contrary, FRR gives an overview of instances where the system has denied access to an authorized person. The point where both FAR and FRR are equal is known as Equal Error Rate (ERR) [8].

   Additionally, more than ACC, FAR, and FRR comparisons is required since the specifics of the intrinsic trade-offs that a system must make when implemented are concealed by these standard performance metrics [15]. These metrics, however, are tied to a specific configuration of the classification threshold. Instead, using Receiver-Operating-Characteristic (ROC) curves to depict results is advisable. These curves plot the relationship between the False Acceptance Rate (FAR) and True Positive Rate (1-FRR), parametrically linked to the threshold value [8]. Evaluating existing research on brainwave authentication is complex due to often incomplete metric reporting (frequently only optimized configurations are presented, without ROCs) and variations in samples, algorithms, experimental conditions, and other performance-affecting factors, which are not uniformly reported or accounted for [16].

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 both its general effectiveness and its usefulness.

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, while the open-set considers the attackers who are unknown to the system. 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[16]. Hence, the authentication systems must be able to identify and reject the known attackers as well as the attackers, completely 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.

4. **Lack of Research on Intra-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. 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 intra-session variability among users can be accounted for a significant drop in the system's performance. Unfortunately, this crucial issue has been not been addressed by most of the researchers working on brainwave authentication systems.

5. **Reproducibility 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 [17]. 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.

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<=30), including studies such as [18, 19, 20, 21]. 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. Additionally, the researchers developing those datasets rarely share their source code and data, hindering the reproduction of the results needed to test new algorithms [17].

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) EEG-based authentication models compare when evaluated under the same conditions?

1.3 Solution Overview

In the preceding section 1.2, we discussed the intricate challenges this study aims to resolve. In this section, we provide a preliminary outline of the proposal that is being contemplated. This thesis aims to develop a robust and scalable benchmarking framework where the performance of different SOA authentication algorithms can be compared, utilizing pertinent evaluation metrics. The following points offer a concise overview of our suggested solutions tailored to address the research questions articulated in the previous section.

1. Our study strategically employs 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 [22]. 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 [16]. Therefore, it is imperative to ascertain whether lowering the FAR threshold to increase security may unintentionally render the system less usable. The utilization of all the above mentioned evaluation metrics in our study, provides an effective solution to the first research question outlined in the previous section.

2. Our approach employs Siamese Neural Networks (SNN) alongside state-of-the-art (SOA) algorithms. SNN 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 [23]. One of the most significant advantages of using SNN 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, SNN can generate a unique embedding for the newcomer and compare it to the existing ones, thus reducing computational time significantly.

3. As discussed in the previous section, understanding both close-set and open-set scenarios is essential to determining the resilience and applicability of brainwave authentication systems. Therefore, our research extends beyond the close-set scenarios and greatly emphasizes open-set scenarios. Open-set scenarios present a unique challenge as the system identifies and rejects the attacker's brain signals that it has not previously encountered. As a result, we evaluate each authentication approach under both threat case scenarios, including the SNN.

4. The foundation of our research relies on four publicly accessible EEG datasets. 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 down our selection to one particular dataset, which offered three EEG recording sessions,

each session conducted at an interval of seven days. As a result, our study utilizes three single-session datasets and one multi-session dataset for evaluating the performance of various authentication algorithms.

5. Our proposed approach 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.

6. Each dataset selected for our research 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. Designing an authentication framework based on such a large population 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.

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 background and core concepts of brainwave authentication, such as common EEG artifacts 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. Additionally, we present a short overview of the workflow employed in our benchmarking framework in this chapter. The practical implementation and mythologies 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. This chapter assesses the performance of various authentication algorithms, the outcomes of the research, and a comprehensive analysis of the evaluation results. In the forthcoming chapter 7, an analysis will be conducted on the replicated results of our framework through a comparative examination of our research with prior investigations in brainwave authentication. Chapter 8 concludes with a discussion of this study's findings and potential future enhancements to the proposed application.

2

Background

2.1 Background on EEG

Developing a direct interface for communication and control between human brains and computers has been a subject of contemplation within the scientific community for an extended period. Numerous research and development initiatives have been employed to actualize this concept, resulting in its emergence as a highly burgeoning field of scientific investigation in recent times [24]. This vast area of research has also led to the development of something now known as Brain-Computer Interface (BCI). BCI technologies have been extensively employed in the health sector, where they are utilized in various tasks such as fatigue detection, sleep quality assessment, and clinical fields, such as abnormal brain disease detection and prediction including applications to seizure, Parkinson's disease, Alzheimer's disease, and schizophrenia [25].

BCI can be classified into three groups: invasive, partially invasive, and non-invasive. The first two classifications of BCIs necessitate the surgical placement of sensors(electrodes) within the brain cortex to capture the minute currents produced due to cerebral responses. Although invasive methods of BCI provide superior signal quality, the surgical dangers and requirement for long-term installation of invasive devices offset any potential benefit of improved signal quality [26]. On the other hand, non-invasive technology utilizes external neuroimaging devices to record brain activities, consisting of functional near-infrared spectroscopy (fNIRS), functional magnetic resonance imaging (fMRI), and electroencephalography (EEG). Non-invasive EEG-based devices have been the most widely utilized modality for practical BCIs and clinical applications due to the relative improvements in signal quality, dependability, and mobility compared with alternative imaging techniques [25]. With the development of inexpensive and easily portable EEG devices such as Neurosky, BCI is not restricted to the health sector but has entered new realms like gaming. Research on brain biometrics has recently garnered a lot of attention in this area, which has been further encouraged by the limitations of using passwords to prove online identity [8].

2.2 EEG Instruments and Data Acquisition

To acquire the EEG) signals, an EEG apparatus necessitates the inclusion of sensors capable of establishing conductive contact with the scalp. This amplifier 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

(a) Medical grade EEG headset from g.tec

(b) EPOC/EPOC+ wearable headset

Figure 2.1: 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.

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.USBamp developed by G.Tech Medical Engineering GmbH[2]. An example EEG headset from G.Tech is illustrated in Figure 2.1 (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 2.1 (b). Consumer devices are cheaper than medical-grade EEG headsets and more friendly but have a poor signal-to-noise ratio [27].

[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/

2.3 Data Acquisition Procedures

The firing of neurons in the brain is significantly influenced by the mental state of individuals, exhibiting a solid susceptibility to both external environmental stimuli and internal self-regulation. Hence, it is imperative to devise a specialized collecting paradigm to gather EEG signals [29]. 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 in a calm, relaxed state and are not performing cognitive tasks. As a result, resting state protocols have been employed in many brainwave authentication studies such as [30, 31]. While the resting tasks offer simplicity in data collection, they are greatly influenced by the outside world, and it is not easy to ensure complete silence in a natural application environment [29].

In contrast to protocols for rest, protocols for cognitive activities are characterized by a higher degree of complexity. One category of cognitive protocols encompasses mental tasks. Mental tasks involve the subject imagining doing something specific (e.g., imagine moving their left and right hand or image closing or opening a fist), causing the associated EEG signals to appear [27]. This approach demonstrates suitability for people across various physical limitations and visual impairments, exhibiting a high degree of applicability [29]. Nevertheless, it is worth noting that motor imagery and mental tasks demand specialized training to generate proper responses, making them challenging to execute [32].

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 [29]. They are generally elicited by exposing subjects to external audio or visual stimuli. EPRs are influenced by the subject's knowledge, level of motivation, and cognitive capacities [33], making them more likely to display distinctive, unique traits helpful for authentication. One notable drawback of ERPs, in comparison to EEG, lies in the increased complexity of the elicitation methods associated with ERPs. EEG can be obtained without the need for any specific stimulation of the user, but ERPs can only be obtained when the user is subjected to a specific and carefully controlled kind of stimulation [27].

2.4 Common EEG Artifacts

After the EEG data is acquired, it is imperative to eliminate any noise intruding the EEG signals to obtain clean and precise readings. EEG signals can be corrupted by various artifacts, either of 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 itself, or from external factors in the environment, including electromagnetic fields from other electronic devices.

2.4.1 EEG Features

- Auto regressive Coefficients: Yule walker and Burg's method
- Power Spectral Density: Fourier Transformation, Discrete Fourier Transformation and welch's periodgram.
- Common Spatial Patterns (CSP)

- Time-Frequency Analysis:
- Wavelet Transform:
- Statistical Features

2.4.2 Authentication Algorithms

- Linear Discriminant Analysis
- Logistic Regression
- Support Vector Machine
- K Nearest Neighbour
- Gaussian Naive Bayes
- Deep Learning

2.4.3 Common Performance Metrics

- Confusion metrics
- Accuracy, Precision, Recall, F1-Score
- EER
- ROC-Curve
- DET-Curve
- FAR and FRR

3

Related Work

The unique attributes and individualistic patterns inherent in brain signals have attracted considerable research interest towards constructing brainwave authentication systems. Consequently, this section will delve into a selection of pertinent studies that align closely with the research objectives of our study. We aim to explore and discuss the findings of these prior works, which have sought to address similar challenges that our research intends to tackle.

3.1 Existing studies exploring cross-session variability

Although studies investigating the effects of intra-class variability across sessions in brainwave authentication are scarce, certain researchers have focused on this area. One of the most extensive works on this area was done by Huang *et al.* [34], in 2022, who had explored EEG variability across sessions, subjects, and tasks. The study contains EEG data from 106 subjects, 96 out of 106 participated in two sessions conducted 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 greater research into EEG variability across sessions and subjects.

While the results of Huang et al.'s cross-session evaluation were inferior, Seha and Hatzinakos [35] in 2020 produced impressive results in a similar study area using steady-state Auditory Evoked Potentials (AEPs) for EEG-based recognition. The study involved EEG experiment on 40 subjects across two sessions held on separate days. The study demonstrated exceptional results even when evaluated across cross-session (multiple-sessions). EER of mere 2-4% was achieved in cross-session evaluation that is 16 times more effective than that of Huang et al.'s work on cross-session evaluation.

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 model's effectiveness and hinders practical application [23]. 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.* [36] presented a notable study employing the publicly available Physionet dataset [37] 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 key 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 training, an aspect often impractical given the limited size of most brainwave datasets [38].

Maiorana [39] 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 30 seconds verification probe, a significantly good result considering the cross-session variability in brain data.

Lately Fallahi *et al.* [23] 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)** [40] and **ERP Core** [41]. Unlike Maiorana's [39] 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 Benchmarking works on Brianwave Authentication

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. As noted in section 2.4.2, algorithms like SVM, LDA, and KNN have been

widely adopted in many brainwave authentication studies such as [42, 1, 43] because of their simplicity and ability to find the discriminant patterns in brain signals. Jayarathne *et al.* [44] in 2020 compared the algorithms as mentioned earlier performance on the EEG data recorded from 12 subjects using an EMOTIV Epoc+ headset. Accuracy was calculated for the different combinations of electrodes, and the best-performing classifier turned out to be KNN with $99.0 \pm 0.8\%$, followed by SVM achieving $98.03 \pm 0.1\%$ accuracy and LDA with $98.01 \pm 0.5\%$ accuracy respectively. Meanwhile, Huang *et al.* [45] opted for a different approach, concentrating on alternative algorithms like NB, NN, and LR. 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. Furthermore, performance metrics like ACC, TPR, FPR, and ROC-Curve were chosen to demonstrate to evaluate the performance of classifiers on the EEG data of 30 subjects. NB demonstrated the worst performance among the classifiers, registering average ACC, TPR, and FPR of 77.96%, 75.71%, and 19.80% respectively. LR has better performance than NB with average ACC, TPR, FPR of 81.59%, 79.04%, and 15.05% respectively. The NN classifier, however, displayed the most exceptional performance, recording average ACC, TPR, and FPR of 82.69%, 81.96%, and 17.38% respectively.

4

Solution Approach

This study aims to develop a benchmarking suite on EEG-based authentication systems for a larger set of participants (N>100) using various open medical-grade EEG datasets. 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. This section comprehensively describes the tasks performed to compare the performance of different brainwave authentication algorithms. on open EEG datasets.

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 randomly classify the users based on their brain data.

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 [15]. 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 considerable amount of time, handling the noisy data.

4. **Overfitting or Underfitting**: Bad quality datasets can lead to overfitting or underftitting while creating the machine learning models. Overfitting occurs when the model is too complex that it starts learning from the noise, and underfitting arises when the model is too simple to understand the intricate patterns in the data. Both of these situations may result in inaccurate predictions.

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

Although consumer devices offer simplicity and are more user-friendly than traditional EEG devices, their EEG data have lower SNR than EEG devices. And considering the potential pitfalls associated with 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 [17]. 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 [46], and can assess instantaneous reactions to short stimuli [29], 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 [47]—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 [8]. 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 dataset 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 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 [48, 49, 50, 51] 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 sections, we provide a concise overview of the datasets that were excluded and included in our study, respectively. The datasets that were excluded from our study had the potential to be included, but ultimately were not incorporated due to the aforementioned conditions [**?**].

[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/`

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

Dataset	Year	#Subjects	EEG Device	#Channels	Sampling Rate	#Sessions	EEG task
BrainInvaders12 [52]	2012	25	NeXus-32	16	128 Hz	1	Visual Stimuli
BrainInvaders13a [53]	2013	24	g.GAMMAcap	16	512 Hz	1	Visual Stimuli
BrainInvaders14a [54]	2014	64	g.Sahara	16	512 Hz	1	Visual Stimuli
BrainInvaders14b [55]	2014	37	g.GAMMAcap	32	512 Hz	1	Visual Stimuli
Gao et al. [56]	2014	30	Neuroscan	12	500 Hz	1	Visual Stimuli
BrainInvaders15a [48]	2015	43	g.GAMMAcap	32	512 Hz	1	Visual Stimuli
BrainInvaders15b [57]	2015	44	g.GAMMAcap	32	512 Hz	1	Visual Stimuli
Mouček et al. [58]	2017	250	BrainVision	3	n.a.	1	Visual Stimuli
Hubner et al. [19]	2017	13	BrainAmp DC, Brain Products	31	1000 Hz	1	Visual Stimuli Auditory Stimuli
Sosulski and Tangermann [59]	2019	13	BrainAmp, EasyCap	31	1000 Hz	1	Visual Stimuli
Lee et al. [60]	2019	54	BrainAmp	62	1000 Hz	2	Visual Stimuli
Simões et al. [18]	2020	15	g.tec	8	250 Hz	7	Visual Stimuli
Goncharenko et al. [61]	2020	60	NVX-52	8	500 Hz	1	Visual Stimuli
Chatroudi et al. [62]	2021	24	g.tec	64	1200 Hz	1	Visual Stimuli
Cattan et al. [63]	2021	21	g.USBamp, g.tec	16	512 Hz	1	Visual Stimuli
ERPCORE: P300 [50]	2021	40	Biosemi	30	1024 Hz	1	Visual Stimuli
Won et al. [64]	2022	55	Biosemi	32	512 Hz	1	Visual Stimuli

4.1.1 Potential Datasets Excluded from the Final Study

- **Mouček et al.** [58]: This dataset was made available for public use in 2017. The EEG experiments were conducted in primary and secondary schools across the Czech Republic, which involved the participation of 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 then presented with corresponding visual stimuli while experimenters observe online event-related potential waveforms and attempt to predict the number being considered.

  While this dataset has, by far, contained the most number of participants, i.e., 250, and also fulfills all the conditions set by us 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. This is because each subject's EEG experiment was terminated as soon as 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, for other subjects, the experimenter was unable to accurately guess the correct number even after three attempts, resulting in a higher number of samples being observed for such subjects. We believed such an unbalanced dataset could be susceptible to bias and overfitting, so

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. [65]	2017	45	actiCap	32	500 Hz	1	Auditory Stimuli
Draschkow et al. [66]	2018	40	BrainAmp, actiChamp	64	1000 Hz	1	Visual Stimuli
Marzecová et al. [67]	2018	18	BrainAmp	59	500 Hz	1	Visual Stimuli
Mantegna et al. [51]	2019	31	BrainAmp, EasyCap	65	512 Hz	1	Auditory Stimuli
ERPCORE: N400 [50]	2021	40	Biosemi	30	1024 Hz	1	Visual Stimuli
Hodapp and Rabovky [68]	2021	33	BrainAmp	64	1000 Hz	1	Visual Stimuli
Rabs et al. [69]	2022	38	BrainVision	26	500 Hz	1	Visual Stimuli
Schoknecht et al. [70]	2022	38	ActiCap, ActiChamp	58	500 Hz	1	Visual Stimuli
Toffolo et al. [21]	2022	24	Biosemi	128	512 Hz	1	Auditory Stimuli
Lindborg et al. [71]	2022	40	BrainVision	64	2046 Hz	1	Visual Stimuli
Stone et al. [72]	2023	64	TMSi Refa	32	512 Hz	1	Visual Stimuli

we chose not to include it in our study.

- **Hubner et al.** [19]: 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** [59]: 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.** [66]: 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 perfect for our study and was originally included in our study. However, we encountered two issues while working on this dataset. Our framework is designed to scrap EEG data from the internet directly. Unfortunately, when retrieving this dataset from the data repository, we encountered an error with the message "Bad Magic Number". 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** [69]: 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.** [18]: 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.** [34]: The dataset in question has been previously discussed in section 3.1, where it was noted that it offers a highly comprehensive analysis of cross-session evaluation. We decided to incorporate this dataset into our study 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.

4.1.2 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** [48]

   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, specifically 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 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** [49]: 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) [73], the N-Back Task [74], the Multi-Attribute Task Battery (MATB) Task [75], and the Flanker Task [76]. These tasks were specifically designed to elicit various cognitive states. The authors employed a 64-electrode Ag-AgCl ActiCap (Brain Products Gmbh)

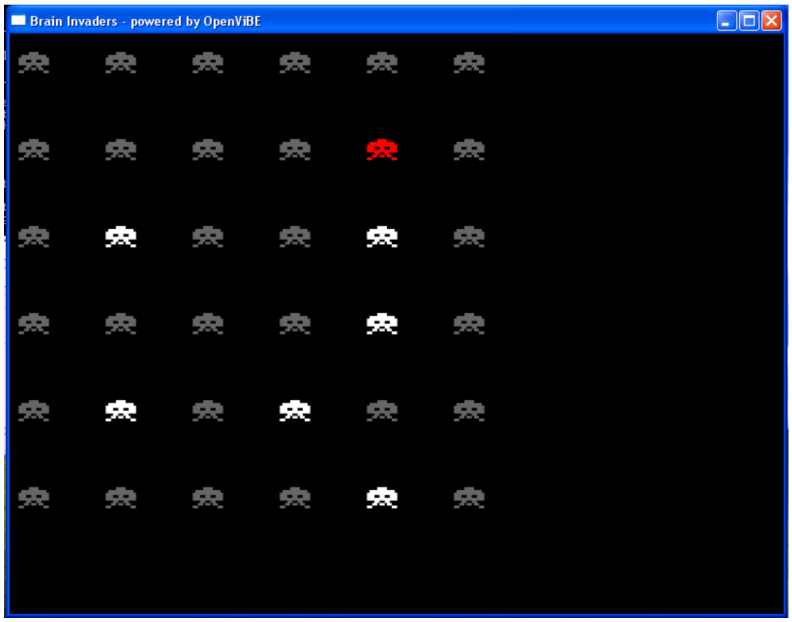


Figure 4.1: Brain Invaders user interface at the game's introductory stage [40]

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 conducted by Eriksen and Eriksen (1974) [76] 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 central target (congruent condition) or in the opposite direction (incongruent condition). The experimental procedure for the flanker task is illustrated in Figure 4.2. Upon the conclusion of the trial, the participant is provided with feedback regarding the outcome of their performance, specifically 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** [68]: This dataset has been used in various brainwave-based recognition studies such as [16, 23, 77]. 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

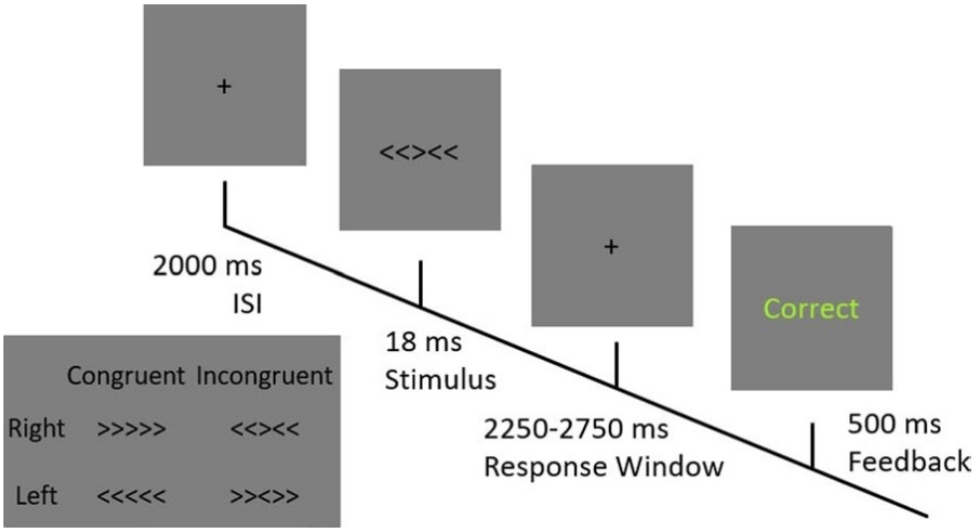


Figure 4.2: Flanker Task: After an Inter stimulus of 2000 ms, one of four possible stimuli (bottom left) is displayed for 18 ms. Participants then have between 2250 and 2750 milliseconds to respond before receiving 500 milliseconds of feedback [49]

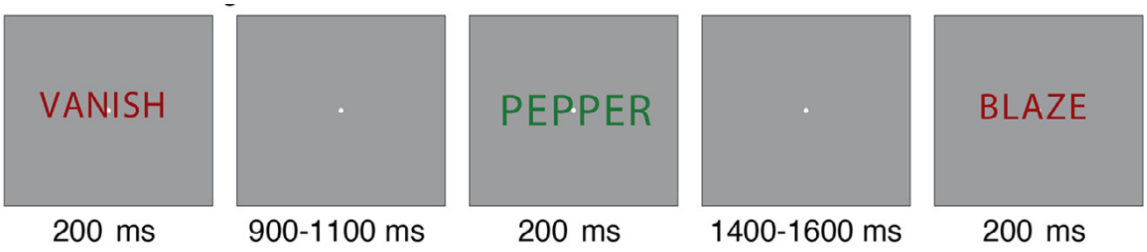


Figure 4.3: The experimental setup for the ERPCORE: N400 task involving a specific configuration designed to elicit and measure the N400 component. [41]

related or unrelated to the prime word. The experimental setup for ERPCORE: N400 is depicted in Figure 4.3.

4. **Mantegna et al. (mantegna)** [51]: The dataset utilized in this study is derived from EEG investigations, specifically 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 Event-Related Potentials (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 a total of 135 sentences that rhyme, with either congruent or incongruent endings. The primary objective of this experiment is to elicit N400 ERPs.

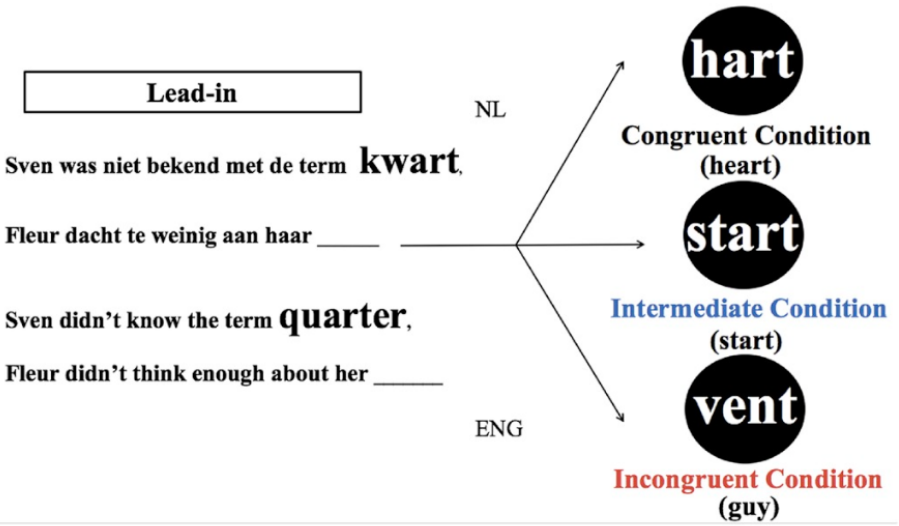


Figure 4.4: Three alternative target words were selected for each sentence pair. In the congruent case, there was overlap in the rhymes, and the target word was easy to guess based on its meaning. In the middle case, there were words that rhymed with the target word, but the target word was not predictable based on its meaning. There was no rhyme overlap in the incongruent case [51]

Figure 4.4 illustrates an instance of a sentence pair.

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

Our benchmarking framework is organized into five modules: datasets, paradigm, evaluation, pipeline, and analysis. The next section contains detailed information on all of the modules described above. 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 chapter 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

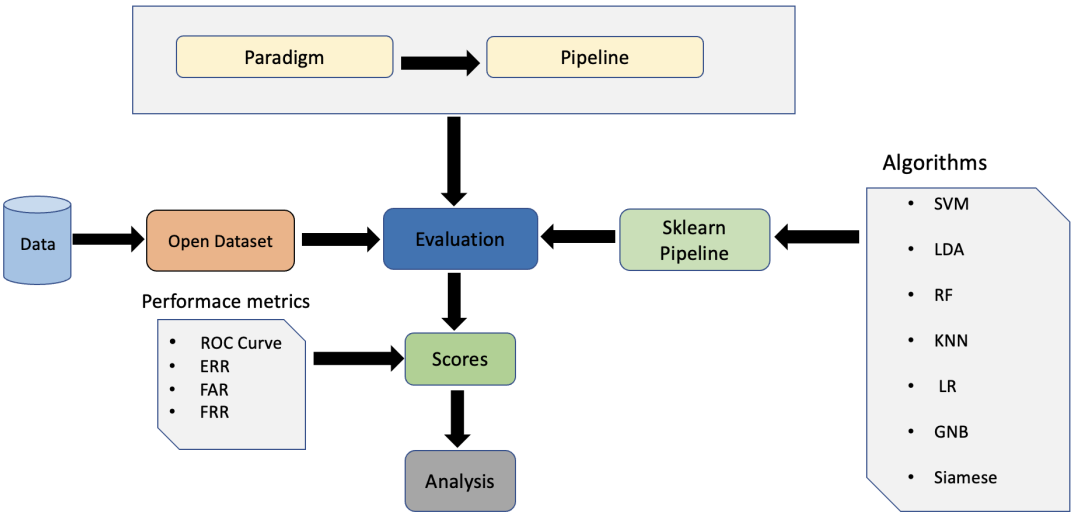


Figure 4.5: Overview of benchmarking suite

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 important to acknowledge that the execution of the procedures above necessitates the utilization of the Scikit-Learn [78] pipeline. This pipeline facilitates the execution of various python pipelines comprising distinct datasets, paradigms, feature extraction methods, and algorithms.

5 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, [17]. 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 common pipeline will be implemented as a wrapper around the scikit-learn pipeline library, known for providing 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 4.1.2, 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, particularly when performing various analyses and evaluating new algorithms. To overcome these obstacles and enhance the efficiency of accessing the datasets, a python interface is developed. The purpose of this interface is to optimise and enhance the process of 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 [79]. 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 [17]. 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.

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 exhibited congruity in their measurement scale. However, the BrainInvaders15a dataset was originally measured on a micro voltage scale. To ensure consistent data scalability across all four datasets, we rescaled the EEG data of BrainInvaders15a.

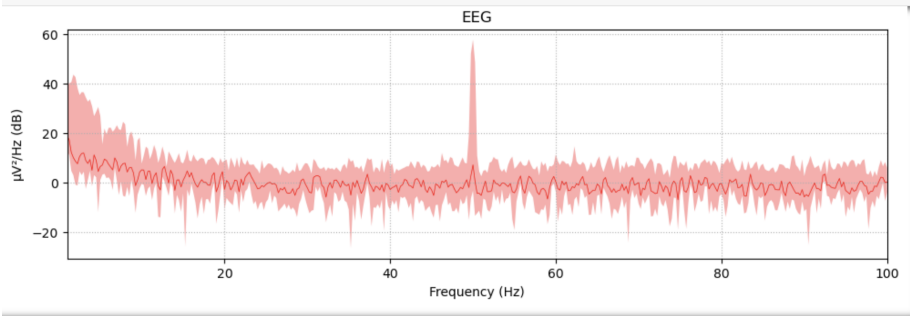
- In accordance with the information presented in Section 4.1.2, it has been established that the Mantegna dataset consists of three distinct categories of events, namely congruent, intermediate, and incongruent. According to the research conducted by Mantegna et al., [51], it was observed that both 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 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 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.

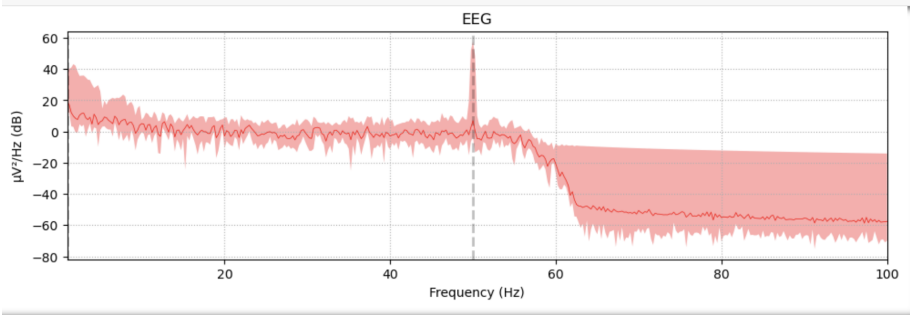
5.2 Pre-Processing

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 [17]. 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 the objective of eliminating the 50Hz line noise and filtering out signals originating from flat channels with frequencies below 1 Hz. Figure 5.1 (a) depicts the unprocessed raw signal, which exhibits a significant signal strength at 50Hz due to line noise. On the other Figure 5.1 (b) illustrates the consequences of implementing a bandpass filter, revealing a noticeable stabilization in the raw signals after the filtration of the data.

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



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

period, which ranged from -200 to 0 ms. Baseline correction is employed to reduce the drifting effects of DC offsets [23]. 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 certain large 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 in order 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 5.2, thresholds of 100μV and 150μV resulted in the exclusion of over 80% of the total EEG data for datasets such as BrainInvaders15a and COGBCI. Consequently, we sought to identify an alternative approach that would effectively eliminate the presence of noisy data while minimising the loss of valuable EEG data.

We implemented a more sophisticated approach to eliminate noisy data by utilizing the *Autoreject* [80] 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 artefacts. We were unable to 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

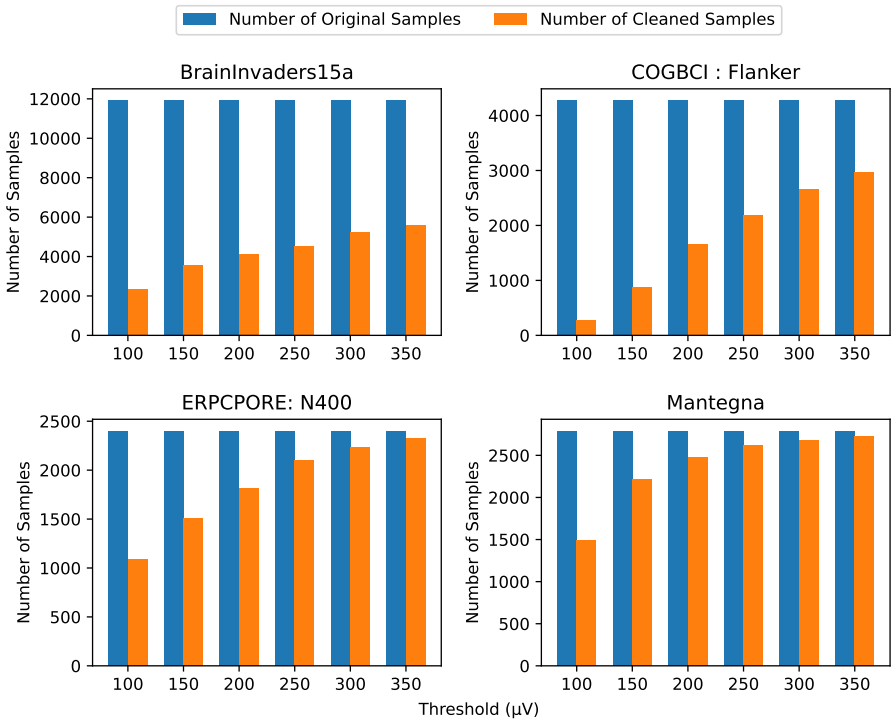


Figure 5.2: Actual number of epochs versus the number of cleaned epochs after conducting epoch rejection with various thresholds. X-axis depicts the sample count whereas y-shows the threshold values in μV

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 5.3 depicts the mean evoked potentials 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

After pre-processing the EEG data, the next step is to acquire discriminant features that represent and encode a user's mental activity using the clean EEG signal [27]. We surveyed many studies presented for brainwave authentication. We found that the Autoregressive (AR) model and Power spectrum (PS) are some of the most widely used methods for extracting features in time and frequency domains [16, 8]. Further, AR's potential to reveal particular inherent characteristics of the EEG signal within a single channel and PS'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.

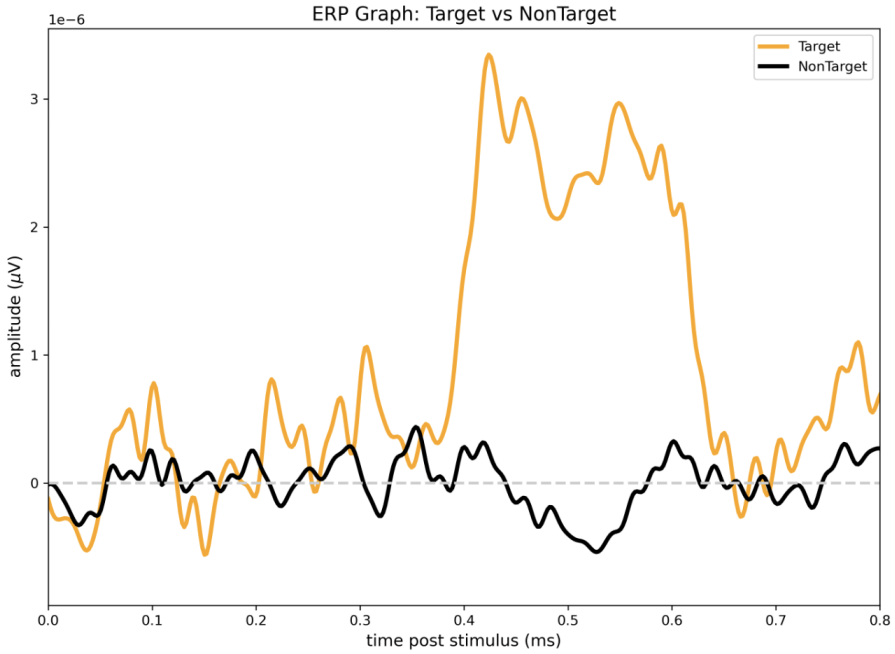


Figure 5.3: The Averaged Evoked Potentials exhibit an increase in amplitude ranging from 250 to 400 milliseconds, which can be attributed to the implementation of the oddball paradigm

- **AR Coefficients**: The AR model is fitted using pre-processed epochs, time series data lasting for 1 second [16]. 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 [81].

  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. It is important to acknowledge that the AR coefficients are computed individually for each channel within the signal. 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).

- **Power Spectrum**: The Power Spectrum (PS) of each epoch is computed across different frequency bands, namely low (1-10 Hz), alpha (10-13 Hz), beta (13-30 Hz), and gamma (30-50 Hz), utilizing the Welchs periodogram algorithm [16]. Welch's periodogram is used to compute the Discrete Fourier Transform (DFT) results [27]. In our study, first the power spectral density (PSD) for each frequency point in 1-second epoch is 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 [82]. We then computed the average PS within the specified frequency ranges. This allowed us to determine the average power spectrum (PS) of the low, alpha, beta, and gamma frequency bands. Similar to the AR features, the PS 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 through the utilization of 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 [16]. Traditional classifiers like LDA, SVM, RF, GNB, 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 both training and testing [16]. The total number of repetitions conducted was 10, and the results of the evaluation metrics are obtained from 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, 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" [16]. 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 simpler

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

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 completely 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 GroupK-fold 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 9 identified as unknown attackers per model [16]. 9 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 [34]. It is also the underlying cause of the significant scarcity of datasets. In our study, we have s single multi-session dataset out of all the open datasets, i.e., COGBCI, which contains 3 recorded sessions at an interval of 7 days. Each session consisted of a full 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 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 [17] 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 sessions of data to ensure an adequate number of 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

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

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, with the exception 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 that were 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 [23], 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 Networks (SNN) is a highly effective deep learning (DL) method for implementing similarity metrics for brainwave authentication, employed already in studies such as [23, 35, 39]. This type of learning allows for accurate predictions after training the network with only a few samples [83], 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 SNN method adopts 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 [84]. Among these, the *triplet loss* approach is particularly well-suited for biometric recognition [23]. As a result, our study utilizes SNN with three CNN branches, which are trained using a triplet loss function. As explained by Schroff *et al.* [83] 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 5.4 illustrates the learning procedure in SNN using triplet loss function.

The Siamese architecture proposed by Fallahi et al. [23] was implemented in our study.

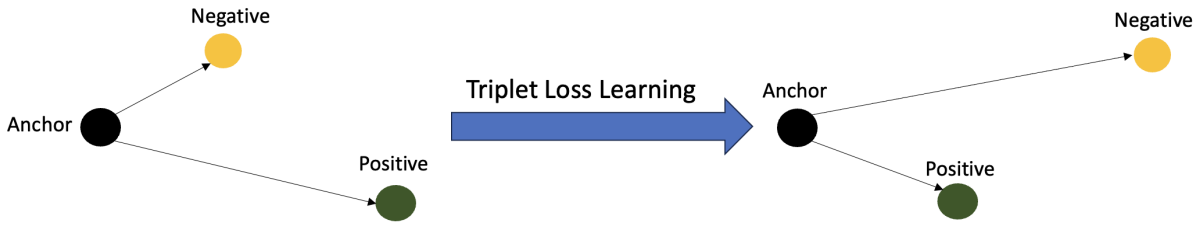


Figure 5.4: Triplet loss function minimizing the euclidean distance between the anchor and positive embeddings while simultaneously maximizing the distance between the embeddings of two different individuals, specifically the anchor and negative embeddings

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 [83] 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 SNN 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 [23]. 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 SNN 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. The process of evaluating open-set scenarios involves the comparison of each subject's brain sample with the samples contained within the testing set. 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 (SNN) is comparable to the approach employed for the cross-session assessment in supervised learning-based classification tasks. Therefore, LeaveOneGroupOut cross-validation was employed for grouping the sessions into training and testing set in each round of cross-validation.

Threat Case Scenarios: A close-set scenario is attained by employing the LeaveOneGroupOut cross-validation technique we discussed previously. 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 sessions data is partitioned into training and testing sets using the LeaveOneGroupOut method, where subjects from the enrollment database that are being verified are excluded. The enrollment database in this case 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 4.5. 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 *subjects, 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 *subjects, 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

6.1 Evaluation Metrics

In this section, we examine the performance of SOA and well as deep learning algorithms on the four open datasets. In section 6.2, evaluation results are presented on both single and cross-session considering seen attackers(i.e., close-set scenario) as well as unseen attackers (i.e., open-set scenario). In addition, we compare the results obtained from within-session and cross-session evaluations to determine the impact of EEG variability on the authentication process. The performance of authentication algorithms varies based on other factors, such as the sample size of datasets, the duration of epochs, the AR order, and the windowing size used to calculate PSD. To gain a thorough comprehension of the system, a comprehensive investigation is conducted to examine the impact of these factors on the overall effectiveness of our authentication systems.

It is essential to compare the performance of the algorithms with appropriate metrics because it is seen that a lot of studies presents 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 in usually imbalanced since most researchers build the single classifier for individual subject. Accordingly, that single user is labeled "authenticated" and the remaining users are marked "rejected" for training and testing the authentication model. As a result, the model is trained on more on the negative samples. This makes it easy for the model to identify rejected users. Therefore, the high value of accuracy represents a biased assessment of the model's performance because of the skewness present in the training data. Hence, we choose not to focus on common metrics like accuracy in our study. Instead, we employed the 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 [16].

6.2 Results

This section provides a complete analysis of the outcomes of our investigation, including a thorough comparison of the datasets and evaluation system in both closed and open-set scenarios.

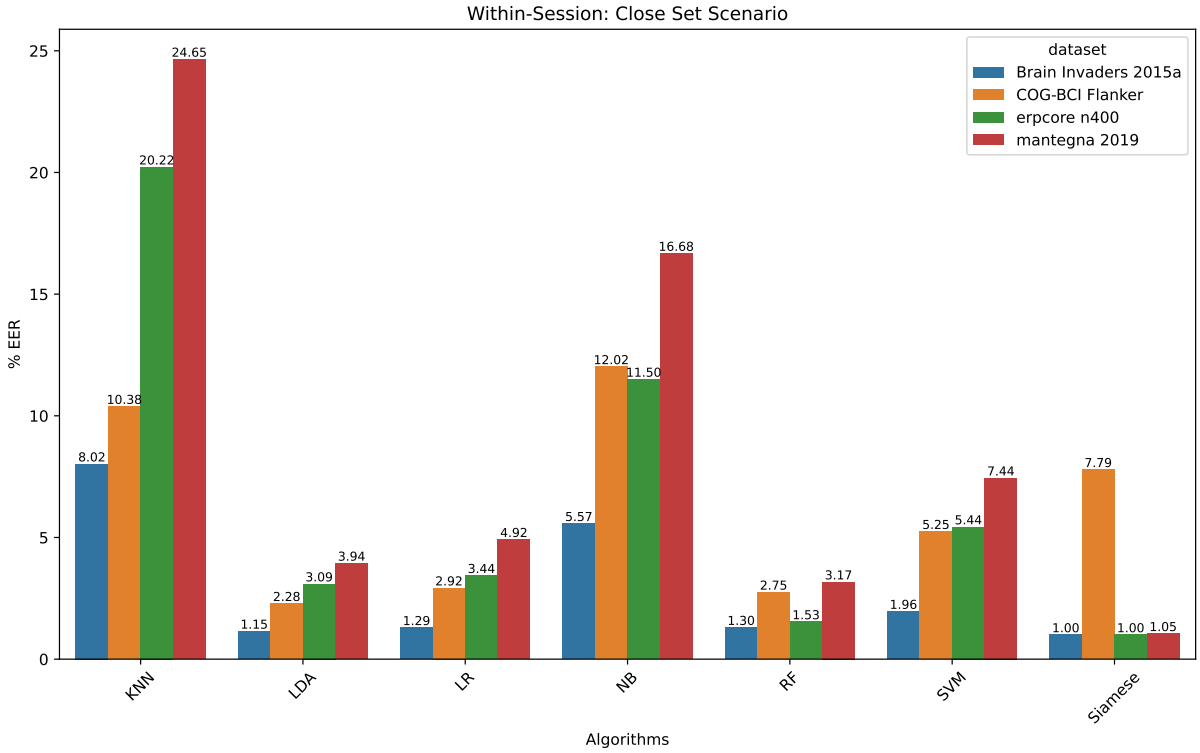
6.2.1 Within-Session Evaluation Results

This section presents the results of our study in the single session setup. 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%. 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 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 GNB 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.

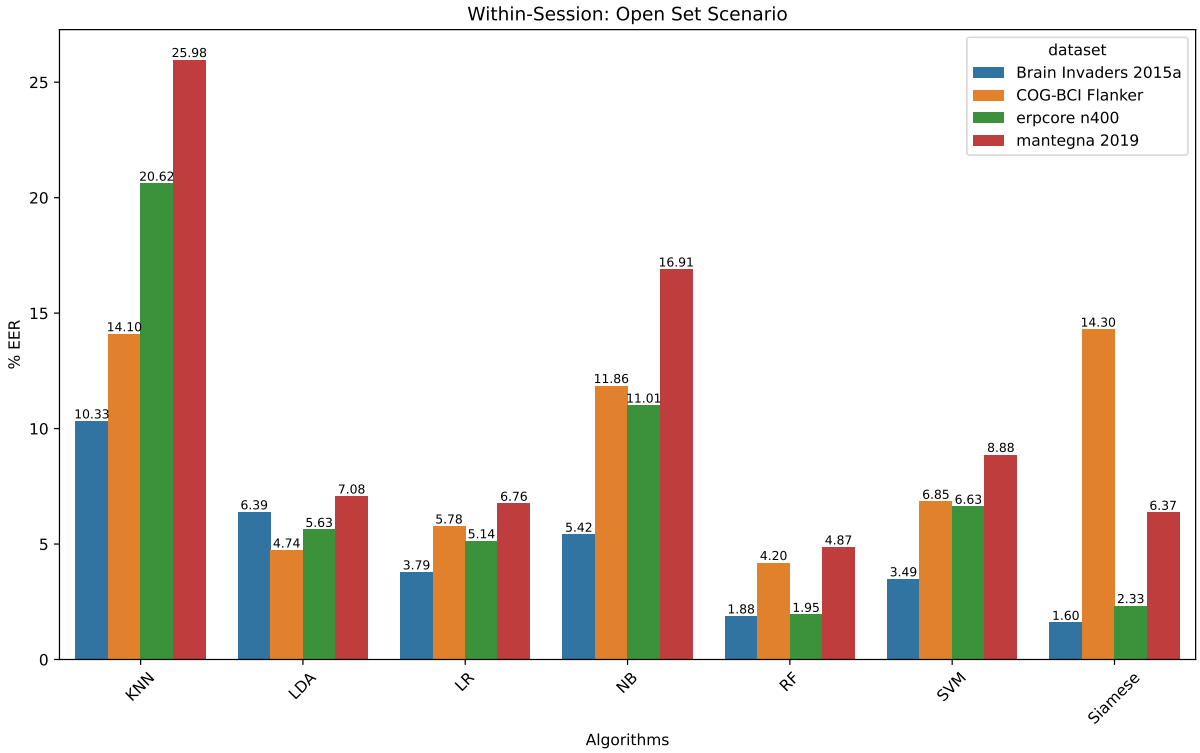
Comparison between datasets: BrainInvaders15a performs better than the other datasets, as seen by the data presented in Figure 6.1. The achieved EER for BrainInvaders15a demonstrates a notable decrease across all classifiers, except the LDA classifier, when used in the open-set scenario. In this case, the EER of BrainInvaders15a is higher than that of 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 [16]. In section 6.2.5, 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.

Comparison between close-set and open-set scenarios: It was hypothesized that 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 Fig 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 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 [16]. As outlined in section 5.4.1, the evaluation strategy of close-set requires training the authentication model with N-1 attackers



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



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

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

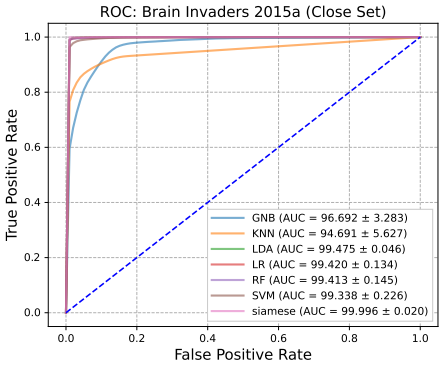
Table 6.1: Average FRR at 1% of FAR for the four datasets in within-session evaluation scheme, comparing different classifiers and threat case scenarios. The values in the table are shown in percentages.

Dataset	Scenario	LDA	SVM	LR	RF	KNN	GNB	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

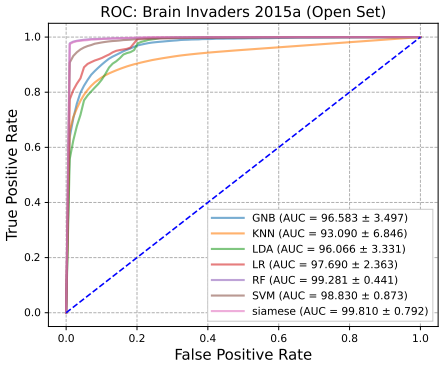
where N represents the total number of users. In contrast, the classifiers in an open-set scenario learn from approximately three-quarters of the N-1 attackers. Consequently, in close-set settings, the attacker spaces are more significant than in open-set, thereby enabling more effective training of 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: Traditional machine learning algorithms such as LDA, SVM, LR, RF, KNN, and GNB 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 they are 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 [85, 83, 86] and Brainwave authentication [23, 39].

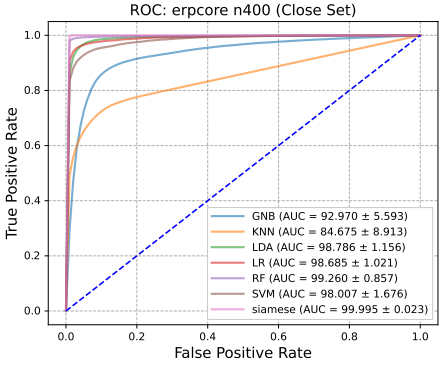
The results obtained in our study have also demonstrated Siamese networks as the best-performing algorithm among all the traditional classifiers. The ROC-Curves of the four datasets are presented in Figure 6.2, showcasing the operational capabilities of the traditional and deep learning authentication models in closed-set and open-set settings. The Area Under Curve (AUC) under ROC-Curve represents a single value representing the system's ability to differen-



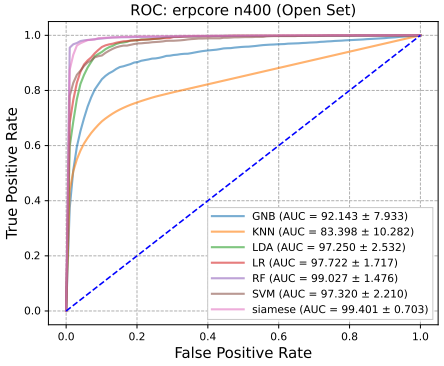
(a) ROC: BrainInvaders15a Close-Set



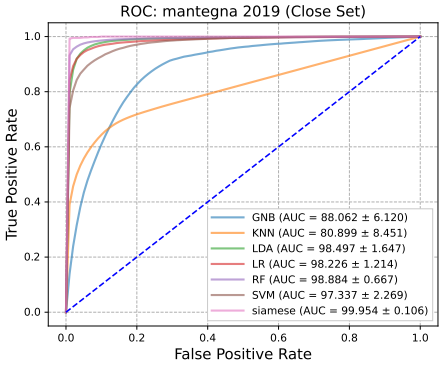
(e) ROC: BrainInvaders15a Open-Set



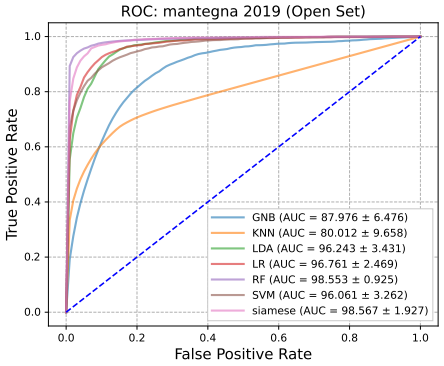
(b) ROC: ERPCORE N400 Close-Set



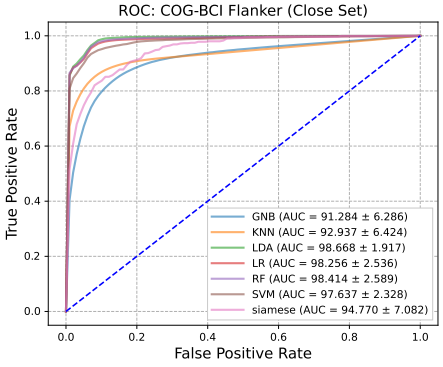
(f) ROC: ERPCORE N400 Open-Set



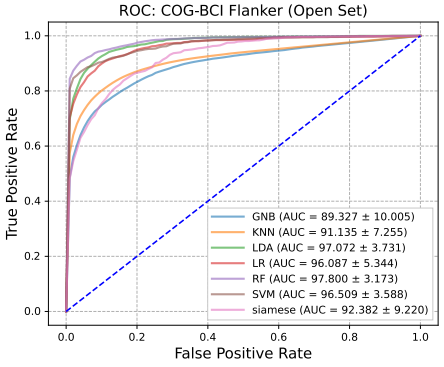
(c) ROC: Mantegna 2019 Close-Set



(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

tiate between genuine users and imposters [16]. A higher AUC implies an improved performance as it indicates the system has a higher True Positive Rate (TPR) value and a lower False Positive Rate (FPR). The Siamese Networks have a higher AUC score than traditional 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 outperforms 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 indicate that Siamese networks are superior to other classifiers in attaining EERs for the BrainInvaders15a, ERPCORE: N400, and Mantegna2019 datasets.

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 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 [87]. On the 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 increase 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, hence 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.

The findings presented in Table 6.1 indicate that the most optimal setup is achieved when employing Siamese Networks on dataset BrainInvaders15a for authentication. Remarkably, the closed-set scenario achieves FRR of just 0.05% at threshold of 1% FAR. The authentication system demonstrates high usability and resilience in open-set situations, as seen by a FRR value of only 2.07%. 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 Cross-Session Evaluation Results

This section focuses on the outcomes derived from conducting evaluations throughout multiple sessions. The cross-session review is run solely on the COG-BCI Flanker dataset, which represents the only dataset available. Figure 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 determined by the cross-session evaluation. The performance of all the classifiers in both the close-set and open-set scenarios was observed to be comparable in cross-session. Therefore, the findings shown in both figures pertain to the close-set scenario. As depicted in Figure 6.3, The attained EER for the traditional 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 Receiver Operating Characteristic (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

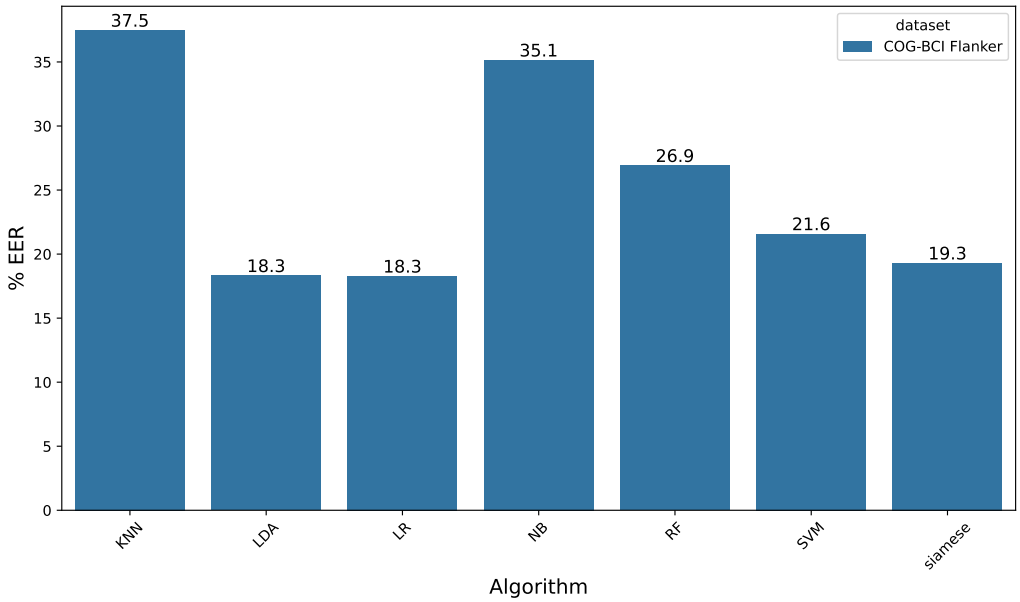


Figure 6.3: Average EER for cross-session evaluation on COG-BCI Flanker dataset, comparing performance across different authentication algorithms

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 evaluation, indicating that the GNB 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 findings from the cross-session evaluation conducted on the COG-BCI Flanker dataset have provided valuable insights into noteworthy observations. We did not achieve the best possible results in cross-session. One possible explanation for the sub-optimal performance in cross-session evaluation is the increased variability and unpredictability due to the time duration between the enrollment and authentication process. As mentioned in section 5.4.2 and 5.4.1, we performed enrollment for each subject by utilizing the brain sample acquired in two sessions and performed authentication on the remaining session's EEG data. This temporal gap may have resulted in changes to the EEG signals, making it challenging for the classifiers to recognize users consistently across sessions. The other factors which could impact the results in cross-session settings are the electrode resetting across the sessions, variations in human brain states, and template ageing [35]. Moreover, the performance of traditional algorithms appears to have been impact more prominently in cross-session evaluation than Siamese Networks.

The results of our cross-session setting align with Arnau-González *et al.* [88] 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 Power Spectral Density across Theta, Delta, Alpha, Beta, and Gamma 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

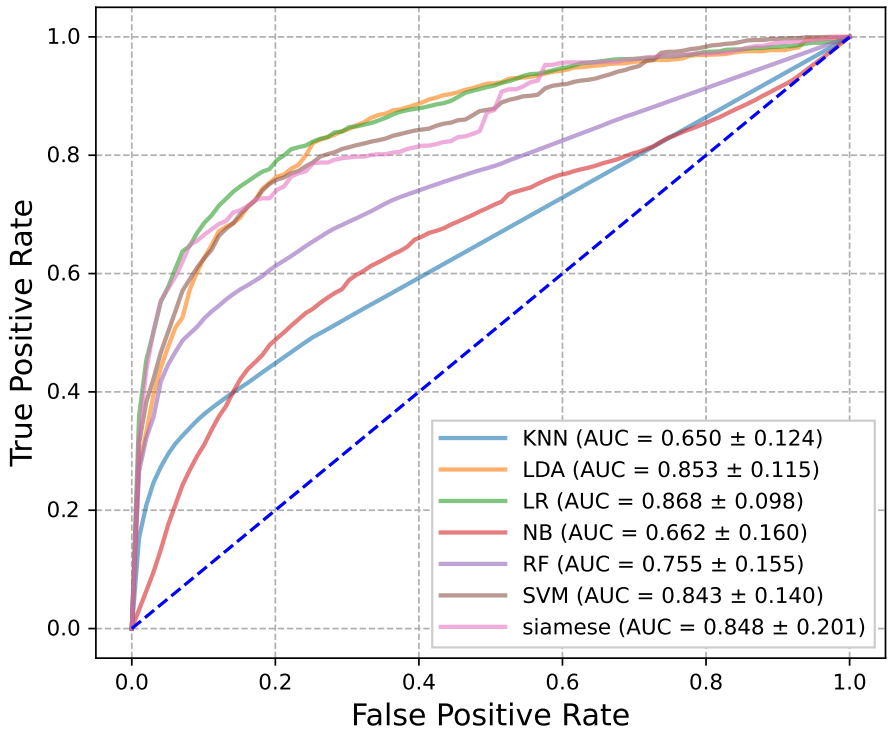


Figure 6.4: Performance comparison of the dataset COG-BCI Flanker in cross-session evaluation scheme. ROC Curves are depicted for different classifiers in close-set attacker scenario.

a cross-session scenario. The accuracy reached in cross-session evaluation was 79%, a substantial decrease compared to the 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 Within-Session Vs Cross-Session

One of the main objectives of this thesis is to comprehensively study the impact of EEG variability across single and multi-session settings. We conducted thorough evaluations of our authentication models in both within-session and cross-session scenarios, as detailed in the preceding sections, and 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. In this section, we will compare the results obtained from both the within-session and cross-session evaluation schemes. According to Table 46.2, the results of the multi-session (cross-session) evaluation are significantly poorer than 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 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 which have comparable EER and FRR at 1% FAR in within-session also experiences 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 increase 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 compared to traditional 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 traditional classifiers. The observation as mentioned above implies that Siamese Networks has the ability to 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 traditional classifiers, thereby presenting encouraging prospects for further investigation in EEG-based authentication systems.

Table 6.2: Average Performance of classifiers on the COG-BCI Flanker [49] dataset, comparing Within-Session and Cross-Session Evaluation

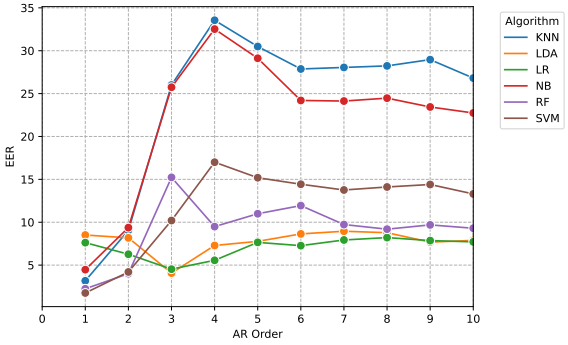
Metric	LDA	SVM	LR	RF	KNN	GNB	Siamese
Within-Session							
%EER	2.28 ± 2.55	5.25 ± 3.50	2.92 ± 3.28	2.75 ± 3.28	10.38 ± 6.92	12.02 ± 7.24	7.79 ± 6.54
FRR at 1% FAR	14.46	19.19	14.05	13.33	32.88	59.56	45.07
Cross-Session							
%EER	18.32 ± 11.47	21.58 ± 13.18	18.28 ± 11.47	26.91 ± 13.67	37.47 ± 10.34	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

6.2.4 Impact of Feature Extraction on Performance

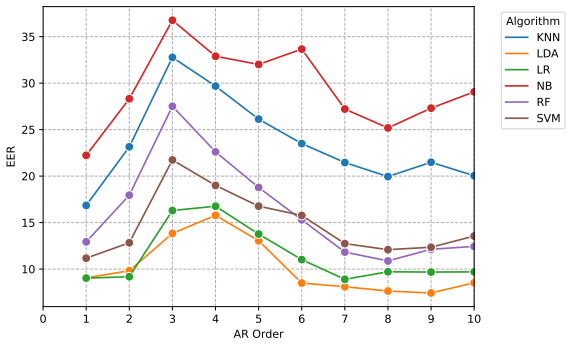
Feature extraction plays a crucial role in developing a resilient EEG-based authentication system. In this section, we will assess the impact of different feature extraction steps on the performance of all datasets. In our study, we extracted features in the time domain by estimating the AR coefficients and in the frequency domain by calculating Power Spectral Density. These features were passed as input to the classifiers for training and testing. Siamese Networks can acquire discriminant patterns from time series epochs data, as described in Section 5.4.2. Consequently, their performance is independent of the AR and PSD features. Accordingly, this section will evaluate the efficacy of traditional classifiers on the four datasets. For each dataset, we extracted 17 unique feature sets (10 AR, 1 PSD, and 10 combinations of both AR and PSD) and evaluated their classification performance. These feature sets included AR coefficients (order=1,2,3,4,5,6,7,8,9,10) of the epoch, PSD of the epoch, and a combination of the two.

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.

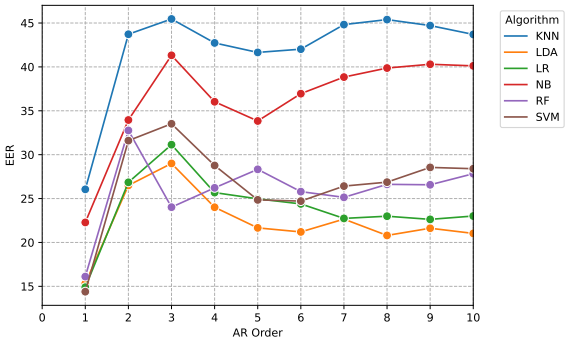
As depicted in Figure 6.6, the classifiers' performance utilizing only PSD features outperforms that evaluated through AR features. The data in Figure 6.6 indicates that the EER remains



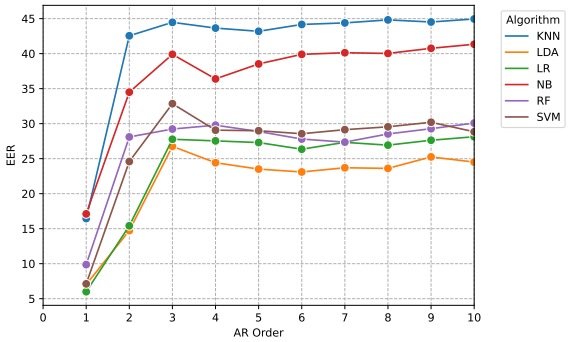
(a) Effect of AR features with orders (1 to 10) on performance of the dataset BrainInvaders15a



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

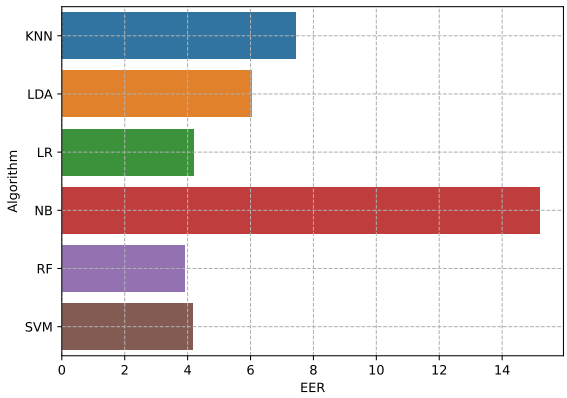


(c) Effect of AR features with orders (1 to 10) on performance of the dataset ERPCORE: N400

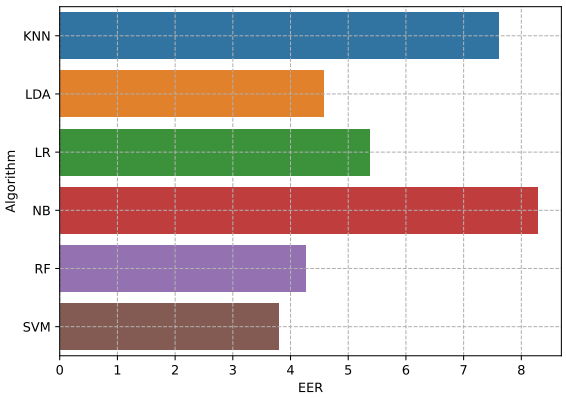


(d) Effect of AR features with orders (1 to 10) on performance of the dataset Mantegna2019

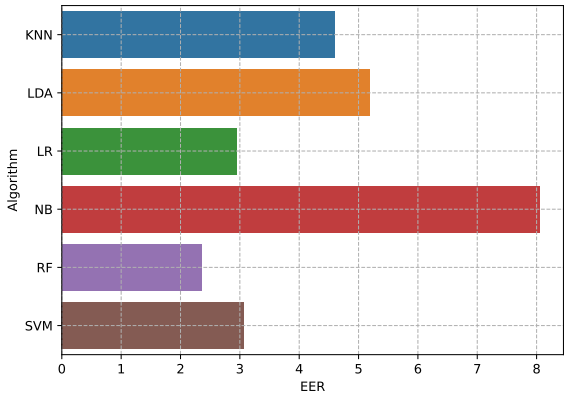
Figure 6.5: Impact of Auto Regressive(AR) Features on the performance of the datasets. Figures (a), (b), (c) and (d) depicts the change in the EER of the traditional classifiers for the datasets BrainInvaders51a, COG-BCI Flanker, ERPCORE: N400 and Mantegna2019 respectively.



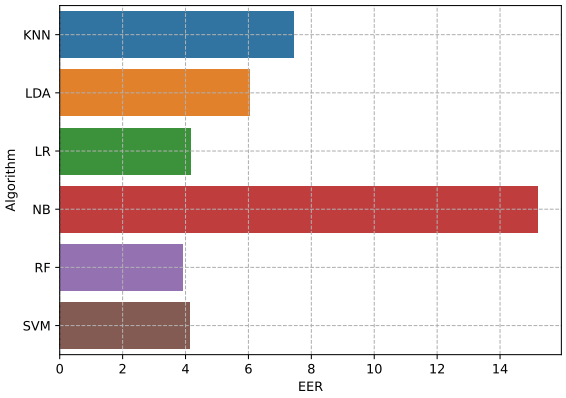
(a) Effect of PSD features on the performance of the dataset BrainInvaders15a



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

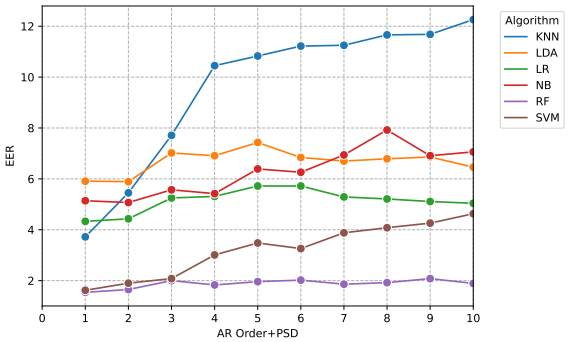


(c) Effect of PSD features on the performance of the dataset ERPCORE: N400

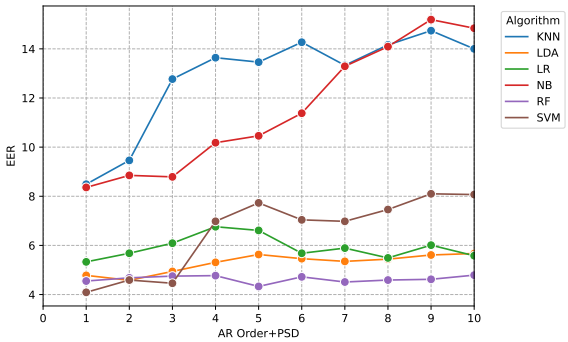


(d) Effect of PSD features on the performance of the dataset Mantegna2019

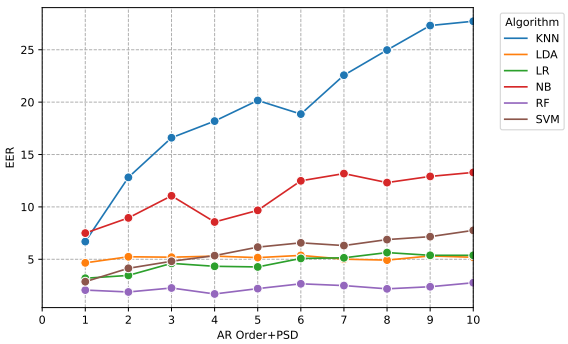
Figure 6.6: Impact of Power Spectral Density (PSD) Features on the performance of the datasets. Figures (a), (b), (c) and (d) depicts the change in the EER of the traditional classifiers for the datasets BrainInvaders51a, COG-BCI Flanker, ERPCORE: N400 and Mantegna2019 respectively.



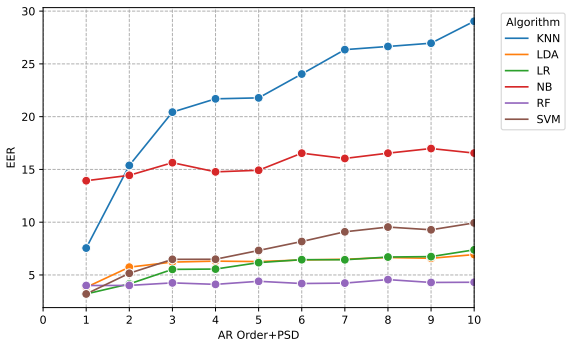
(a) Effect of the combination of AR (from order 1 to 10) and PSD features on the performance of the dataset BrainInvaders15a



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

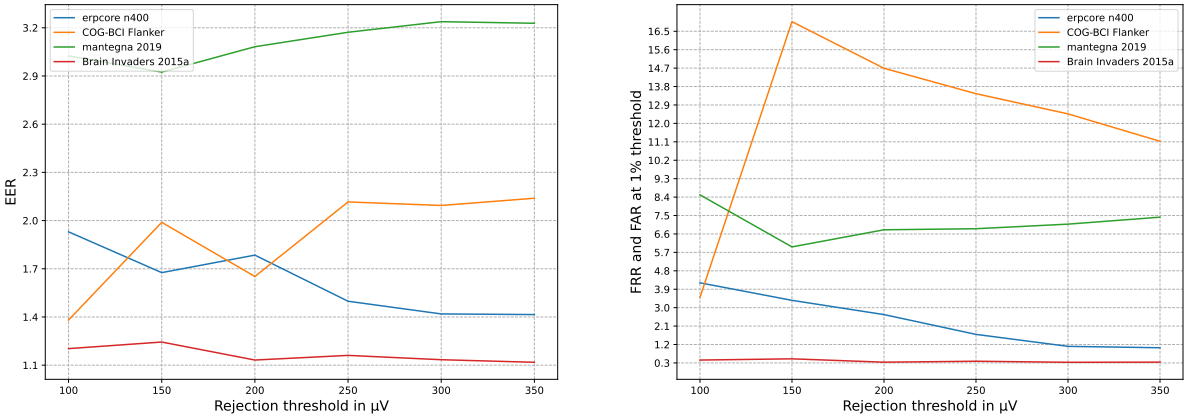


(c) Effect of the combination of AR (from order 1 to 10) and PSD features on the performance of the dataset ERPCORE: N400



(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.



(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. Figure 6.6 shows the EER and FRR at 1% FAR for classifier RF.

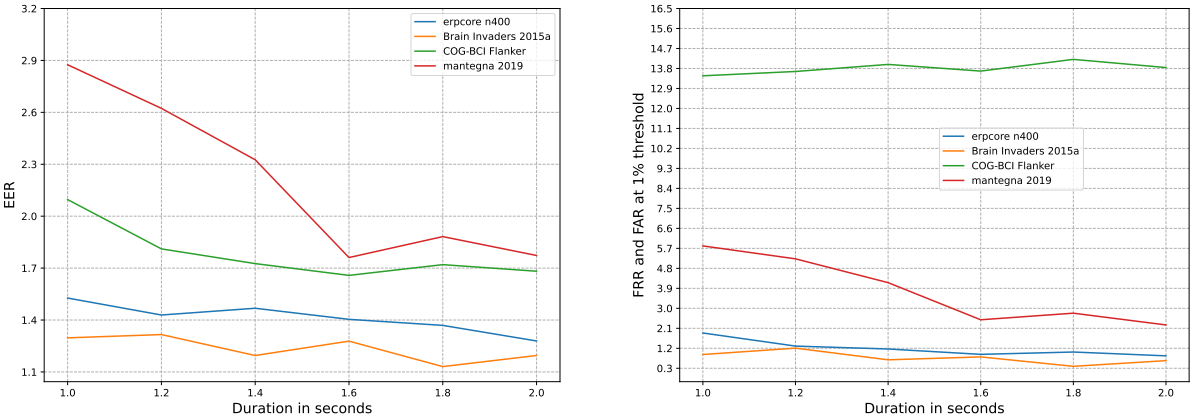
below 16% across all classifiers and datasets. Conversely, the utilization of AR features, as shown in Figure 6.5, results in a notably higher EER, reaching up to 45% for the KNN classifier on datasets like ERPCORE: N400 and Mantegna2019. This observation indicates that frequency domain features capture distinct EEG patterns across subjects more effectively than time domain features. It is noteworthy that, once again, the RF classifier stands out as the best performer.

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 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.

6.2.5 Effect on Performance due to Epochs Rejection

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 previously mentioned in section 5.2, the artefact rejection process involves the utilisation of the peak to peak rejection approach. The observation was made that varying rejection thresholds have an impact on the quantity of the dataset that triggers an alert. Consequently, this section will evaluate the influence of various rejection criteria on the efficacy of the best performing traditional classifier, namely RF, as well as the deep learning approach known as Siamese Networks. The evaluation will be performed on the four datasets using the within-session evaluation scheme within the context of the close-set scenario.

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. Conversely, a constant reduction in FRR at a FAR of 1% was observed as the rejection threshold increased from 100μV to 150μV. This suggests a positive correlation between the number of



(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. Figure 6.6 shows the EER and FRR at 1% FAR for classifier RF.

samples and the classifier's performance on ERPCORE: N400, indicating that as the number of samples increases, the classifier's performance improves. 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 performance of the classifier 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 and experimental conditions. In order to enhance the flexibility of our framework, we have devised a design that allows researchers the choice to specify their own 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 approach in diverse EEG authentication scenarios.

6.2.6 Effect on Performance due to Epoch Duration

In this section, we analyze the effect of different epochs duration on the performance of our authentication system using RF classifier and same pre-processing and feature extraction pipeline. The durations of the epochs were 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 this 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 scenario.

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. In contrast, ERPCORE: N400 and COG-BCI Flanker datasets exhibit only marginal shifts in performance. This is evidenced by the nearly unchanging EER and FRR values at 1% FAR as epochs duration increases.

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.

7 Discussion

7.1 Comparison with existing works

8

Conclusion and Future Works

8.1 Conclusion

8.2 Limitations

8.3 Future Works

Bibliography

[1] 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.

[2] 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.

[3] 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.

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

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] Joseph N. Mak and Jonathan R. Wolpaw. Clinical applications of brain-computer interfaces: Current state and future prospects. *IEEE Reviews in Biomedical Engineering*, 2:187–199, 2009.

[25] Xiaotong Gu, Zehong Cao, Alireza Jolfaei, Peng Xu, Dongrui Wu, Tzyy-Ping Jung, and Chin-Teng Lin. Eeg-based brain-computer interfaces (bcis): A survey of recent studies on signal sensing technologies and computational intelligence approaches and their applications. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 18(5):1645–1666, 2021.

[26] Francisco Velasco-Álvarez, Salvador Sancha-Ros, Esther García-Garaluz, Álvaro Fernández-Rodríguez, M Teresa Medina-Juliá, and Ricardo Ron-Angevin. Uma-bci speller: an easily configurable p300 speller tool for end users. *Computer methods and programs in biomedicine*, 172:127–138, 2019.

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

[30] 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.

[31] 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.

[32] 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.

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

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] Barjinder Kaur, Dinesh Singh, and Partha Pratim Roy. A novel framework of eeg-based user identification by analyzing music-listening behavior. *Multimedia tools and applications*, 76(24):25581–25602, 2017.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

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

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

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

[60] 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.

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

[62] 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.

[63] 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.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] 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.

[69] 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.

[70] 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.

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

[72] 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.

[73] 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.

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

[75] 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.

[76] 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.

[77] 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.

[78] 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.

[79] 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.

[80] 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.

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

[82] 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.

[83] 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.

[84] 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.

[85] 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.

[86] 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.

[87] 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.

[88] 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.

A

Appendix

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
```

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 COGBCIFLANKER 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
```