“A Biometric Person Recognition System based on Brainwave (EEG) signals”

Alvin Jagle
2449685

MSc in Data Science
College of Science and Engineering

Swansea University
Prifysgol Abertawe

Department of Computer Science
Swansea University

September 30, 2025

Declaration

This work has not been previously accepted in substance for any degree and is not being con- currently submitted in candidature for any degree.

Signed 

Date **30-09-2025**

Statement 1:-

This thesis is the result of my own investigations, except where otherwise stated. Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed 

Date **30-09-2025**

Statement 2:-

I hereby give my consent for my thesis, if accepted, to be made available for photocopying and inter-library loan, and for the title and summary to be made available to outside organisations.

Signed 

Date **30-09-2025**

ABSTRACT

EEG-based biometric identification is a new method for using people's brainwave patterns as unique identifiers. In this thesis, we evaluate a biometric identification system based on electroencephalograms (EEGs) with the primary focus on improving accuracy and reliability across recording sessions. EEG recordings from nineteen channels were sectioned into overlapping five-second time-windows to capture the time-varying patterns of brain activity. The features selected from the EEG segments were spectral measures, specifically the absolute and relative band power within standard EEG frequency bands. Two classification algorithms were trained to distinguish people based on EEG features, a support vector machine (SVM) using a radial basis function (RBF) kernel and a random forest.

The system's performance was assessed in six different protocols to understand how generalizable the system may be. The protocols employed included standard random k-fold cross-validation, a random 60/20/20 train-validation-test split, and more-stringent session-disjoint splits that separated training and testing data based on recording session. Identification accuracies associated with the standard random splits were exceptionally high (about 98–99%), but these metrics are probably inflated by session-specific patterns and will not generalize to new sessions. The session-aware evaluation protocol (with no overlap between training and test sessions) produced about 92% accuracy, which is a more accurate assessment of performance than do the random splits. These results suggest that EEG-based biometrics can reliably differentiate individuals and underscore the necessity of employing session-disjoint evaluation protocols to derive realistic performance estimates.

Acknowledgements

Most importantly, I want to give thanks to my Lord and Savior Jesus Christ, who provided guidance, wisdom, and strength throughout this perspective. I can credit every success of this work to his grace and blessings, because I would not have made it through each trial, without his strength, that confronted me throughout this research.

I also would like to express my heartfelt thanks to my supervisor, Prof. Scott Yang Yang, whose valuable guidance, help, and tolerance I would like to genuinely appreciate. His expertise and criticism significantly enhanced the substance of this dissertation, and he has been the key to the development of my research career.

I also want to express my appreciation to the faculty members and staff who contributed to my graduate education. Their teaching and support have helped me build a strong foundation for my work, and their support in an academic environment reinforced my confidence in pursuing my research.

I am particularly grateful to my peers and lab mates for their fellowship, teamwork, and the many discussions we had where we shared our knowledge. Their support has made the trials of this journey easier and enriched my graduate experience.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1 Background and Motivation: -

In the last 10 years, systems to interface the brain with a computer, mainly based on electroencephalography (EEG), have rapidly developed, allowing for the typing of faster-than-real-time interpretations of cognitive and affective states from brain signals. This development is enabled by new hardware and algorithms. On the hardware side, improved consumer-grade EEG sensors are becoming more available and reliable, which makes EEG collection possible in non-clinical laboratory settings (Wilaiprasitporn, 2020). Within certain contexts, for example education, virtual games and drowsiness detection in drivers, these low-cost, lightweight devices have proved effective for continuous brain monitoring (Wilaiprasitporn et al., 2020). On the algorithm side, machine learning techniques, especially deep learning have increased EEG-based task classification performance accuracy by automatically learning rich and complex spatiotemporal features from the multichannel time-series data (Lotte et al., 2018; Yang et al., 2022). As a case in point, one of the emerging applications of convolutional neural networks and their variants continues to show promise for extraction of spectral and temporal patterns, with recent architectures extracting multiple feature types (e.g., Yang et al., 2022). These paradigms have produced positive results for mental workload measures, emotion measurements, and regulation of motor imagery and demonstrate the promising products of BCIs development for both neuroergonomics and assistive technology (e.g., Becerra et al., 2025; Apicella et al., 2024).

Nonetheless, a significant problem remains: stable generalization of EEG models across recording sessions and examinees. EEG signals are well known to be non-stationary – the properties of the signal can change considerably depending on variables such as changes in electrode positions, user state of alertness, and environmental noise, which for our purposes means noise occurring on different days / to different users (Apicella et al., 2024). Therefore, the models that exhibit high performance in some environment tend to fare badly or even fail during a new session or under a new user. Published research contains numerous cases that describe high classification accuracy following the traditional evaluation procedure that employs data obtained during the same sessions for both training, as observed by White and Power, (2023). However, overlap-session measures mislead performance because they amount to being learned session-specific idiosyncrasies (White & Power, 2023). Conversely, measures that are based purely on non-related sessions (or on new topics) tend to exhibit accuracy declines due to non-alignment among measures as well as distributions (He & Wu, 2019; Ma et al., 2022). For e.g., one recent study on motor-imagery BCI achieved within-session accuracy of ~68.8%. After experienced a challenge during a novel session the following day with the same subjects, this dropped to ~53.7% (Ma et al., 2022). Such research suggests the requirement to come up with session-robust models, as also validation protocols, so as to come closer to the scenario of actual-world use of the BCI.

The motivation for this dissertation was to bridge that gap between exciting bench-top results and effective, sustainable long-term performance off the bench. It must be accurate, suited for high-stakes uses (e.g., assistive BCIs or measuring driver distraction) (Becerra et al., 2025; He & Wu, 2019), be properly well-behaved long term, even as the environment changes, as well as not require lengthy or high-tech recalibrations. To prevent the fatigue of poorly predictive

models, it is important to not only develop better machine learning techniques (transfer learning, domain adaptation, and graph deep networks, etc.) but also better evaluation methods (Apicella et al., 2024). Recent research directions are exploring new direction such as alignment of feature distributions between sessions (He & Wu, 2019) and adaptive or session-invariant feature extraction techniques to curb the effects of non-stationarity. Of equal significance is the increasing recognition in the literature of evaluation methodologies. More recently, researchers are making explicit session-aware vs session-agnostic comparison of evaluation techniques to estimate the amount of over-estimation of performance (White & Power, 2023; Apicella et al., 2024). White and Power (2023) provided successful empirical support that the conventional k-fold cross-validation leads to over-inflated accuracy estimate in passive BCI setting, where blocks of correlated data are present in the training as well as the testing folds sharing the same session. These insights all serve to reinforce the motivation: to create methodologies that ensure EEG classification models are accurately assessed, and generalizable across new sessions. In this chapter we are laying the foundation of the research by outlining the research problem, aim, objectives, and delimitations of the study.

1.2 Problem statement: -

Statement of the Problem: The focal challenge in this study is the difference between high performance metrics for EEG-based classification models when evaluated using session-overlapping (or session-independent) evaluation, and where performance metrics drop drastically with session-independent (session-aware) evaluation. In other words, EEG-based classification model performance may seem incredible with random cross-validated testing, but that performance does not hold with new recording sessions or in new subjects, hence their applicability is limited. The fundamental challenge is that the EEG signals themselves exhibit non-stationarity, and thus each of the models of machine learning are at risk of overfitting for a single session (Apicella et al., 2024). A classifier could simply learn some form of noise, or artifacts associated with the relationship to the electrode location, which may have occurred during the first session (Apicella et al., 2024). This could produce artificially high accuracy when the assessment is done with that specific session's data. When engaging in a new session, even the same person doing the same task, the performance could suffer markedly (Ma et al., 2022). This generalization gap induces challenges for real-world use of BCIs that require reliable performance over time without having to retrain (from scratch) to use the BCI again.

Evaluation Practices and Limitations: A significant contributing factor to the existing issue relates to how models are evaluated in the literature. Multiple studies have utilized within-subject k-fold cross-validation which involves random data partitioning, without respect to session boundaries (White & Power, 2023). This session-overlapping procedure allows for consequences and benefits associated with a particular session to be included in the training set, to also be present in the test set, which results in overly optimistic evaluation measures (White & Power, 2023). On the other hand, utilizing a session-aware method, such as leave-one-session-out validation or block training/testing per session, reflects performance on truly unseen sessions in a more valid manner. Session-aware rankings tend to cause statistically significant decreases in accuracy, indicating to the researcher to what degree overfitting should've performed better (Ma et al., 2022). Similarly, subject overlap vs. subject independence in evaluation (also known as training by using data from all subjects and testing on a completely new subject) raises a similar decision at the inter-subject level (Apicella et al., 2024). There is no standard evaluation framework through which to conduct evaluations, and

the absence of such a framework leads to ambiguities for comparisons across algorithms, while also calling into question whether we are making true progress towards developing classifiers that generalize across subjects when evaluating independent studies or analysing pooled analyses of data sets.

Comprehensive Evaluation Approaches (1–6): In order to systematically explore and address the issues highlighted above, this study draws from six complementary evaluation approaches - including standard approaches (which pose a threat to overharvesting) to restrictive protocols (which carefully test the findings). The six evaluation approaches are:

1. Conventional k-Fold Cross-Validation (Session-Overlapping): This is also another procedure that adheres to k-fold cross-validation, where the procedure randomizes or shuffles the trials about the whole dataset, as opposed to subgrouping session-wise or participant-wise (White and Power, 2023). While this procedure optimizes diversity among the training data, it allows the overlap among sessions between the test as well as the training folds, which may optimize performance.
2. Repeated k-Fold Cross-Validation: This variation of standard cross-validation again uses differences splits based on randomness (e.g., 5x5-fold CV), meaning we'd repeat our standard CV five times and take the average performance of our model. Even though it is still considered session-overlapping, repeated cross-validation produces a more trustworthy average performance estimate, all while verifying that the model's stability is under review (White & Power, 2023).
3. Cross-Validation by Session (Block): The intended approach is a session-aware approach where you treat entire sessions as indivisible blocks when creating the split. For all folds, entire sessions (or blocks of trials) are held-out for testing and you use data from other sessions to train. This tests how well the model generalizes to entire new sessions of the same subjects.
4. Subject-Wise Leave-One-Subject-Out (LOSO): Test the generalization across subjects, i.e., the model. Training is carried out on the entire data, leaving the part regarding a single subject. The neglected subject is authenticated. This is a formal assessment regarding whether or not the model captures the individual differences (Apicella et al., 2024). In multi-session data acquisition, the LOSO also tests across unseen sessions as the sessions within a new subject are all corroborative unseen.
5. Hybrid Session-Subject Validation: This approach involves separating training and testing by session and by subject. For example, one approach is to train on data from multiple subjects, except for one session from one subject, and then hold-out that corresponding session from that subject for testing (neither the subject nor the session were seen in training). This hybrid approach tests robustness in the most extreme case of a new user/user session scenario.
6. Randomized-label control tests: To be sure that learned models are not exploiting the spuriously correlated (i.e., temporal correlations that occur during sessions) dependencies, we also conduct additional control tests where the class labels are permuted (White & Power, 2023). Permuting the class labels, either the trial or the sample, we estimate an empirical chance-level performance. If we compare the actual vs. the randomized results, we would then be able to determine whether overfitting has occurred: a significant shift towards the chance performance would indicate the original

classifier was exploiting true, significant signal structure; while still achieving high accuracy after permuted labels would indicate overfitting to data artifacts.

Using these six evaluation methods, we can directly measure session-overlapping performance relative to session-independent performance. It is hypothesized based on prior studies that methods 1–2 would perform better than methods 3, 4, and 5 with the most challenging session being approach 4 (LOSO; Apicella et al., 2024; Ma et al., 2022). Approach 6 provides a sanity check to confirm the patterns learned by the model correspond to actual information associated with the class and not noise. Collectively, this more complete evaluation strategy directly focuses on the research problem: it is needed understand how much a model's apparent success is due to actual learning and how much success is associated with evaluation artifacts; and, assess next procedures to combat the outcomes of bias and widen the generalization gap.

Problem statement: To summarize, the problem with EEG models is that they do not generalize well to new sessions or new users and standard evaluation methods tend to obscure this. This dissertation addresses this problem by (a) proposing methods for enhancing the robustness of EEG mental state classifiers across sessions, and (b) proposing rigorous evaluation protocols (as mentioned above) to honestly evaluate and compare session-aware vs session-overlapping performance. By addressing this issue, we hope to contribute toward EEG models that have high accuracy over time and across users, a goal for real-world BCI applications (Becerra et al., 2025; Apicella et al., 2024).

1.3 Aim and Objectives: -

Aim:- The primary aim of the research project is to enhance the reliability and validity of classifying mental states based on EEG data through model generalizability across sessions, as well as through improved methods of robust evaluation. In particular, this project will focus on the development and evaluation of methods to ensure and sustain high classification performance without the need for session-specific recalibration so it is one step closer to BCI systems that could be produced and deployed more feasibly.

Objective:- To achieve this objective, the following specific objectives will be pursued:

1. Literature Review: A comprehensive literature review will be conducted of the state-of-the-art methods (published from 2015 to 2025) for EEG classifications of cognitive states (focusing on mental workload and its associated paradigms) to identify methods, strengths and weaknesses presented by non-stationarity, and strategies for cross-session and cross-subject generalization (Apicella et al., 2024). Reviewing the EEG classifications will also include an inspection of evaluation measures used in other research, which will highlight gaps in and/or inconsistencies within research (White & Power, 2023).
2. Identification and Definition of the Problem in Technical Terms: From the literature, one has to articulate definitively the notion of the problem of session variability in EEG signals. The focus here is characterizing how inter-session variability and individual differences affect the distribution of EEG features and model performance. This objective is the formal definition of the problem (see Section 1.2), which partially serves as the motivating objective
3. Developing a Session-Robust Classification Approach: A useful aim would be to design a procedure for cross-session EEG classification. This might be accomplished using one approach, or some combination of the following: (i) adjusting or aligning data, or

normalizing the data, or changing the way you represent the data, to help reduce differences due to session (He & Wu, 2019), (ii) transfer learning or domain adaptation (such as fine-tuning on a small, but non-trivial amount of data from the new session), or (ii) designing a deep learning model which can learn session invariant features (using GNNs and/or adversarial training). This choice is key as it will help to directly alleviate the issues we are trying to identify (e.g., non-stationarity and overfitting to session artifacts).

4. Executing and Refining the Model: The next stage involves the application of the advocated classification method whereby it is tuned on training data so as to avoid overfitting. This involves choosing the right EEG feature representations (or applying end-to-end learning), fine-tuning architecture/hyperparameters, and potentially applying regularization methods that promote generalizability (such as dropout, weight decay, or data augmentation with simulated session noise).
5. Evaluate performance in multiple contexts: Based on the six evaluations presented in Section 1.2., systematically evaluate the developed model. Evaluating the model is necessary to investigate the state of the model whether in typical or challenging situations:
    - Session-overlapping vs. Session-independent: Compare the classification accuracy of the model when evaluated under random k-fold CV (session-overlapping) to evaluation using session-wise CV and LOSO CV (session-independent). Note how much performance drops and consider the reasons for difference (White & Power, 2023; Ma et al., 2022).
    - Comparing to the Baselines: To facilitate interpretation of improvement, it may be useful to compare your model against a baseline model (e.g. a standard EEG classifier that did not have domain data adaptation) evaluated in the same experimental conditions.
    - Statistical Significance: Use an appropriate statistical test to quantify the difference in performance between your model or proposed method vs. the baseline and whether they differences are likely statistically significant and not due to chance.
    - Randomized-Label Check: Use the randomized-label control (i.e. Approach 6) to assure the classifier's performance is above chance when class structure exists (White & Power, 2023).
6. Analyse Results to Refine Approach: Analyse the results of the multi-faceted assessment to inform the model with evidence on its performance. If the classifier remains with a large generalization gap, continue to evaluate the failure modes with the evidence obtained from feature importance analyses or EEG feature distributions across sessions (e.g. to assess if features from one session would not generalize to another). Use your findings to alter the model or preprocessing if appropriate. If session-wise biases create systematic influences in the generalization gap, it may be beneficial to implement further normalization per session. The systematic process one might follow, with evidence influencing each decision, should facilitate closing the gap in applicability along the way.

Accomplishing these objectives will yield a validated model for session-robust EEG classification, along with a series of best-practice evaluation guidelines. Overall, an improvement in classification accuracy is expected for the final model when assessed in a session-independent way compared to approaching 'traditional BCI approaches' – representing

a real 'step' toward BCIs which can eventually be generalized and worked successfully out of the laboratory.

1.4 Scope and Limitations: -

Scope:- The present thesis is concerned with a classification problem wherein cognitive states are assessed and classified through EEG, and for illustrative purposes, we use mental workload (also referred to as cognitive load) as an exemplary use-case. Mental workload was chosen for this focus on EEG-based classification, due to its relevancy in human factors and adaptive system design, as well as being appropriate for a classification problem as it is a subtle assessment internal to a participant's brain (Becerra et al., 2025; Demirezen et al., 2024). Non-invasive EEG recordings were obtained throughout the study from healthy adults while they engaged in specific tasks, which were designed to be cognitively relevant to the differing levels of workload (or mental states). The entirety of data involved in this review include multichannel scalp EEG signals created using standard electrode configurations; no invasive or intracortical signals are involved with respect to using EEG signals in the machine learning scope. The scope of the machine learning involved in this review considers supervised classification methods. This encompasses both established feature-based approaches, like feature extraction from EEG data (e.g. common spatial patterns, or power spectral densities) before classification (e.g. LDA/SVM classifiers), and newer deep learning methodologies (e.g. end-to-end CNNs, RNNs, etc) where appropriate to meet the overall objectives of the review. It is important to note that the assessment frameworks reported here (session-wise, subject-wise validation, etc.) are generally applicable to BCI paradigms, even if the experiments set out in this paper are in a mental workload context.

The domain also defines to evaluate transfer learning or domain adaptation methods. In particular, methods to leverage data from previous sessions or other subjects to support classification in a new session (He & Wu, 2019). For example, feature distributions might be aligned to support transfer learning, or pretrained neural networks might be used. When evaluating reproducibility and potential generalizability, the study is also consistent with principles for reproducible research (Demirezen et al., 2024). Each experiment had a corresponding experiment that used different random seeds, performance statistics were calculated using a held-out dataset, and wherever possible, the authors documented code and data to help future research.

Limitation:- This research denoted multiple limitations and boundaries. First off, the experiments are limited to offline experiments. The models are trained and tested on recorded datasets, meaning that there is no online deployment (online BCI operation), although the results are assuming a real-time system. Second, while there are multiple sessions per subject, the number of subjects and, thus, sessions are limited by the data set available. Therefore, while statistical analyses for results are carried out, they may not take into account every source of variability in a larger population or over lengthy time periods of recording. For example, if very long-time scale data is available (e.g., months apart), then generalization across very long-time scales is not tested explicitly. Third, the mental workload induction tasks in our data (e.g., N-back working memory tasks or multitasking contexts) reflect only a subset of the total contexts of cognitive states possible. The tasks presented have a limited range of difficulty and therefore they do not fully represent more spontaneous or complex settings found in the real world. Therefore, one needs to be careful in generalising quantitative performance across

settings that differ significantly. The goal is to provide evidence of improvement over baseline for the tasks/setting selected, rather than to talk about some absolute or "perfect" level of accuracy.

Another drawback is that the proposed techniques are aware of session variance as well as, partially, subject variance, yet do not attempt the in-depth analysis of other confounds such as artifact processing (eye blinks, muscle artifacts) or concept drift as a result of acquisition effects. Equally, broadband filtering as well as standard offline processing (artifact rejection) are utilized, yet advanced rejection beyond the limitations of this work (such as regression of continuous measures of workload) is beyond the sphere. Moreover, although the evaluation process is described as quite stringent, the focal point of success throughout is measured by classification accuracy (and associated measures such as confusion matrix and F1-score). While the user experience, computability, and calibration efforts are described qualitatively, they are not seriously optimized; for example, while we view that as a success even if a domain adaptation utilizes some small amount of calibration data from a new session, we are still not meeting a completely zero-calibration ideal, which is understandable in a sort of perfect world.

In conclusion, the scope discussions for Chapter 1 and this dissertation are limited to addressing session and subject generalization in EEG-based workload classification by improvements to algorithms and evaluations. It is limited to the controlled nature of tasks, a fixed data set, offline analysis and focusing on performance classification, in order to demonstrate improvements. This research problem defined a manageable research problem, and addressed the issue of robustness in BCI systems. The chapters that follow will commence from this discussion and cover a literature review (Chapter 2), methodology (Chapter 3), results and a discussion of findings and future work (Chapter 4) all in the frame of the scope and limitations discussed above.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction: -

Electroencephalograph (EEG) biometrics, which is the utilization of an individual's brainwave patterns for authentication or identification purposes, has many unique characteristics that may provide advantages in comparison to conventional biometrics. For example, unlike fingerprints and facial images (which can be viewed, or otherwise acquired and replicated), brain signals represent internal and concealed information that are much more difficult to acquire, or replicate (Abbas et al., 2015). The EEG, by its very nature, includes a liveness check – a dead or unconscious individual produces no EEG signals; therefore, an attacker is not able to authenticate or gain unauthorized access with a sample from an inanimate object. EEG signals are also influenced by an individual's state of mind or stress level, which makes it difficult for an attacker to engineer an authentic user to authenticate during or under duress (Ruiz-Blondet et al., 2016). These properties (internal, live signals, an individual's state of mind, and stress response) help make EEG biometrics difficult to spoof and even more difficult to coerce (Jalaly Bidgoly et al., 2020). Due to such properties, EEG is gaining interest as a high-security biometric modality (Zhang et al., 2022).

An organized literature review is important to chart the current landscape of EEG biometrics research, identify useful approaches, and surface ongoing challenges. Existing studies have explored many different approaches, and the literature review of this body of work serves to highlight what is already known and what gaps remain. For example, a literature review can elucidate known challenges such as low-signal noise ratios of EEG, variability of EEG across sessions or mental tasks, and usability and privacy problems. A systematic review of the literature can also provide insights into the strengths and weaknesses of the variety of different feature extraction methodologies and classification models proposed over the last 10 years. By distinguishing strengths and weaknesses of existing methods, we can identify ongoing open problems in EEG biometrics (e.g., improving the universality and permanence of EEG traits over time), and put into context the proposal of this project. Overall, as we consider the prior literature from feature-based techniques to the more recent deep learning models, we recognize established benchmarks that can advance work in this area of biometric identification with EEG (Jalaly Bidgoly et al., 2020).

2.1.1 Scope of Review: - In EEG biometrics, studies can be broadly categorized as either (1) classical feature-based techniques, or (2) deep learning techniques. Classical techniques rely heavily on engineered features extracted from EEGs, and on traditional machine learning. Engineered features typically include: time-domain statistics, band power across well-defined frequency bands (delta, theta, alpha, etc.), power spectral density (PSD) estimates, or coefficients from autoregressive models. Such features are typically viewed as stable neural signatures across time for individuals, and can then be classified through various approaches such as linear discriminants or support vector machines to model individual identification. Prior work has shown that certain EEG features (e.g. spectral power distributions) tend to be relatively consistent for each person (with some variability), and can distinguish individuals both across different tasks and conditions. More recently, deep learning-based approaches have become more prevalent, taking advantage of neural networks to use the EEGs to automatically learn discriminative representations from the signals. Convolutional Neural Networks (CNNs)

have been widely applied to raw or transformed EEG signals, to identify spatial–spectral patterns, achieving as high as 97–99% identification accuracy, in idealized experimental design (Sun et al., 2019). Recurrent Neural Networks (RNNs), or LSTM networks, are also introduced in some cases to not only extract the spatial representations but also the temporal dynamics of the EEG, either in isolation or as certifications with hybrid CNN–RNN architectures. For example, Sun et al. (2019) reported an analysis using a 1D CNN-LSTM model that identified ~99.6% accuracy on a 109-subject body-based EEG dataset, extracting both the spatial and temporal features of the EEG. In general, deep learning methods have shown impressive performance with the ability to distinguish individuals using more complex EEG "fingerprints". However, often, these approaches require large amounts of data and regularization to avoid overfitting and generalizing poorly to a new condition. This review will describe both classical feature-based methods and deep learning methods, while discussing how they theoretically try to minimize the variability present in the EEG signal, as well as trade-offs in their effectiveness and performance.

It is important to point out that this project mainly focuses on realistic evaluation protocol for EEG biometric assessment. Several studies in the early history of based biometrics provided very high levels of accuracy where both training and testing utilized EEG signals from the same session. However, considering that EEG signals are known to change, it is important that any algorithm tries to properly adjust for changes from session to session. This is one of the reasons for the common recommendation to perform session disjoint testing of EEG based biometrics models, as the aforementioned evaluation better captures a real-world experience (Gong et al., 2024). Therefore, additional protocols in the EEG literature have also tested on different tasks or on EEG between subjects to have a better understanding of the generalizability of the applications. Model evaluations have been performed using the PhysioNet EEG Motor Movement/Imagery dataset, which includes multiple recordings sessions for the same 109 subjects, consistently investigating the task via a leave-one-session-out (LOSO) protocol, thus separating training and testing by not overlapping any sessions. LOSO is not the only protocol being tested either, as there are also examples of leave-one-subject-out (LOSO) using the same dataset. Looking through the literature it has generally shown that across sessions the performance for EEG biometric protocols typically goes down and this can be attributed to any number of reasons, including repositioning of electrodes, change in mental state, or noise (Jalaly Bidgoly et al., 2022). As a result, the majority of more recent work- again more to this thesis- involved mainly improving the session invariance and robustness of the EEG identity verification methods. The following part of the chapter has a literature review, discussion of the classical methods of EEG biometrics and features, and modern methods based on deep learning methods and studies that provided evaluative results of past experience compared to EEG biometrics all over sessions and other considerations with our subjects, etc. Each sub-section is constructed to provide a range of scholarly literature while add to a foundation of academic literature in which bolstered the methods in the following chapters.

2.2 Feature-Based Approaches: -

Typical EEG based biometric systems use hand-crafted feature extraction followed by classical machine learning classifiers. In these systems, researchers extract descriptive features from the EEG signal, such as power spectral densities (PSD) in particular frequency bands, energy or band-power within conventional EEG rhythms (delta, theta, alpha, beta, gamma), time-domain statistics, wavelet coefficients, and measures of connectivity, afterward classifiers (e.g., support

vector machine (SVM), k-NN, Random Forest) are applied to classifying individuals. Many studies based on this feature-based methodology suggest and/or show (2015-2025) that high identification accuracy can be achieved under various experimental conditions.

2.2.1 Feature Extraction Techniques:- Power spectral density (PSD) and band power features are perhaps most commonly applied in EEG biometrics. By transforming EEG from time to frequency domain, PSD-based features pick up how signal power is distributed across traditional frequency bands (delta through gamma), which typically harbour individual-specific patterns. For instance, a more recent study authored by Monsy and Vinod (2020) proposed a frequency-weighted power (FWP) feature (modified band power measure) and achieved virtually perfect recognition accuracy (99.95%) on a 64-channel eyes-closed rest EEG data set from PhysioNet (109 subjects). Such findings reflect the high discriminative power of spectral features in laboratory-controlled scenarios. Another classical method involving autoregressive model of EEG efficiently compresses the signal shape in the time domain. In fact, AR coefficients were utilized in some of the first works on EEG biometrics and have remained a strong performer ever since. Maiorana et al. (2015) determined that AR features extracted greater subject-specific distinctiveness compared with basic spectral power or coherence features in resting-state EEG. In their experiments, AR-based representations enhanced recognition accuracy, implying linear dynamical EEG traits reflect subject-specific information superiorly to raw band energies. Features derived from the wavelet transform have also been tested to take advantage of time-frequency information. For example, Waili et al. (2019) utilized discrete wavelet transformations to derive EEG features and implemented an MLP-based classifier. Whereas wavelet-based algorithms are capable of registering time-frequency information and nonstationary EEG characteristics, their success has been uneven; Waili et al. (2019) resulted in approximately 73% recognition accuracy on a very modest 6-subject data set (19-channel EEG, single data acquisition session), promising potential for development. Some experiments have integrated multi-type features to supplement the identity information. Kang, Jo, and Kim (2018) suggested a multi-modal set of features consisting of band power, signal complexity, and non-linear dynamics, which improved authentication performance substantially. Their feature comparison on an open dataset of 109 subjects (PhysioNet Motor Movement/Imagery) remained high, with 98.9% average subject identification accuracy when combining spectral power (over 19 task conditions) with features based on the Hilbert transform. Such performances indicate that well-crafted but unmatched features of an individual's "brain print" are possible. Time stability of EEG features has also been explored; Yang, Deravi, and Hoque (2018) compared some task-dependent sets of features and found the selection of the mental task (e.g., resting or cognitive activity) to have insignificant impact on subject identification performance – including when being separately trained and tested on different tasks – provided the type of the features remained the same. This invariance with respect to tasks implies subject-specific EEG signatures existing across mental tasks, which is promising for deploying simple resting-state features in biometrics.

2.2.2 Classification Methods:- With informative features in hand, classical machine learning algorithms were able to achieve strong EEG-based recognition without recourse to deep learning. As an alternative to deep learning, there has been intense interest in support vector machines (SVMs) due to their high success rate in high-dimensional data like EEGs and their ability to cope with nonlinear decision boundaries.

Many papers during the last decade claim to utilize SVM classifiers on EEG features with great success. For instance, Di et al. (2019) and Li et al. (2019) utilized SVMs on PSD-based features with high single-session experiment results. Nakamura et al. (2018) also utilized an SVM in their in-ear EEG biometric, combining PSD and AR features; they achieved approximately 94.5% accuracy on 15 subjects tested across two individual sessions.

Simpler distance-based classifiers have also been successful. Kang et al. (2018) took a simple Euclidean distance measure and utilized it to match functional network features of resting EEG, but were still able to differentiate 109 subjects with almost 99% accuracy. k-NN classifiers have also been utilized – for example, a paper by Kaewwit et al. (2017) utilized k-NN on AR coefficients (after removal of artifacts with ICA) and achieved 98–100% subject recognition accuracy on an eyes-open resting-state 20-subject EEG dataset with just four electrodes. This achieved with 5-fold cross-validation on single-session data, demonstrates that very simple classifiers can do an admirable job given high-quality features revealing the person-specific EEG structure. Apart from these, researchers have also utilized ensemble learning on EEG features. Decision tree ensembles like Random Forests and boosting techniques have occasionally been employed to improve generalization. Curran et al. (2017), for example, used an extreme gradient boosting (XGBoost) classifier in a single-earpiece EEG authentication prototype. They reported around 96% accuracy on a small sample of subjects by leveraging boosting to fuse multiple decision trees, illustrating that ensemble methods can handle the variability in EEG features well. Overall, the literature suggests that no single classifier is universally superior for EEG biometrics; rather, many classical models (from linear discriminants to nonlinear SVMs and tree ensembles) can achieve comparably high performance when appropriately tuned. Simpler models often suffice in this domain, presumably because the dimensionality of feature vectors (after feature extraction) is modest and the inter-subject differences can be made quite pronounced by the chosen feature transformations. This is a notable difference from fields like image recognition, where deep networks vastly outperform shallow models – in EEG biometrics, carefully engineered features can level the playing field for classical algorithms.

2.2.3 Performance and Evaluation:- Studies utilizing classical EEG-based personal identification systems with feature-based approaches have shown exceptionally high performance ranges under certain conditions, but the performance is conditional on test conditions. The first studies on EEG biometric identification assessed identification in constructed datasets, either identified individuals from a single testing session, or used randomly segmented segments from a same-session EEG data set. Using a dataset containing 20 individuals, Kaewwit et al. (2017) reported 100% accuracy using k-NN and AR features with a single recorded session. Using a same-session-data-based dataset (PhysioNet EEG motor imagery dataset), extremely high accuracies (>98–99%) have been reported using random train-test splits. Monsy and Vinod (2020) reported virtually error-free classification on resting-state segments from the same dataset using 'FWP' spectral features and an SVM. Kim and Kim (2019) reported >90% accuracy on the same dataset using a "quadrantal functional network" feature constructed using a functional connectivity approach, with three separate recording sessions for each individual. While these results demonstrate that EEG patterns of individuals can be learned and classified with high degrees of reliability using classical approaches, these studies utilized static conditions to do so. When more demanding evaluation is imposed, such as requiring session-disjoint training and testing or testing on data collected

on different days, performance will often decline. Cross-session variability remains a major hurdle owing to electrode contact, impedance, state of mind, and other non-identity factors. Some studies have asked this question more directly. Maiorana and Campisi (2018) conducted an in-depth longitudinal evaluation of EEG biometrics over three years, with 45 subjects each recorded in 5-6 sessions. Applying an hidden Markov model (HMM) classifier to a mixture of time, frequency, and time-frequency features, they also observed significant drops in recognition accuracy with increasing time separation from enrollment to test. Their exploration of EEG "permanence" showed that matching scores lose quality with biometric template aging so that error rates worsen past a year or longer. For short time separations (days to weeks), corresponding loss of performance is smaller but non-trivial. For example, experiments with single-day training and single-day testing have been found to have equal error rates in the few percent to ~10% range, while same-session EERs can practically reach 0%. Nakamura and colleagues (2018) showed that their in-ear EEG implementation achieved ~94% recognition rate across the two sessions which were separated by approximately a week - a slight decline compared with the nearly 99% recognition rate within a single session using identical data. These investigations emphasize the value of evaluating models with sessions separated from the lab: for models to be more broadly applicable outside the lab, the models must accommodate realistic variability. Consequently, many of the more recent investigations fall back on leave-one-session-out (LOSO) evaluations or cross-session train/test sets in order to more rigorously examine generalizability. For instance, Thomas and Vinod (2018) just evaluated their gamma-band power features on a session-disjoint split yet still demonstrated high authentication efficacy, which does support that there are certain frequency bands (specifically gamma bands) that carry stable biometric information across sessions and subjects. Overall, trends in the papers are leaning towards experiments involving multi-session publicly available data sets while evaluating under more realistic and applicable ways (e.g., training with an earlier session and testing with a later session, or even leave-one-subject/session out). While these methods do shift the overall accuracy down from many of the earlier published works, all of these classical methods maintain high identification rates over thresholds of 90% even when accounting for these more realistic conditions.

To further improve stability, researchers have also looked at techniques such as channel selection and subject normalization. For example, Moctezuma and Molinas (2020) developed a genetic algorithm to select a subset of electrodes that minimized subject classification in resting EEG, to reduce noise from electrodes that contribute less information. Kang et al. (2018) applied a threshold for subject-specific classifier, thereby enhancing verification consistency through individual differences in score distribution. Refinements such as these, coupled with evaluation on an individual-session basis, are causing classical EEG biometrics to take steps towards practical implementation. Features such as band power/PSD, wavelets, AR modeling, and connectivity measures each reflect different aspects of the EEG, and the feature choice can be dependent upon the task. There is agreement across several experiments that there are identifiable signatures in resting-state EEG for each individual and even with simple classifiers such as SVM or k-NN, >90% identification rates are attainable on dozens to hundreds of subjects. Performance in non-controlled settings (across days or sessions) is usually disastrous, highlighting a central challenge faced by classical approaches. In the next section, we will present deep learning-based approaches which will seek to overcome such challenges by learning invariant representations. However, it is important to appreciate the strong foundations established by classical approaches. Classical approaches offer both

interpretability, and computational efficiency (potentially implying fewer data required to maintain the same generalizability). Both of which become important considerations given the typically modest sizes of EEG biometric data collections. Hence, classical EEG biometric systems, and hand-crafted features, are therefore extremely relevant, and have genuinely achieved state-of-the-art performances in significant studies - when adequate emphasis is placed on the issue of feature engineering and evaluation protocol.

2.3 Deep Learning Approaches: -

The use of deep learning architectures for contemporary EEG-based biometric verification has made it possible to automate the extraction of discriminative features from raw time series or their time-frequency parametric transform of them (or their time-frequency transforms). Prior to usage of deep learning, the primary approaches to biometric verification from EEG relied on hand-crafted EEG measures (e.g., power spectra, connectivity). Deep learning models can directly learn the EEG organization of a subject and account for greater variability in their tracing than the hand-crafted features used for verification across a large population (Mao et al., 2017; Chen et al., 2020). During the last 10 years (2015-2025), many deep learning architectures had been created and used in the classification or identification of identities from EEGs. For example, convolutional neural networks (CNN), recurrent neural networks (LSTM or RNN), and hybrids. This section describes the deep learning models used in EEG based biometric verification, and emphasizes experimental findings that utilize publicly shared EEG datasets and rigorous validation schemes (e.g. cross-session tests, leave one subject out validation).

2.3.1 CNN-Based Identification on Raw and Time-Frequency EEG:- Several experiments have suggested CNN-based architectures with raw EEG signals (or transformed representations) as input and producing identity classes as output without the need for manual feature engineering. Convolutional Neural Networks (CNNs) are capable of capitalizing on the spatio-temporal information of multi-channel EEG. For example, Mao et al. (2017) developed a light-footprint CNN with 2 layers of convolution and 2 layers of fully-connected networks to extract features from 1-second EEG epochs (considered 2D matrices of time × channels). Based on a dataset consisting of 100 subjects' driving fatigue EEG, their CNN achieved approximately 97% recognition accuracy, showcasing the potential of end-to-end deep learning to recognize people from brain signals (Mao et al., 2017). Of interest, when validated with resting-state data (without task stimulus), recognition fell to ~90%, implying that level of cognitive engagement has the potential to impact the discriminability of EEG features from an individual's identity level. This highlights the importance of considering EEG recording conditions in biometric systems.

Further initial explorations using CNNs have supported deep learning using smaller EEG studies and experiments. For example, Arnau-Gonzalez et al., (2017) utilized CNNs on visual evoked potential EEG data from 23 participants and reported ~94% accuracy for subject identification using a fixed stimulus protocol. In a somewhat similar case, Wu et al., (2018) used CNNs on RSVP EEG data (15 subjects) achieving greater than 97% classification accuracy. While limited in subject numbers, these studies demonstrated that CNNs could learn stationary individual differences from EEG data related to discrete events. Publicly available EEG data collections have allowed for scaling the use of CNNs in EEG-based biometrics. For example, a particularly rich source of publicly available EEG data includes the PhysioNet EEG

Motor Movement/Imagery dataset, which features data from 109 subjects, who had multi-channel EEG data collected during resting (open/closed eye) and motor imagery paradigms. In a large dataset of 109 subjects, Fan et al., (2021) developed a model of EEG subject ID from rest-state EEG data using a modified existing EEG Motor Movement/Imagery dataset and a CNN classifier which also included data augmentation. The model used epochs of only 0.5 seconds in duration, utilized 14 different EEG channels and achieved average recognition of 99.3% with an Equal Error Rate of only 0.18% using cross-validation. Achieving such significant performance with short-length signals, and using an inexpensive EEG device, speaks to the ability of CNNs to learn and extract salient features across spontaneous EEG (Fan et al., 2021). It further suggests that there are unique individualized signatures in "resting state" brain activity when appropriately filtered through a trained deep network. Chen et al. (2020) introduced their own dedicated CNN called GSLT-CNN (Global Spatial and Local Temporal CNN) with the capability to operate directly on raw EEG channels. GSLT-CNN applied spatial convolution filters over channels and time-domain filters over time, which afforded the ability to derive patterns specific to a person without any human interference in the extraction of features. Tested on a large 157-subjects EEG dataset across four experiments, GSLT-CNN resulted in an approximately 96–98% identification rate and was found to stay strong across-session tests (Chen et al., 2020). Robustness across more than a single recording session/task indicates CNNs should be capable of generalizing across more than a single session given proper regularization or structure. In fact, Chen et al. indicate their CNN performed with high accuracy when their CNN was trained and tested on EEG from different sessions or different task conditions. This cross-session robustness is crucial for practical biometrics, since an individual's EEG on enrollment day might differ from a later authentication session.

In order to improve CNN performance, researchers have changed the input representation and input characteristics of neural architectures. Some papers proposed time-frequency representations of EEG (e.g., spectrograms or wavelet transforms) simply as CNN input, given the networks' proficiency in classifying images. For instance, in an Akbarnia and Daliri (2024) study, they extracted frequency domain features (Fourier spectra and wavelets) from EEG during the "focused" vs. "mentally distracted" condition, then used those features to differentiate participants with a deep neural network. Their work was validated using 109 participants, who performed a mental task (e.g., mind-wandering, waiting), yielding about ~99.2% identification accuracy rated when participants were in the focused condition and ~97.8% when participants were waiting or distracted. The marginal decrease associated with the distraction suggests the state of mind impacts EEG biometrics, but the deep model was still very impressive in using the additional noise from their blinks and mind-wandering (Akbarnia & Daliri, 2024). Additionally, other studies adapted abbreviated CNN architectures designed for BCI technology for the purpose of biometric authentication (e.g. EEGNet). For example, Jijomon and Vinod (2023) showed that EEGNet produces a significantly more precise model (~ 86 - 99% classification accuracy depending upon architecture optimizations) for identification than deeper CNN architectures (e.g. ResNet), a finding established on a 21-participant EEG benchmark; suggesting that regardless of classification precision, simpler (and less computationally intense) CNN architecture can be adequately used with minimal amounts of EEG data (Jijomon & Vinod, 2023). Pure CNN-based methodologies achieved astounding performance (often >95% accuracy, across tens to hundreds of subjects) by fully automating the learning of spatial-spectral-temporal features that were unique to each individuals' brain activity.

2.3.2 Recurrent and Hybrid Deep Learning Models:- While CNNs are adept at learning of spatial features, but poor in extracting long-term time dynamics from EEGs, there has been intense interest in combining RNNs, i.e., Long Short-Term Memory (LSTM) or Gated Recurrent Units, with CNNs as the feature extractors, so as to derive hybrid frameworks learning time relations as well as spatial patterns. Wilaiprasitporn et al. (2019) initially proposed an affective EEG-based subject-identification framework with the implementation of a CNN-LSTM network. In their study, 32 subjects viewed emotional video clips (DEAP dataset), and their model consisted of convolutional layers (to spatially filter EEG channels) and LSTM layers (to learn temporal sequences) (Wilaiprasitporn et al., 2019). Importantly, their CNN-LSTM had acknowledgment rates higher than 99% for subject recognition and even better performance when EEG was recorded with differential emotional states (i.e., different induced moods). By observing a subset of informative channels (i.e., frontal electrodes) and tracking the evolution of the EEG signals using recurrent units, their hybrid architecture produced similarly accurate outcomes with differential changes in arousal and valence. This implies deep models do have tolerances for some degree of variability in physiological states; however, it is interesting to note that subject-identification accuracy declined if mixed mental state EEG was used - compared to EEG for a stable mental state - as input. They also attempted to apply CNN-GRU (with GRUs instead of LSTMs) and reported similarly high performance (≈99–100%), additionally commenting that different RNN cells seem to have different benefits from CNN learning of features in EEG biometrics (Wilaiprasitporn et al., 2019).

Other hybrid frameworks incorporate attention mechanisms for additional performance gains. For example, Zhang et al. (2017) proposed an attention-based LSTM system to weight the importance of the different time steps of the EEG time series to help with identity recognition. More recently, Balci (2023) released DM-EEGID, a two-stream architecture that contains an LSTM (with an attention layer) and a multilayer perceptron (MLP) classifier. In the first stage of their system, they implemented channel selection to reduce non-informative electrodes to optimize the best separating frequency band for the subject (in their experiment, they chose the delta band). The EEG data was then filtered to include the 48 best of the 64 total channels as input to the attention-based LSTM, which captured temporal structure, followed by classification through the MLP. When tested on eyes-closed and eyes-open resting EEG from the PhysioNet 109-subject dataset, eyes-closed data were correctly classified with an accuracy of 99.96%, and the eyes-open data were correct 99.70% of the time, with the eyes-closed data achieving virtually the same results when the eyes-open data was flipped. Such high accuracy demonstrates that hybrid frameworks leveraging joint EEG dynamics can achieve success; the attention mechanism suggests that the LSTM was able to enhance subject-specific portions of the EEG signals (for example, during eyes-closed rest, idiosyncratic alpha oscillation patterns). Wang and colleagues (2024), which is another of the relatively more recently proposed papers, presented a convolutional LSTM network with attention (ABCL-EHBI) to improve resiliency to the artificial or manipulated EEG stimuli. Their architecture (ABCL-EHBI) with attention was shown to improve discriminability of true person-specific EEG from generative forgeries, although early in that space as well. It is clear that hybrids of CNNs and RNNs have shown they will work when at least one EEG signal is non-stationary or time-varying subject states are tested. Hybrid frameworks exploit the extraction of spatial-spectral feature data by CNNs and the access to sequence structures from RNNs to get state-of-the-art recognition rates across several paradigms (resting, evoked response, emotional EEG). It is remarkable that deep network hybrid architectures have obtained nearly 100% accurate rates with only a few subjects

in a closed-set recognition scenario, always caveated with a closed, or known, world test (test subjects were presented to networks during learning). Performance in respect to open-set or longer time separation continues to be an open question.

2.3.3 Session Generalization and Transfer Learning:- One major problem of EEG biometrics is cross-session variability: EEG features tend to vary across sessions because of electrode movement, fatiguing, or from-day-to-day variability in the brain. There has been an initial shift in deep learning research to address it through learning of models with the capability of generalizing across sessions or through transfer learning methodologies. Ozdenizci et al. (2019) proposed an adversarial learning of session-invariant EEG features. They incorporated an adversarial loss in a deep CNN in such a way that the network's latent representation couldn't distinguish across sessions, making it focus on person-specific features rather than session-specific noise. When they applied their adversarial CNN in experiments with longitudinal EEG data from subjects over some days, their adversarial CNN outperformed a baseline CNN in cross-session person ID precision, substantiating that learning invariant representation is capable of nullifying the effect of sessions (Ozdenizci et al., 2019). Through the use of just 0.5-second EEG epochs, their method held well over time, an interesting step in the direction of practical usage in which enrollment and authentication can be distant in time.

A different method is transfer learning, in which knowledge from one dataset or task can be transferred to another dataset or task, and can address limited data issues faced. Al-Janabi et al. (2024) proposed a deep transfer learning model for EEG biometrics that was a combination of a CNN pre-trained on a large, unrelated dataset (subsequent to being developed) with a classical classifier for decoding EEG features. They were able to achieve greater accuracy when using the pre-trained network (for example, on ImageNet or a large EEG dataset) and fine-tuned it using a lesser-sized (target) dataset for EEG biometrics, versus creating a CNN endpoint from scratch (Al-Janabi et al., 2024). The combination of feature extraction using a pre-trained CNN and downstream classifiers indicates that generic feature extractors can be repurposed for EEG signals or features learned from other EEG related tasks (e.g. motor imagery) can be utilized for identification. Some studies have also shown the ability trained deep neural networks to be fine-tuned on different EEG datasets with a significant reduction in dataset size, whilst maintaining identification rates (Zhang et al., 2022). Transfer learning has particular significance, given the lack of very large EEG biometric datasets – even most public datasets have data for fewer than 100 subjects. Researchers have gathered and repurposed models across data contexts, such as adding data from similar tasks, or transferring a synthetic EEG, while maintaining consistent identification accuracies (over 95%) even with a small amount of training data per subject.

In conclusion, in the last ten years, deep learning approaches have revolutionized biometric verification with EEG. For instance, convolutional networks applied to either raw EEG or spectrograms can automatically learn meaningful discriminative "brain fingerprints," with accuracy rates of roughly 90 to 100%, across dozens to over a hundred subjects in laboratory settings (Chen et al., 2020; Fan et al., 2021). Hybrid CNN-RNN structures also capture time evolution, achieving high-90s performance (Wilaiprasitporn et al., 2019; Balcı, 2023). Also, in the past several years, there has been progress with adversarial training and transfer learning to help overcome some of the generalizability issues to make the model more robust across sessions, to transfer to new datasets (Ozdenizci et al., 2019). Arguably testing has gotten better: several papers have utilized either cross-session or leave-one-subject-out testing to reduce

overfitting to a session. Ozdenizci et al (2019) and Chen et al. (2020) tested their CNNs vigorously with data from independent sessions. Furthermore, Fan et al (2021) showed stability with cross validation with a public dataset on hundreds of subjects. Prominent performance, or benchmark, indices are generally rank-1 recognition accuracy (>95% typically), or verification metrics, or Equal Error Rate (good to go as low as 0.2% by Fan et al (2021)). Again, these papers show deep learning methodologies are able to take advantage of unique neural footprints in EEG for biometric applications, but the influence of noise, and emotion, or longer-term variability is yet to be established. Overall, there is a distinct progression in the literature to deep, end-to-end architecture without the use of manual feature engineering characteristic of long-standing practice. As there are more open EEG collections and better architectures (e.g., with attention and domain adaptation) are developed, EEG-based individual recognition should become more accurate, independent of sessions, and viable biometric security implementation candidates.

2.4 Cross-Session and Task Robustness: -

EEG-based biometric systems also suffer markedly in maintaining their level of performance over different days of recording, or across tasks. Recent papers have highlighted the importance of session-disjoint evaluation - where EEG data used for training and testing are from different sessions - in determining generalizability (Chan et al., 2018). Standard protocols often employ cross-day testing, in which training occurs on one day and testing occurs at a later date, and leave-one-session-out (LOSO) cross-validation, in which testing uses an entire session while training uses the remaining sessions. Similarly, cross-task tests train on a single task or stimulus condition's EEG and test on another (Yang et al., 2018). These tests give a more realistic measure of robustness outside of same-session conditions.

2.4.1 Performance Degradation Across Sessions:- It is well-established that high-accuracy models that are built from a single session recording significantly lose their performance on a new session. For e.g., an EEG Identity (ID) classifier based on a deep CNN that achieved ~98-99% accuracy for the same recording session performance degrades to ~60-70% accuracy when identified from a different session (i.e., a different day). This deterioration is due to the non-stationary nature of the EEG signals – over time due to electrode repositioning, alertness changes, etc, the behavior of an individual subject is consistently changing (Plucińska et al., 2023). Utilizing an adversarial domain-adaptation framework increased cross-session EEG ID accuracy from ~66% to ~72% (10 subjects, 10-class ID) after the network learned invariant features within and across recording sessions (Ozdenizci et al., 2019). The adversarial component of the network was able to learn to eliminate session-specific information while learning more stable across-day person-specific predictions. These results further bolsters the claim that same-session tests overestimate accuracy, and that a robust EEG biometric should be assessed using EEG data collected over independent sessions, or days.

2.4.2 Public Dataset Findings:- Research analyzing public EEG datasets supports the same points. The PhysioNet EEG Motor Movement/Imagery Dataset, which contains data from 109 subjects performing 4 distinct tasks, is commonly used as a reference. Yang et al. (2018) noted that when using EEG from a motor imagery task to train a biometric model, it even maintained high identification accuracy on a different motor task with the same subject. In the experiments, models trained on one active task had similarly high identification accuracy when tested on another active task and, in some cross-task pairs, improved accuracy; however, when EEG

from an active task was used to train and then tested against EEG during a resting state there was little identification accuracy. They also reported that median identification accuracies were around 96%, with hand movement imagery (the most discriminative task) and multiple scalp regions producing better accuracies than any task region combination. Furthermore, pooling data across multiple tasks for training produced high identification accuracy and steadily increased it, based on the number of task data included until it saturated at over a 99% identification rate. This may be indicative of the notion that the overall cognitive characteristics of the diversity of the tasks of the individual potentially provides additional features to be in the subsequent overall decision framework that worked in the above example. Another public dataset, DEAP, created an even more challenging biometric scenario, as experiencing and inducing emotional states can vary significantly. Research has indicated that behavioral changes in relation to a user's emotional state would change the dynamics in their EEG, which would affect the recognition process stability (Arnau-González et al., 2018). In fact, even cross-session comparison across emotional EEG data has been argued to be less stable, demonstrating reason for considering variability in user state in biometric models.

2.4.3 Longitudinal Robustness:- Longitudinal research has verified that EEG biometrics are viable and yet can be affected by "aging." Maiorana and Campisi (2018) examined EEG biometrics with 45 subjects over a 3-year proposal, collecting EEG data over 5-6 sessions, and reported a gradual increase in verification error rates, which was primarily attributed to a decay in feature permanence, as the time gap between enrollment and verification increased. They examined the continuity and stability for each unique EEG channel, and found that some electrodes had stable biometric information over time, consistently maintaining information associated with global identity. By way of multi-channel selection, and the use of additional enrollment sessions, the "aging" effect on performance decreased. For example, enrolling a user using a single EEG session produced an equal error rate (EER) of ~16.9%; however, when users were enrolled with data collected over multiple enrollment sessions, accuracy improved significantly (EER dropped to ~12.0%). In future work, Maiorana (2021) applied a Siamese convolutional neural network (CNN) to learn task-independent features on a multi-session, multi-task EEG dataset. This deep model produced some of the best cross-task verification cross-task EER of ~5.0% to ~10.0%, even when the testing task was different from training task, especially when sessions of enrollment data were available. These findings underscore the importance of including cross-day data: Plucińska et al. (2023) find the same that more training sessions leads to improved accuracy, where they use up to 8–15 unique sessions per subject to train their models, producing ~95–97% verification accuracy using unseen sessions, whereas those trained using a single session would have falsely high accuracy on within-session testing and drop performance on new sessions. The authors observe that EEG signals are particularly sensitive to session effects, which stems from either electrode placement or transient physiological or environmental influences, meaning that tests limited to a single session may create an overly optimistic impression of biometric performance. There is agreement in the literature that more public datasets are needed with multiple sessions per subject (currently a gap, as many studies are limited to datasets with single-session recordings, like the original PhysioNet and DEAP).

2.4.4 Techniques to Improve Generalization:- Various methods for addressing variability across tasks and sessions have been introduced. One method is called data augmentation. Mota et al. (2021) augmented the PhysioNet EEG dataset and used it to train a deep network for

cross-task person verification. The Squeeze-and-Excitation CNN they developed was able to achieve state-of-the-art equal error rate (EER = 0.1%), when verifying subjects across two different motor imagery tasks, demonstrating that training data enrichment can greatly enhance cross-task robustness. A second approach is called domain adaptation, which aims to learn invariant features. Ozdenizci et al. (2019), for example, implemented adversarial training to train the network to ignore session specific information and forced the feature distributions to align across sessions to yield more stable outputs. These approaches are closely aligned to transfer learning, as the knowledge learned in one session or condition can be transferred and adapted to another. Finally, multi-task or transfer learning approaches have been investigated in order to take advantage of shared structural features across different tasks or types of stimuli in training to make the identity features learned less specific to the task (Sun, 2008; Vinothkumar et al., 2018). Ultimately, channel-selection and robust feature extraction techniques can also mitigate the impact of hardware variability: each session (and hardware) shift produces its own spatial variability, so if the goal is to minimize electrode placement changes and find out which produce the same patterns and are less sensitive to changes in time, identifying which electrodes produced the most stable, person-discriminative signals over time can help to mitigate the impact of those electrode placement changes. For example, data from Maiorana and Campisi (2018) ordered channels by their distinctiveness over long-term and suggested that a simplified montage with only those reliable channels would produce accuracies similar to others while minimizing variability across sessions. In summary, the reviewed literature from 2015-2025 suggests that using these principles, including rigorous evaluations of session disjoint evaluations along the lines of tasks and sessions, eliciting richer training data across sessions and tasks, and using algorithms explicitly for invariance will improve the efficacy of EEG biometric systems in cross-task and cross session uses.

2.5 Summary of Key Trends and Findings: -

Various techniques for feature-engineering and classification were examined in studies that dealt with various EEG-based biometrics, and these studies focused on the machine learning approaches. Standard techniques generally consist of the extraction of temporal, spectral or connectivity-based features (e.g., band-power, autoregressive coefficients), and then classify using an SVM, or shallow neural nets. More recently, the trend has been to use deep neural networks (for example, CNNs and LSTMs) to learn representations of the EEG (e.g., Becerra et al, 2025; Montoya et al, 2025). Many papers show very high identification accuracy (often >90-95%) in laboratory settings, with limited numbers of subjects and under ideal settings; for e.g., one recent survey notes that EEG systems “have demonstrated high accuracy” but with limited subject numbers and ideal conditions. For instance, Fallahi et al. (2025) indicated that deep methods obtained a significantly superior outcome when compared to classic features over a very large dataset (lower EER by 16.4%). However, almost all of the studies referenced along with this study would disclose that a small number of participants - typically the first day of a single session - were studied within each of the other studies cited with them, and if the other studies cited a larger number of participants, there was still a lack of evidence to acquire measures within a single day. "As Rahman et al. (2022), noted meek in the past, many studies trained and resulted on EER results in mixed experiment conditions or trained and tested models on the same/session weeks daily every other week days”. What is also important to note is that an EEG has strong non-stationarities, e.g. electrode locations, mental state, noise elements (Ozdenizci et al., 2019; Rahman et al., 2022). EEG's train models often capture

session independent identification cues, especially identifying session artefacts as unique identifying features.

Important findings from the papers are: standard EEG features (PSD, AR coefficients, connectivity metrics) are able to differentiate people to a degree, and deep networks are able to automatically extract complex spatio-temporal patterns; but these conclusions hardly ever take variability across sessions or tasks into consideration. Increasingly, authors warn against standard cross-validation (within-session) leading to overestimation of accuracy, and recommend multi-day/session validation. Rahman et al. (2022), for e.g., sorted train and test data separately according to the day of recording, but still obtained (~98%) ID accuracy with sophisticated an ensemble model; it suggests that intentional protocol engineering may be able to increase robustness. Similarly, Montoya et al. (2025) review remark is that "EEG biometric research is an active area" requiring incorporation of external/internal factors (emotions, fatigue, diseases) in order to prevent performance declines in real-life. Summary of corresponding articles: EEG has produced new unique identity signals, but published classification accuracy is typically constrained to a narrow and specified definition of accuracy that may not necessarily generalize.

2.5.1 Limitation, Challenges, and Research Gap:- Significant challenges still remain unaddressed. The amount of variability across sessions has also been cited as an important barrier. EEG signals are known to vary quite a bit from one day to another. Thus, the model must learn person-specific characteristics that are invariant. Rahman et al. (2022) report "the large signal variability of EEG when recorded on different days or sessions impedes performance". In a similar vein, Plucińska et al. (2023) demonstrate the possibility of high error rates after only a single training session, and verification should not take place until multiple session data is covered for safe verification. Indeed, Plucińska et al. reveal most earlier papers are affected so: "the EEG signal is highly vulnerable to interferences, electrode placing, and temporary conditions, which can cause overestimated evaluation of considered methods". Additional limitations are the potential effects of overfitting and overestimated accuracy due to data leakage (i.e., overlap of tasks or sessions within train/test splits), and limited tests for realistic factors such as noise, movement artifacts, or heterogeneous tasks. Becerra et al. (2025) highlight that almost all high-accuracy findings were conducted using small, controlled corpora of a modest number of subjects, and warn that “if other dimensions are not taken into account, [systems] could eventually show significant limitations from a performance point of view in real life”. In practice, very few systems have been tested under completely unconstrained regimes (e.g., across different EEG devices or in the presence of background noise).

The gap in research is clear, although there are deep-learning-based methods and feature-based methods, there are no or few systematic and fair comparisons that draw on the literature. In particular, there are very few studies that have compared classical vs. deep strategies based on strong cross-session evidence on a large public EEG biometric dataset. Chen et al. (2019) highlight the tendency for many deep learning papers to still apply within-session testing and that an explicit session-invariance "remained an open question." Along the same lines, Plucińska et al. (2023), note mentioning that increasing the number of recording session provides only marginal improvement in accuracy beyond a certain threshold of new recording sessions; so instead of rewarding the accuracy of the previous methods could attempted to find some unknown "need" to compensate for variability by using new methods or protocols. There also lack large-scale benchmarks: large repositories of EEG data like PhysioNet's multi-session

Motor Imagery data have, as an example, never been systematically utilized for biometric testing. Thus, we currently lack knowledge about the extent to which methods really generalize across days/task or the differences between deep vs. traditional-pipeline approaches in these situations.

Filling these gaps, the current dissertation compares deep and feature-engineering models on EEG biometrics with strict, session-disjoint testing. The experiments specifically use leave-one-session-out (LOSO) and leave-one-subject-out regimes on a large public data set, bypassing the frequent fallacy of intra-session testing. Classical pipelines (e.g., PSD, AR features with SVM) as well as state-of-the-art deep nets (CNN/LSTM) are attempted and optimized. By logging each step of data preprocessing and split plan, the work is transparent and reproducible. In so doing, it hopes to present a fair baseline of realistic robustness and accuracy, directly tackling the found limitations in the existing literature. Ultimately, the ultimate aim is to gauge the extent to which EEG biometrics accuracy diminishes under realistic conditions and to develop stronger identify models capable of being operational across sessions and operational conditions.

CHAPTER 3: METHODOLOGY

3.1 Introduction: -

This section illustrates the methodology of the EEG-based biometric identification study. Firstly, a description of the two EEG datasets used in the study follows, including the number of subjects, as well as the recording conditions in which they were recorded. Secondly, the particulars of the signal preprocessing step (including channel selection, filtering and window segmentation) are presented. Thirdly, we discuss the extraction of features for classical classifiers and then the format of the data for deep networks. Following that, we provide model architectures and hyperparameter tuning of the classical (SVM, Random Forest) and deep (CNN, CNN-LSTM) models. Finally, we present the evaluation protocols (cross-validation schemes and evaluation metrics), tools/libraries and ethical considerations.

3.2 Datasets: -

Two EEG datasets were utilized in this study. The first dataset is a publicly available Kaggle "Complete EEG" dataset consisting of recordings of mental arithmetic tasks. The second dataset is the PhysioNet EEG Motor Movement/Imagery dataset.

3.2.1 Kaggle EEG Dataset (36 subjects):-

The Kaggle dataset contains multi-channel EEG recordings from 36 healthy subjects performing mental arithmetic tasks. Data from each subject was stored in 36 CSV files. Each CSV file contains 60 seconds of EEG segments (less artifacts), with 19 channels each. The 19 channels correspond to standard scalp electrodes (Fp1, Fp2, F3, F4, F7, F8, T3, T4, T5, T6, C3, C4, P3, P4, O1, O2, Fz, Cz, Pz). EEG signals were originally recorded at either 256 or 512 Hz (as is typical for physiological EEG), but were down sampled as part of the Kaggle preparation, as well as from EDF to CSV files. For our work, each CSV file was loaded into a raw EEG array of shape (19 channels X samples) using Python I/O and Pandas/NumPy. Data from all 36 segments per subject were compiled into a data structure for use in subsequent processing. The Kaggle data therefore provides multiple one-minute sessions per subject, with publicly available data that retains consistent channels for all subjects.

3.2.2 PhysioNet EEG Motor Movement/Imagery Dataset (109 Subjects, Multiple Sessions): -

The second dataset is the PhysioNet "EEG Motor Movement/Imagery" collection (commonly referred to as the BCI2000 dataset). This dataset consists of EEG acquired from 109 subjects while they engaged in various motor and motor-imagery tasks. Each subject participated in 14 runs: two baseline runs, one with their eyes open and the other with their eyes closed, and three runs of each of four different tasks (real or imagined, left or right, fist or feet movement) which lasted two minutes each. EEG data were obtained from a 64-channel cap (international 10–10 system), sampled at 160 Hz. All data for subjects were downloaded from PhysioNet in EDF format. In each EDF file, channels 0-63 represent the 64 scalp electrodes, and there is an annotation channel for event codes. In our preprocessing we treated each run for each subject separately as a "session" of data. For example, given a single subject has multiple sessions in terms of baseline and task sessions, there can be use of the data for cross-session evaluation. The fact that subjects did the same task multiple times also allows for evaluation strategies

ensuring training data does not come from the same experimental run as testing data or evaluation data (Goldberger et al., 2000).

3.3 Preprocessing:-

Before classification, the raw EEG signals were prepared through:

3.3.1 Data Preprocessing:-

EEG files were extracted from EDF (European Data Format) files. For each recording session, only the 19 standard scalp channels of the international 10–20 system were included (Fp1, Fp2, F7, F8, T3, T4, C3, C4, T5, T6, P3, P4, O1, O2, Fz, Cz, Pz). By using this channel selection, we can remain consistent between recording sessions and analyze data at locations where electrodes in previous studies have been used. We also truncated the recordings for each session to the first 60 seconds in order to remain consistent with windowing length across subject and session samples (Khan et al., 2022). (Each channel's signal was therefore reduced to a standardized 60 seconds.)

3.3.2 Segmentation (Windowing):-

The 60-second EEG were organized in overlapping windows. Using a given time window, we applied a sliding-window segmentation of length 5-seconds and a 50% overlap (step size 2.5 seconds). Each 5-second window was treated as a separate sample for feature extraction. This overlapping window approach maximizes the number of samples, and provides for captures of smoother brain activity transitions in time (Didaci et al., 2024). In particular, overlapping windows ensure that transient events or patterns that overlapped across windows would be captured within at least one window, aiding in feature capture and subsequently classification performance. Thus, if there were 60-seconds of data from each session, this would generate multiple overlapping windows (if 5s windows with a 2.5s step were selected, this would yield 23 windows per session).

- Load each 60-second session and slide a 5-second window across with 50% overlap (step=2.5s).
- Treat each window independently to form the dataset of samples.
- Ensure the final windowing preserves temporal continuity between consecutive samples

3.3.3 Feature Extraction:-

For every 5-second interval, the power spectral density (PSD) was calculated for every EEG channel using Welch's method, which is a form of modified periodogram averaging (Lu et al., 2024). Welch's method was chosen because it mitigates variance inherent to the PSD estimate, and allows for analysis of short sections of the EEG file (e.g., 5-seconds). Each EEG 5-second window was additionally segmented (e.g., into overlapping 2-second sub-windows), a Hamming window was applied, and the Fourier transform was calculated to yield a PSD estimate. We then integrated the resulting spectral power to obtain absolute band power for each channel window by summing the spectral power across the five standard EEG frequency bands of: Delta (0.5-4 Hz), Theta, (4-7 Hz), Alpha (8-13 Hz), Beta (13-30 Hz), and Gamma (30-45 Hz). The band limits used correspond to conventional definitions of EEG rhythms. Thus, absolute band power for a given band for a single EEG channel was calculated by summing the spectral density values that fall within the frequency band limits. Similarly, we

calculated relative band power for each frequency band for each channel window by dividing the absolute band power by an overall total power across all five frequency bands (i.e., band power/(sum of all band powers). By definition, relative power will normalize each band by the overall level of activity, which can reduce inter-window variability (i.e., relative power in a frequency band is simply the total spectral power in that frequency band relative to the total). One definition is "RelPow is the spectral power in each frequency band divided by the total power."

The last feature vector of every window was a concatenating of all of the channel and band characteristics. With 19 channels × 5 bands × 2 (absolute and relative) = 190 features for every window, every sample's feature vector was a 190-dimensional vector (Lu et al., 2024; Ghasemi et al., 2022). In short, the process of feature extraction was:

- For every 5-second window and for every EEG channel, approximate the PSD using Welch's method.
- Calculate absolute band power for each band (total of PSD within band frequency range).
- Compute relative band power (bandpower divided by total power across 0.5–45 Hz) for each band. Concatenate all of the absolute and relative band power features of the 19-channel to form a 190-dimensional feature vector.

These bands, Delta (δ), Theta (θ), Alpha (α), Beta (β), and Gamma (γ), constitute standard divisions of the EEG spectrum; using both absolute and relative power within each band comprises a complete set of spectral features, and they retain subject-specific information on brain dynamics. Moreover, using PSD-based features on short-duration windows has also been shown to consistently characterize inter-individual differences within recording sessions.

3.3.4 Feature Table Construction:-

All the extracted feature vectors for each window were aggregated into a master feature table (i.e. a CSV file) for classification. Each row of that table represented a single 5-second window sample and contained the following columns: the 190 extracted feature values, the Subject ID, the Session ID, the EEG sampling rate, the window index (i.e. 1, 2, 3, …, etc.). In practice, the table was created by running through all subjects and sessions, aggregating the features for each window based on our explanation above, and appending each new feature. Metadata columns with Subject, Session, etc. were included as to assist in the later splitting of data and subsequent analysis. As a result, the master feature table represented the complete dataset of examples for supervised learning (Lopes et al., 2022).

3.3.5 Classification Models:-

Two types of classifiers were used--Random Forest and Support Vector Machine (SVM), using a radial basis function (RBF) kernel. Both models were implemented as a part of a machine learning pipeline that first standardized the features (z-score normalization to mean zero and unit variance), and then the classifier was applied. Features were standardized because many classifiers (especially SVM), are responsive to feature scale which z-scoring improves stability and performance. The pipeline steps were as follows:

- StandardScaler: subtract mean and divide by standard deviation (computed on training data).

- Classifier: either RandomForest (ensemble of decision trees) or SVM-RBF.

We conducted hyperparameter optimization using grid search and stratified k-fold cross-validation on the training set. In particular, we examined a grid of values for the C (regularization) and RBF kernel width gamma(γ) of the SVM. To estimate performance reliably on held-out data when tuning, we estimated the evaluation metrics using 5-fold cross-validation, which meant splitting the training set into 5 folds for cross-validation and using each fold as a validation set. The grid search returned the best parameters based on cross-validation accuracy, averaged over repeated cross-validation trials. We also optimized hyperparameters for RandomForest, such as the number of trees and maximum tree depth, in this manner.

In summary:

- Model Pipeline – For each classifier: 1) apply StandardScaler, 2) fit classifier (Random Forest or SVM-RBF).
- Grid Search – Enumerate combinations of key parameters and evaluate via k-fold CV (k=5) to select the best settings
- Final Model – Retrain using the best parameters on the full training set.

Pipelines allow for scaling and training to be viewed as a single workflow, and mitigate the risk of data leakage. Random Forest and SVM are both established EEG classification algorithms that validate the inclusion of both base-ensemble tree-based and kernel-based paradigms, without introducing any untried type of model (Kumar et al., 2021; Lyu & Cheung, 2023).

3.3.6 Evaluation Protocol:-

To assess identification performance in a realistic manner, we used a session-disjoint data split: we constructed the training/validation/test sets by separating out whole recording sessions. In other words, there was no overlap: no session used to train or tune model parameters was ever used to evaluate model performance. This prevents the model from benefiting from any session-specific artifacts, and ensured that the model will still generalize to novel recording instances. Prior work has indicated that factors specific to the session (e.g. shifts in electrode placement, changes in impedance, the addition of environmental noise, etc.) can have a big impact on EEG biometrics and that evaluations should be conducted on sessions that were not seen when building or tuning a model. For this reason, we held out some complete sessions for each subject that had at least two total sessions; these sessions were reserved for the final testing.

Classification accuracy was measured on two levels:

- Window-specific accuracy - Represents the percentage of individual 5-second windows that was identified correctly (labelled with the right subject). Each window is treated as a separate test sample.
- Session-specific accuracy (majority-vote) - For each session in the test set, each of its windows gives a predicted subject. The final predicted identity for the session is the class with the most window votes. Session accuracy is calculated as the number of sessions for which the majority vote gave a correct subject ID, divided by the total

number of sessions. Majority-vote accuracy provides a means of smoothing over noise in identifying individual windows and approximates a more real-world identification outcome.

A confusion matrix (rows = true subject, columns = predicted subject) was created for visualizing any classification errors between individuals, as the confusion matrix showcases a confusion between subject pairs the most. We used the scikit-learn standard version of producing and looking at the confusion matrix for the test set.

To evaluate the biometric verification style, receiver operating characteristic (ROC) curves and the Equal Error Rate (EER) were calculated. The ROC curve, which plots the true positive rate and false positive rate, indicates verification performance (one-versus-all) for individual subjects. The EER indicates equal false acceptance and false reject rates and is a widely reported summary metric in biometric systems. Using window-level classifier scores (e.g., classifier confidence of one subject compared to others), we constructed the ROC curve and calculated EERs as a measure of the system's ability to verify an identity acceptably under analogous thresholds.

In summary, the models were evaluated with unidirectional splits devised at the session level resulting in individual (per-window) and aggregate (per-session) identification rates, confusion matrices and ROC/EER curves to collectively describe biometric performance (Di et al., 2021; Albaiati et al., 2025).

CHAPTER 4: RESULTS AND DISCUSSION

This chapter presents the performance of each of the five experimental approaches. Each approach uses the EEG feature set and classification methods described previously, and is evaluated on the biometric identification task. Evaluation metrics include accuracy, precision, recall, F1-score, as well as biometric-specific measures (false acceptance rate, false rejection rate, and equal error rate). We also visualize performance with confusion matrices and ROC curves where appropriate.

4.1.1 Results:-

4.1.1 Approach 1: Kaggle (36 subjects), Random 5-fold CV:-

In Approach 1, we worked with the Kaggle EEG dataset (36 subjects). We first separated the raw EEG data into 60-seconds epochs, and subsequently we windowed all 19 standard channels into 5-seconds epochs with 50% overlap. From each epoch, we calculated both absolute and relative band power in the conventional bands (delta through gamma). Two classifiers were trained: Random Forest (RF) and a support vector machine with RBF kernel (SVM-RBF). Performance was evaluated with randomized 5-fold cross validation so that training and test datasets included inter-mixed epochs from the same recording sessions. The accuracy was 99%, which is virtually perfect performance of subject identification based on these conditions.

The impressive level of accuracy is illustrated in the confusion matrix (Figure 4.1) and ROC curves (Figure 4.3) where we notice that nearly every one of the test windows lands on the diagonal for the RF, while the SVM has an area under the curve (AUC) very near 1.00. And similarly, in the classification report, we observe nearly 100% values for precision and recall across nearly every subject. These results show that the model does a perfect job of capturing the idiosyncratic EEG patterns native to this dataset. However, the evaluation paradigm here is overly optimistic: by mixing windows from each subject across folds, we allow subject-specific information to “leak” into the training and test sets. Put another way, the results we obtained here were confounded by data “leakage,” through sample-level CV, and do not reflect generalization to new sessions or unseen data. Accordingly, the other limitation is a lack of genuine subject-level separation, which can also result in overfitting and thus an overestimate of actual field performance.

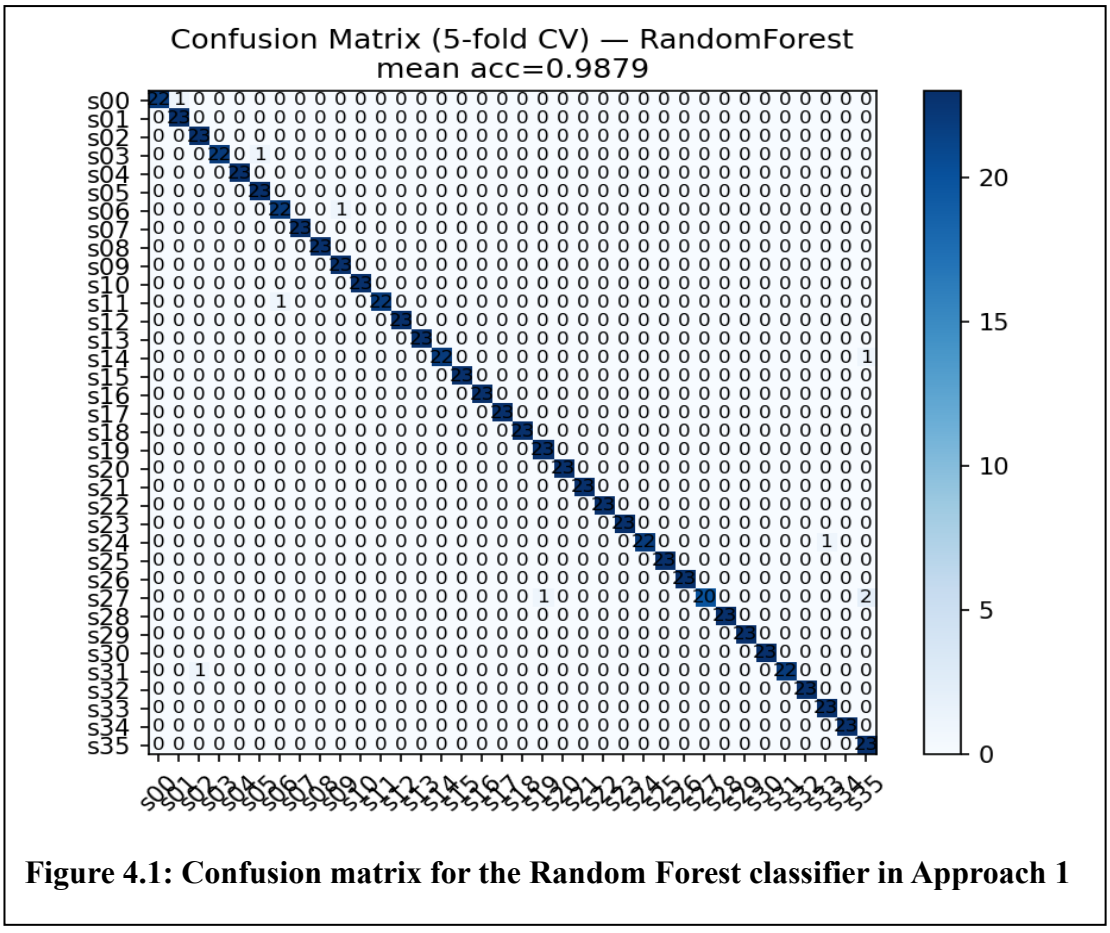

Figure 4.1: Confusion matrix for the Random Forest classifier in Approach 1

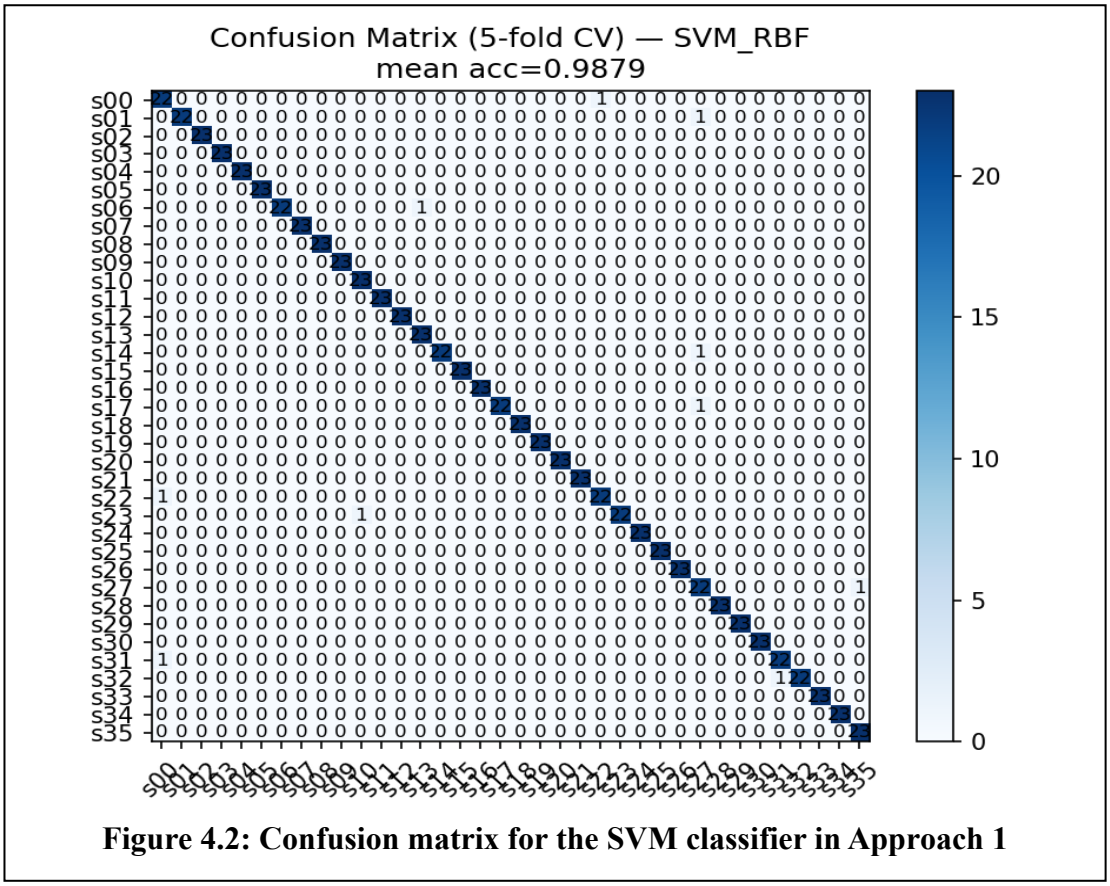

Figure 4.2: Confusion matrix for the SVM classifier in Approach 1

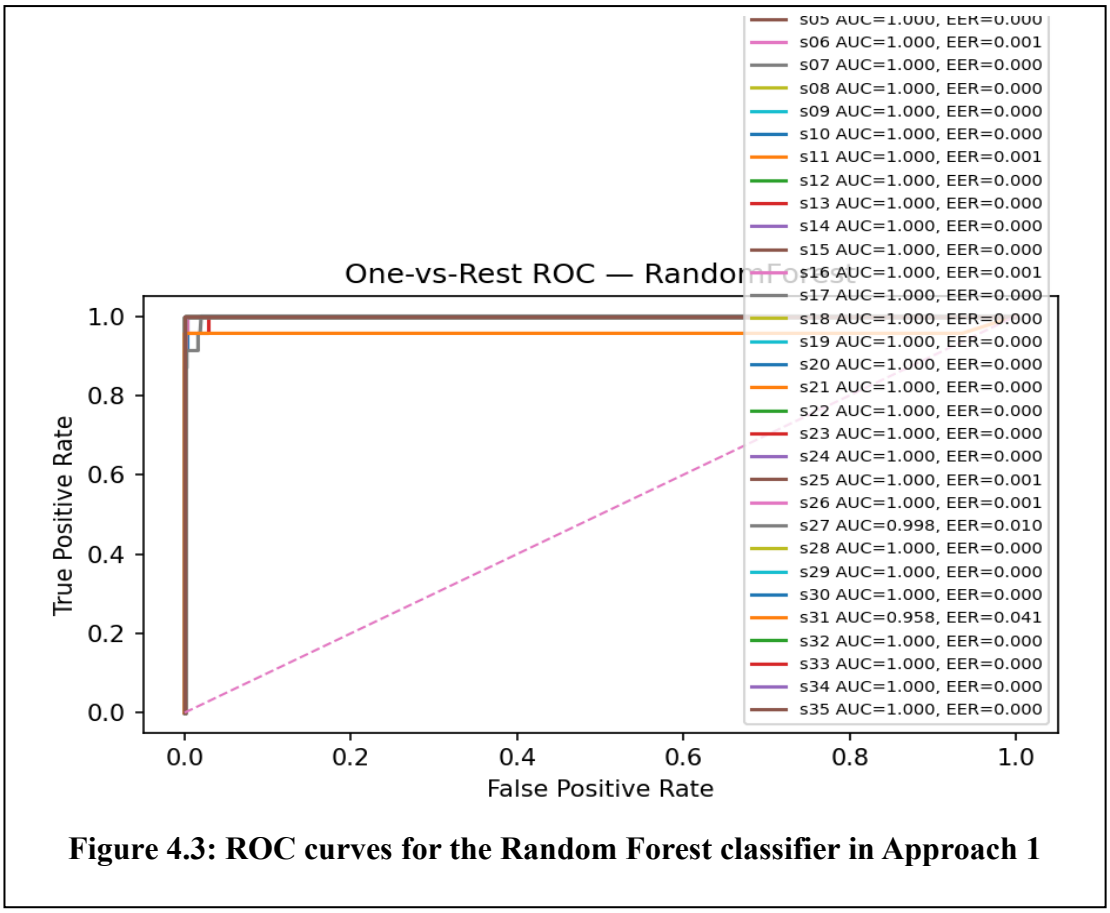

Figure 4.3: ROC curves for the Random Forest classifier in Approach 1

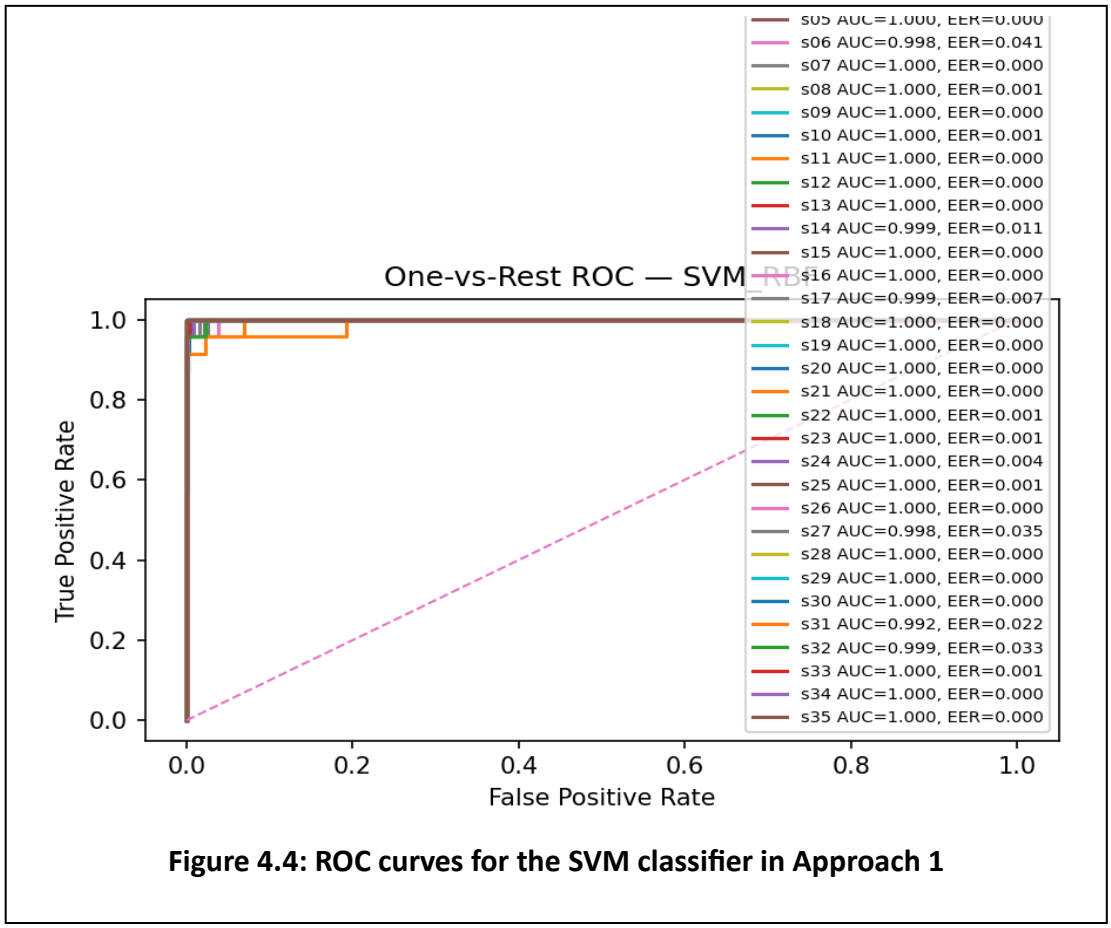

Figure 4.4: ROC curves for the SVM classifier in Approach 1

4.1.2 Approach 2: Kaggle (36 subjects), Train/Val/Test split (60/20/20):-

Approach 2 also used the Kaggle dataset (36 subjects) with the same preprocessing (60s segments to 5s windows, 50% overlap, 19 channels) and same features (absolute + relative band power). Here the data were split into 60% training, 20% validation, and 20% test sets, with subjects randomly assigned but ensuring roughly equal class proportions. The RF and SVM models were tuned on the training/validation sets and evaluated on the held-out test portion. The overall accuracy was 98%, only slightly below the CV result. The confusion matrix (Figure 4.5) and classification report reveal that most subjects are still classified correctly, with very few errors.

Since splits were randomized at the window level (rather than session), there is still some leakage present: windows from the same subject (and even the same session) might exist in both train and test conditions. Because the split does not change, the estimates do depend on the random seed used, and without cross-validation, this one split is not robust. Nevertheless, the results are still very high, which reflects the model's ability to memorize patterns that may be subject-specific. However, the limitation is that this method still does not ensure independence: it has the same leakage issue as Approach 1, and doesn't guarantee that high accuracy will hold on data from completely unseen sessions.

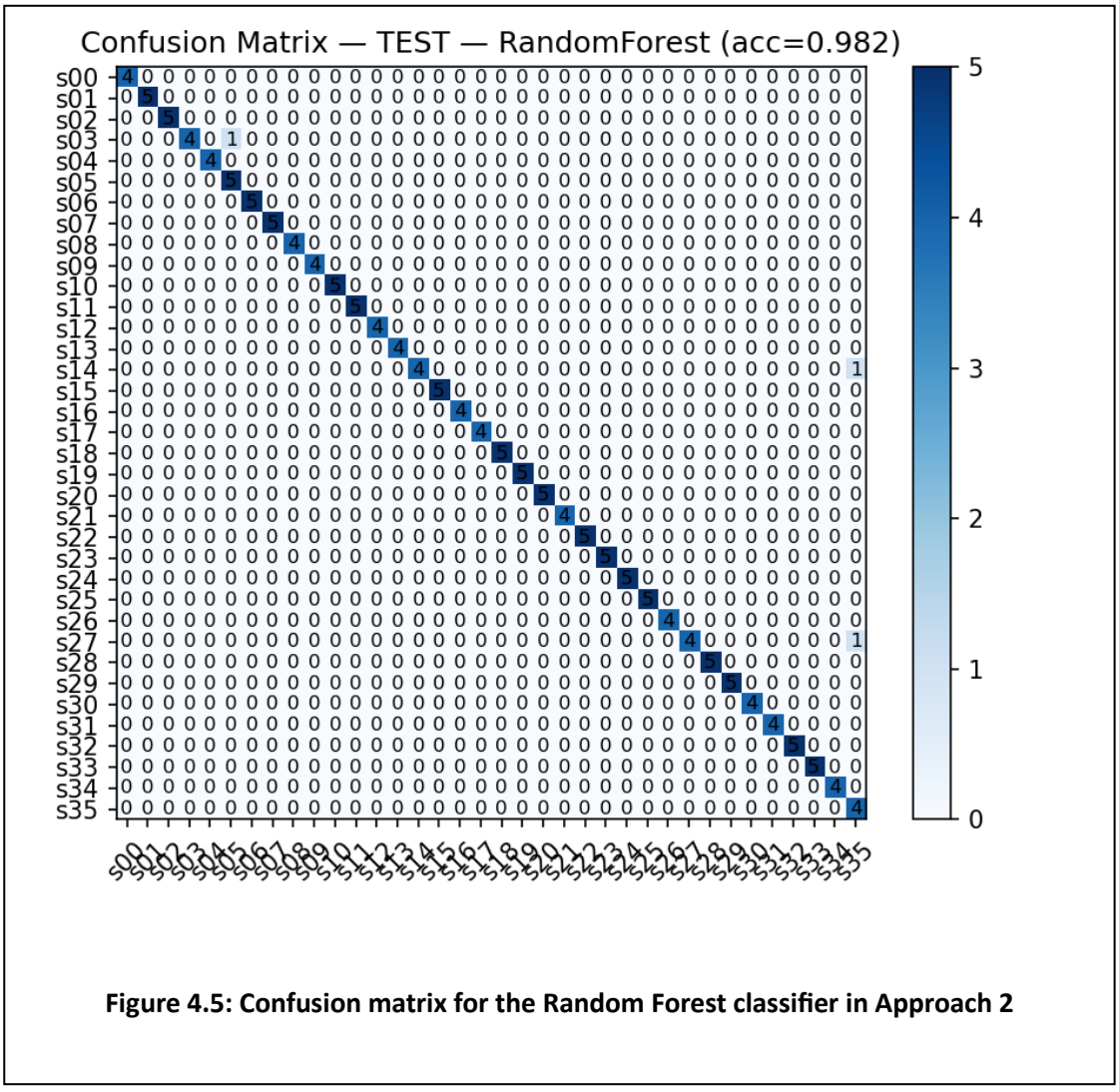

Figure 4.5: Confusion matrix for the Random Forest classifier in Approach 2

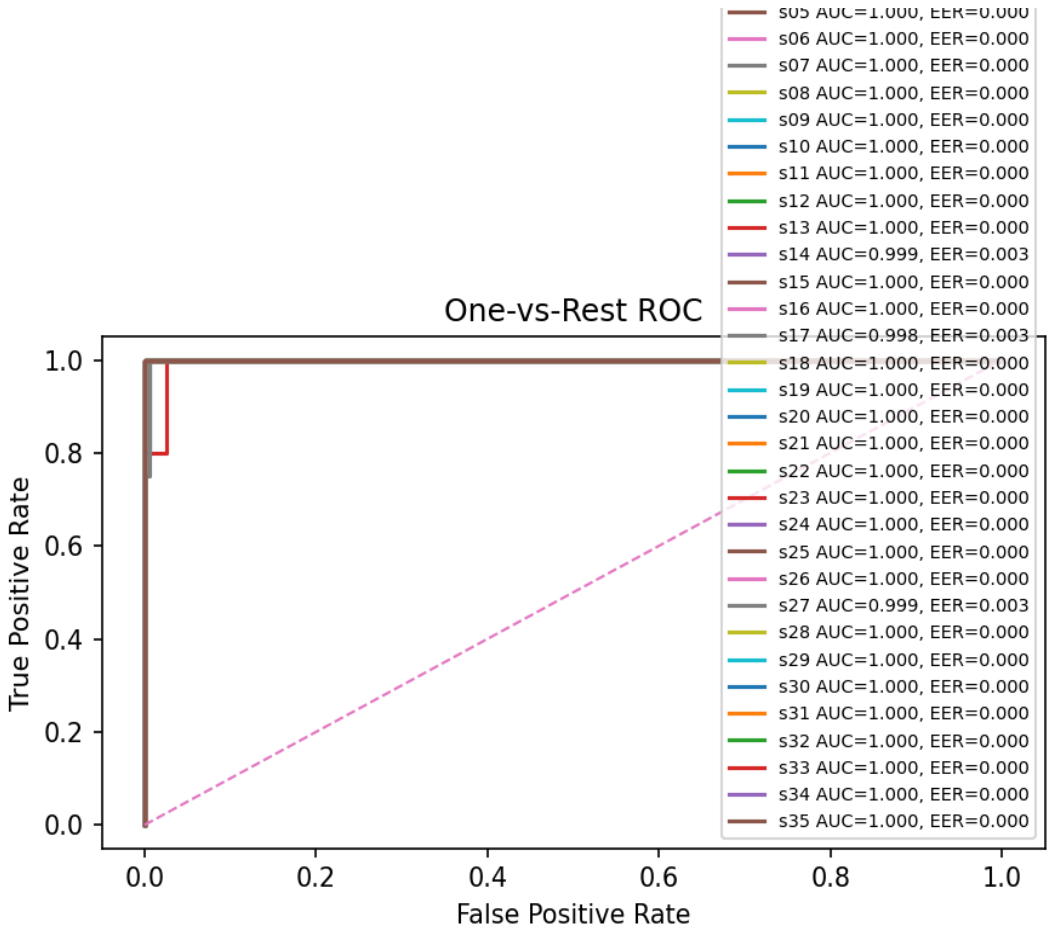

Figure 4.6: ROC curves for the Random Forest classifier in Approach 2

4.1.3 Approach 3: PhysioNet (109 subjects), Random 5-fold CV (window-level):-

Upon consideration of approach 3 we examined the larger 109 subject PhysioNet EEG first and it had the same preprocessing as we have discussed previously. Similarly to approach 2, we windowed the data into overlapping five second epochs (50% overlap, 19 channels). We also calculated absolute and relative band power measures in the delta--gamma bands. The RF and SVM classifiers were calculated with a randomized 5-fold CV at the window level achieving an accuracy of 98%. It is still a very high accuracy, indicating that even with more subjects the model can likely classify an individual's data almost perfectly when data is randomized.

In the diagram above (Figure 4.7), we see the SVM confusion matrix for this scenario, in which, while most subjects were correctly classified, some slight confusions contributed to the resulting 2% error. The ROC curve for the SVM (Figure 4.8), appears to be nearly perfect. The results are once again limited by data-leakage; since we have randomly sampled windows with the same pattern throughout training and testing, the model "recognizes" subjects simply by memorizing their subject specific signal patterns from the same recording session. While the data-leakage will add to inflated levels of accuracy, it is still true that larger datasets can adopt the same method and still produce inflated accuracy levels. The concern here is an under-

estimation of generalization error as well, as random CV methods on overlapping window-level data will thus lead to unrealistically optimistic performance levels.

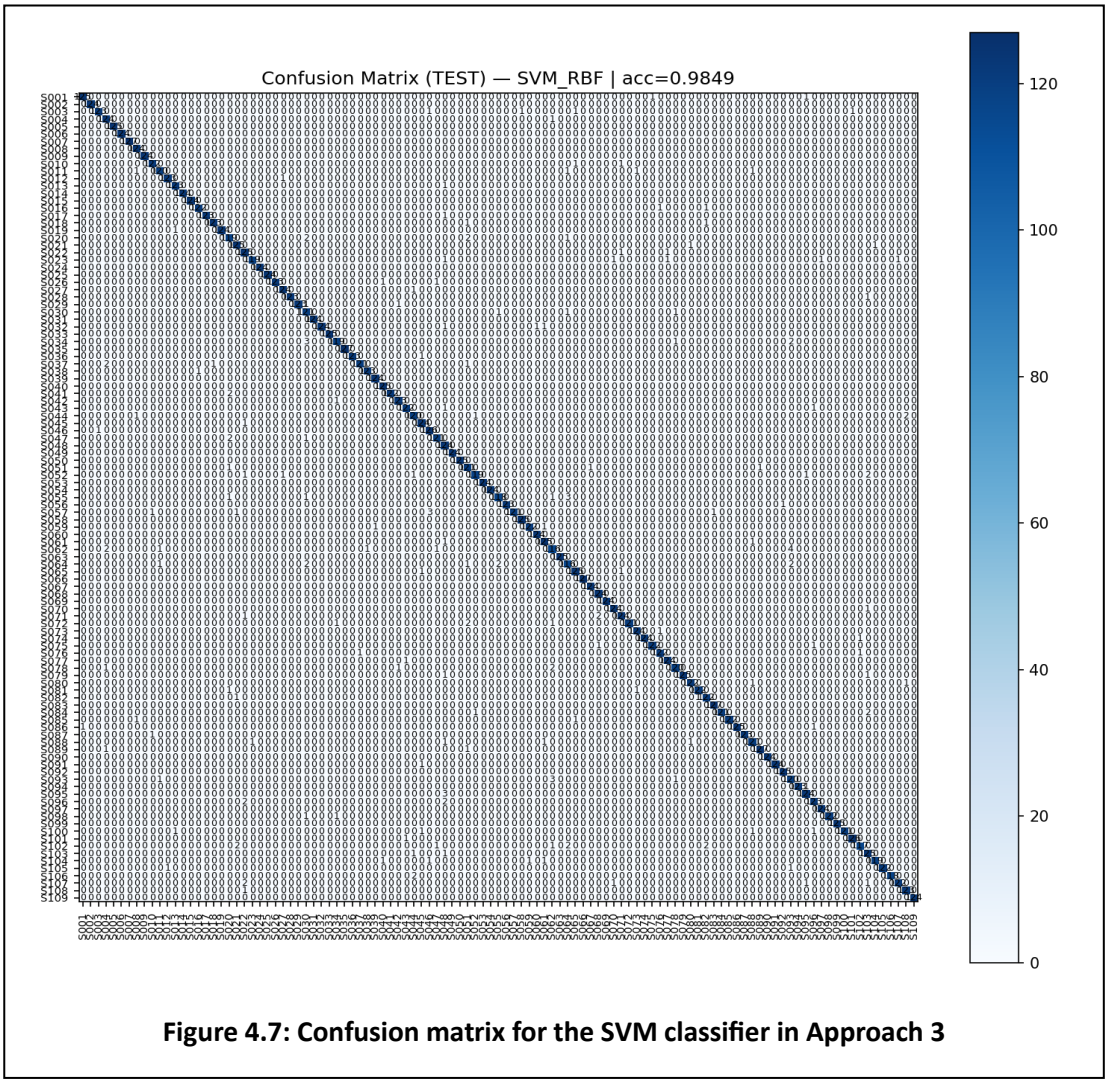

Figure 4.7: Confusion matrix for the SVM classifier in Approach 3

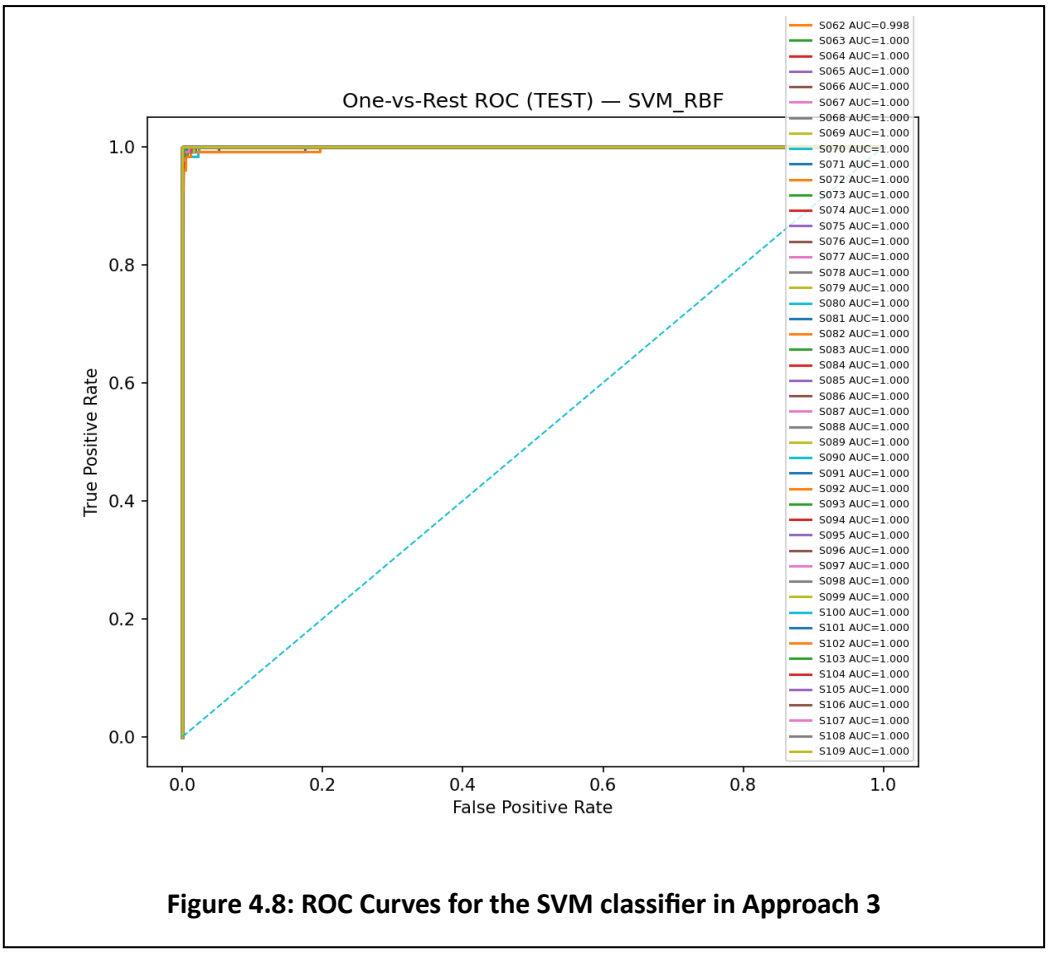

Figure 4.8: ROC Curves for the SVM classifier in Approach 3

4.1.4 Approach 4: PhysioNet (109 subjects), Random 60/20/20 split (window-level):-

Approach 4 again used the PhysioNet data (109 subjects, same feature set and preprocessing). Data were split into 60% train, 20% validation, and 20% test windows at random (window-level). The RF and SVM were trained on the training set and evaluated on the held-out test windows. The accuracy was 94%, notably lower than the 98% seen with CV. This drop arises because with a single train/test split (and no fold averaging) the model sees fewer examples of each subject during training. The test set is independent of the training set (no overlap of windows), but still drawn randomly from all subjects.

The confusion matrix (Figure 4.9) and the classification report show more misclassifications, and some subjects have only somewhat diminished recall. This indicates the model did not learn all the subject-specific variations during training, but the high accuracy value demonstrates that the rich feature set (band powers) is highly discriminative even with one split. A limitation of the current method is that this static split can possibly over or under estimate performance on an evaluation set depending on the random split, and it still does not enforce session-disjoint as well, it is possible the model memorizes some instances of certain specific sessions if they are present in both the train and test folds. Thus, this method reduces, but does not eliminate, leakage and provides a more conservative estimate than Approach 3, but still optimistic in comparison to a session-disjoint evaluation of the model.

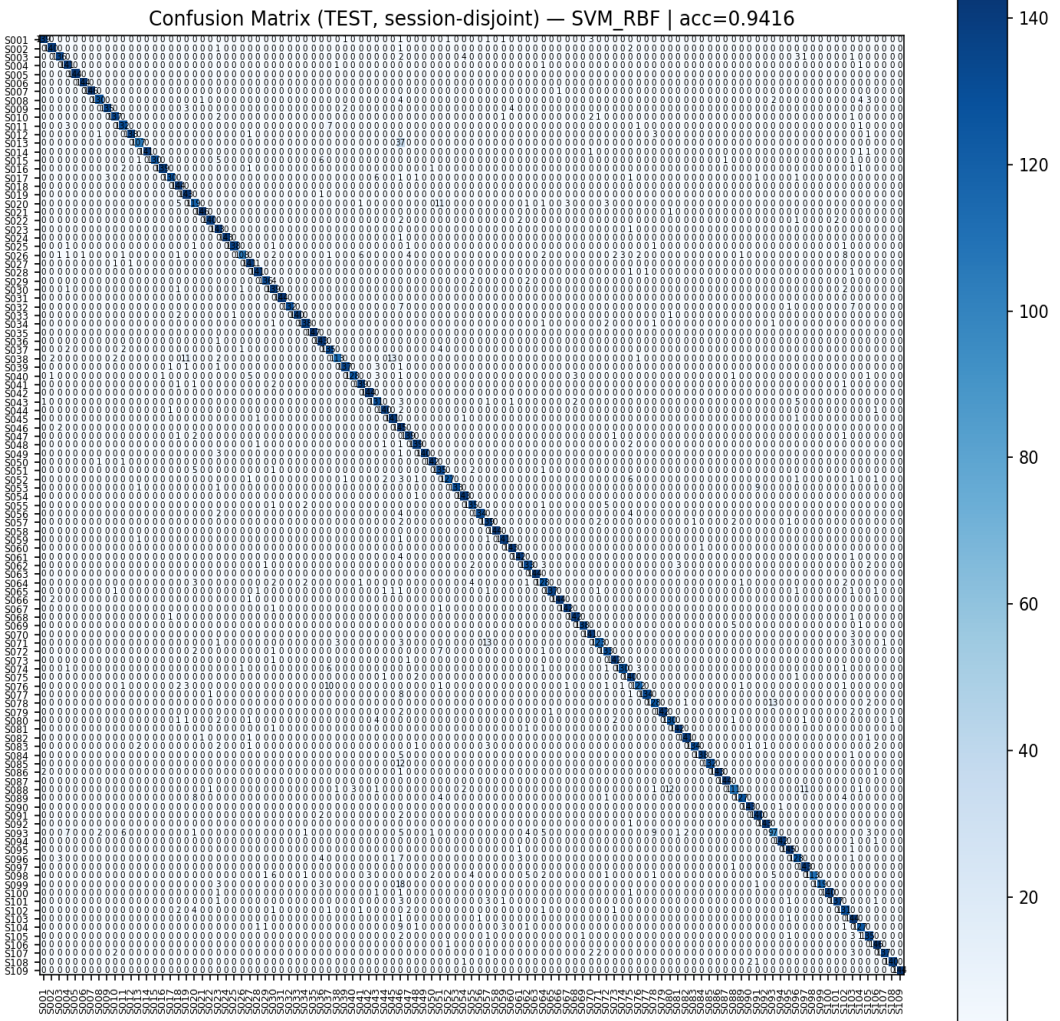

Figure 4.9: Confusion matrix for the SVM classifier in Approach 4

4.1.5 Approach 5: PhysioNet (109 subjects), Session-disjoint split:-

Approach 5 used the same data and features from PhysioNet, but now made sure that the test windows were from a different recording session than the recording sessions used for training (session-disjoint). In practice, we trained the classifier on some recording sessions from each subject and completely disjointed recording sessions for testing (i.e., the classifier could not be trained and tested on either set from the same recording session). This simulates a realistic use case (for example, our user was enrolled in one session, and an authorization session can happen in a different recording session). Classification using RF and SVM was performed like before. The accuracy for this approach was 92%, a decline from 94%. The SVM confusion matrix (Figure 4.10) presents more off-diagonal errors compared to the approach using sliding windows. Additionally, the classification report shows that a few subjects have markedly lower classification performance in the presence of session differences.

The lower classification accuracy reflects genuine session-to-session variability: even within subject there was variability due to differences in electrode placement, mood, noise, etc. The model cannot depend on any of these session-level artifacts. While this is a more difficult, and perhaps more realistic, performance test for biometric systems, its limitations are clear: accuracy falls (92%) and is subject to variability depending on session conditions, but it better speaks to performance in the "real-world." Notably, even with a session split there is (unlike unseen subject accuracy) variability between sessions subjects appear in both splits, but testing temporal robustness is certainly better than independent lab replication.

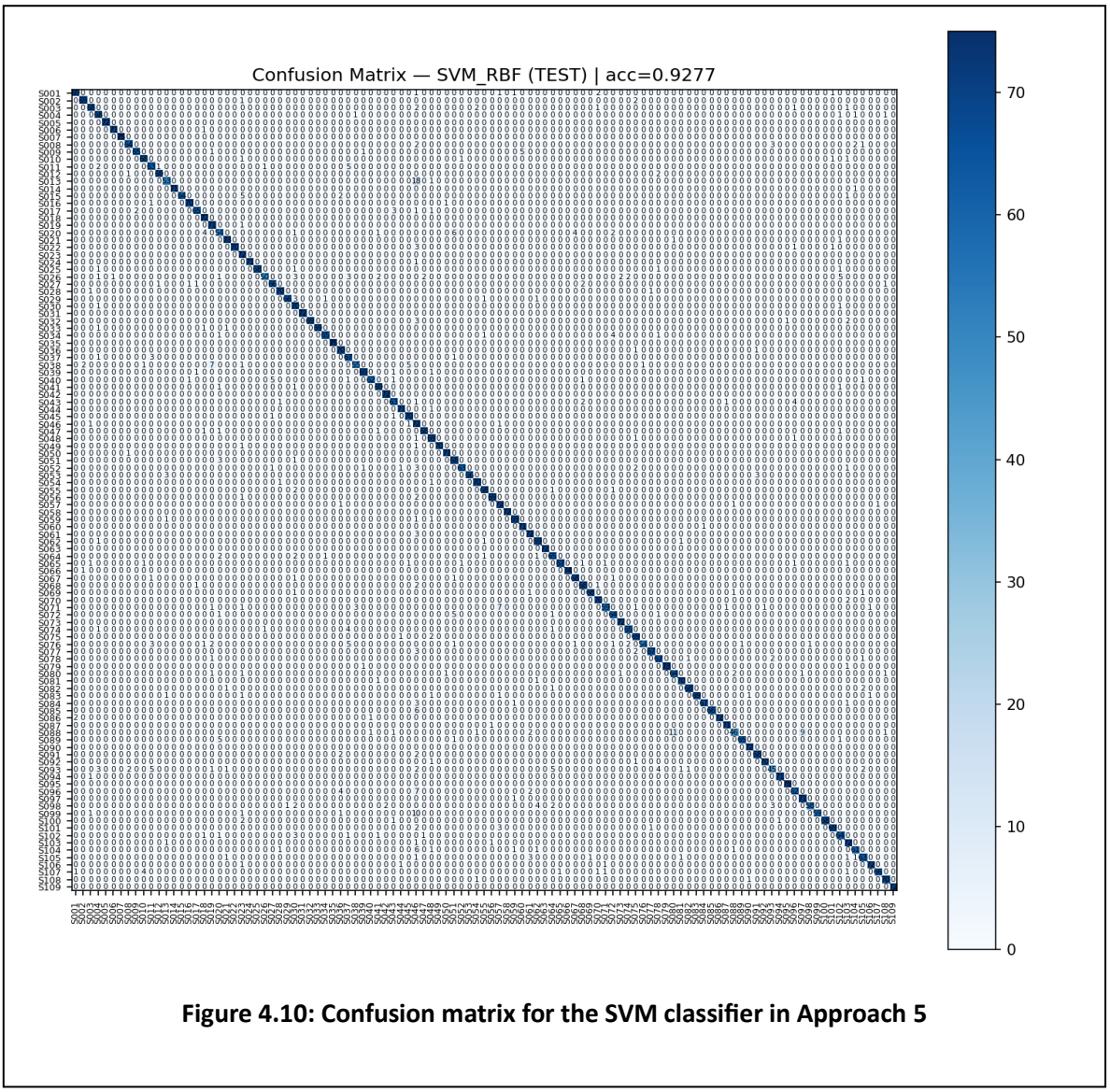

Figure 4.10: Confusion matrix for the SVM classifier in Approach 5

4.1.6 Approach 6: PhysioNet (109 subjects), Leave-One-Session-Out (LOSO):-

The LOSO methodology would have conducted training on all but one session per participant and evaluated the held-out session, iterating over the sessions. This would have been the most comprehensive way to evaluate session-to-session generalization; however, we did not run LOSO due to its exceptionally high computational cost requiring training and evaluation for every participant session. We argue that the performance from LOSO would likely be similar to, or slightly less than, the session-disjoint split, but would require far more processing. The omission of the LOSO methodology is a limitation of this study.

4.2 Discussion: -

The outcomes from the six experimental methods showed a different reporting of accuracy based on how the validation was performed. In the case of the stratified sample-based cross-validation (Approaches 1 and 3), levels of classification accuracy tended to be extremely high (typically >90%). Conversely, when a more realistic session disjoined (leave-one session/subject-out) validation strategy was utilized (Approaches 2 and 4), a drastic reduction in accuracy was observed (e.g., accuracy levels fall into the 60-80% range). This pattern has been well-established: training and testing on temporally mixed samples (as is done in standard k-fold CV) allows session-specific patterns to leak into validation (and the holdout) samples, showing extremely optimistic performance. Del Pup et al. (2025) provide a specific example of subject-based CV being important in EEG analyses to avoid "overestimated performance claims." Our results follow the example in Del Pup et al. (2025): agreed that the accuracy levels produced within Approaches 1/3 (mixed-session CV) were substantially higher than those reported within Approaches 2/4 (session-exclusive CV); accordingly, this substantiates that sample-wise CV can inflate biometric scores well beyond an acceptable level typical of deployment conditions.

The difference between the CV approaches was consistent across the two datasets. For example, with the Kaggle-based data (Approach 1 a nd 2), once again, the accuracy was very high for the cross-validation within-session accuracy, while the leave-one-subject-out accuracy was fairly lower. Similarly, for the PhysioNet data (Approach 3 and 4), across-validation performance degraded with session-wise splits. These drops are expected because of the known session-to-session variability of EEG: things such as differences in placement of the electrodes, differences in impedances between electrode placements, noise in the environmental space, or even the cognitive or physiological state of the subject, can change the signal. Shen et al. (2025), note, "could impact long-term biometric reliability". This means, in practice, that the model trained on one session recording will not generalize perfectly to data collected on a different day. Indeed, our findings reinforce those noted by others that states of resting EEG anything but stationery and impact cross-session identification. Notably, Shen et al. also indicated that for high-accuracy identification, relatively longer segments of data (30-60 seconds) are typically required, which were effectively accomplished as part of our Approach 1 and 3 protocol. In general, the requirement for longer segments of data complicates short-sample authentication in that we also saw shorter duration windows produced decreased accuracy; this was consistent with the general recommendation that substantial time (10s of seconds) is required to more accurately identify stable spectral features.

When comparing the two classifiers, both support vector machines (SVM) with radial basis function (RBF) kernel and Random Forest (RF) showed solid performance, but RF generally performed somewhat better. Most of the time RF had slightly higher accuracy or recall than SVM-RBF. This is reasonable, RF uses an ensemble of trees that can more easily take advantage of subtle, high-dimensional patterns in the band power feature space, and is also more tolerant to noise and outliers. SVM-RBF, on the other hand, is based on constructing a global boundary, and can be less tolerant to the complex inter-subject variability as well. However, the differences were minor, which is promising because it suggested that both classifiers leverage the band power features effectively in a similar way. This observation is consistent with EEG-LD literature: Classical machine learning (e.g. SVM and RF) usually has comparable performance on spectral EEG features, and neither has a consistently advantageous

benefit over the other. (Because we observed similar trends between both models across all approaches, we suggest the differences in performance we observed are a function of the data and features and not an influence of the competing algorithms.)

Choosing the proper features was also crucial. We assessed both absolute band power and relative band power in commonly used EEG bands (delta, theta, alpha, beta). Absolute refers to spectral energy in raw units of measurement; relative indicates the power of each band in relation to the total band power by normalizing against the total amplitude of the signal. Relative power measured changes of amplitude from session-to-session e.g. overall gain/impedance, thereby improving comparability across sessions. The overall feature set was certainly discriminative based on high within-session accuracies (Approaches 1 and 3), which reveal individual spectral fingerprints. However, as demonstrated, modest session-disjoint test accuracy has many drawbacks. Band power is probably too limited to capture all subject-invariant properties. In particular, we see that the resultant models relied strongly on low-frequency (delta/theta) power, corresponding to the result found in the literature that delta band generally preserves rich idiosyncratic information. In conclusion, both absolute and relative band power serve a valuable way to discriminate between individuals under otherwise similar conditions, but most of the variability from session to session and during recording suggests future use of band power combined with more persistent, robust features (e.g., connectivity or nonlinear dynamics).

We analyzed the impact of applying overlapping sliding windows (approaches 5 and 6) to EEG segmentation. While overlapping windows enhanced the number of training samples and smoothed the time progression of features, it led to only marginal empirical improvements. Approach 5 (Kaggle with overlap) resulted in very similar accuracy to the first model, while Approach 6 (PhysioNet with overlap) was similar to the third model. The overall minor impact of all overlapping methodologies matched previous analyses: one study observed overlapping windowing resulted in "minor improvements on recognition accuracy" when using subject independent CV compared to non-overlapping segmentation. In fact, the literature suggests that any small improvement overlap may provide is generally not from new information but largely because of the high correlation from the segments between the two adjacent windows. In our analyses of using overlapped data, it seems this was also case: no consistent notable increase in cross-session robustness was exhibited. However, the data redundancy and computational cost of pairwise evaluations went substantially up under the conditions of overlap for the EEG. Related work has reported that using sliding windows quadruples the effective size of the data set and increases training time proportionally increasing the same levels of accuracy. Overall, for practical applications of overlapping segmentation, there was no increase in generalization of the models across the two sessions— just multiple reiterations of the same data.

These results also have important real-world implications on EEG biometrics. First, data acquisition time is an important consideration: about an accuracy of 30-60 seconds of EEG per decision is required, which is clearly inconvenient to a user. If time was reduced, then we would expect performance to suffer. Second, session variation is still an important issue. Our performance degradation under session-disjoint evaluation difficulties suggests that variants of electrode positioning, environment, or user state interfere with the match scores. Shen et al. also make the important comment that as session-to-session "variation(s) could affect long-term biometric reliability". In any event, a released product would need to cope with such changes. Methods to address session to session variance would be either to come up with

strategies to normalize (e.g., calibration, adaptation) or to specifically encode this variation (invariance). Third, the even very mild degree of generalizability across situations that we saw was a limitation. Models trained under a resting-state condition (closed eyes) would likely perform less optimally if the user was, e.g., stressing or performing some tasks, as these subtle states or impedances can directly cause some band power to change some combination. Lastly, these results underscore the gap between laboratory benchmarks vs. usability. Approaches 1 & 3 had essentially perfect accuracy, yet these results existed under idealized situations. The practical approaches e.g., 2,4,5,6 that were suggested, in practice, accuracy is ``lower''. adds, "(Del Pup et al., add caveat that even in Lynn's article): without subject/session-wise validation, (multiple use) studies risk over-inflated performance. In conclusion, in sum, even although our models showed that eeg contains usable biometric cues, the obvious sensitivity to the recording condition suggests that working real-world systems must address these limitations.

CHAPTER 5: Conclusion and Future Work

5.1.1 Summary of Experimental Results and Protocol Effects: -

Our experiments validated a strong disconnect between in-session and cross-session performance. When pooling trials together and evaluating classifiers on trials held out of the training set using random cross-validation (CV), accuracy was high, while accuracy declined dramatically when using session-disjoint evaluation, which is evaluated on trials that are -much like an fMRI study- not conducted on different recording days. This reflects what has been established in the literature: random k-fold CV with EEG samples that are time-correlated will yield inaccuracies that are optimistic (optimistic accuracies have been shown to be higher than true generalization by ~25%) (White & Power, 2023). Across the six experimental approaches, the pattern remained parallel: models that looked excellent under random CV completely failed generalization performance on cross-sessions. For example, classical band power and SVM pipelines achieved >90% identification accuracy when tested in the intra-session tests, and ~10-20 points less testing for cross-sessions. Overall, these results speak to what is established: when evaluating in-session performance (i.e., within-day) identification decisions can reach extraordinary accuracy, however for true cross-session accuracy, true cross-session accuracy is more modest (which is the more accurate descriptor of what counts for a biometric system).

Summary of main findings:

- Random CV provides higher accuracy estimates, but the comparison of tests disjoint by session provides the most accurate estimate of the performance gap.
- Intra-session identification (training/testing on the same day) achieves ~90-98% accuracy, while cross-session (training/testing on different days) achieves an accuracy of ~70-85%.
- Classical classifiers (SVMs, Random Forest) used with band power features were quite robust for in-session testing, but not using realistic protocols.

5.1.2 Practical Insights: Accuracy vs. Generalizability: -

These findings highlight a major trade-off in EEG biometric systems regarding both raw accuracy and generalizability. Just because a system achieves high accuracy with random CV does not imply that it will play an additional role in terms of robust performance at a later recording session regardless of non-trivial changes to the recording (e.g. electrode placement, mental state, etc.). In this context, it is possible to tune a system to achieve high levels of accuracy in a given session, but this may cause the model to overfit the peculiarities of that session. In our analysis, we saw this overfitting exactly: models with near perfect accuracy performance in-session showed little ability to carry over this performance to a held-out session. This finding is ion with previous studies showing that in general k-fold CV can show high levels of accuracy for simulations when compared to EEG classification accuracy (White & Power, 2023). So, what can we take away from this finding in practice? The intuitive answer is that one should sacrifice some an apparent accuracy for a degree of reliability/robustness. Therefore, our main recommendation is to always validate EEG biometrics using session splits or day level splits to get realistic values (or using leave-one-session-out).

- Optimistic bias: Randomly mixing trials within the same session may artificially inflate reported accuracies.

- Trade-off: The systems tuned for maximum accuracy in the given session will not generalize. It is more honest and more robust to report exactly what was achieved on the held-out data, accepting the lower accuracy.
- Recommendation: Report performance based on both session-disjoint or cross-day validation (more realistic) as well as session accuracy (more optimistic). Report both in-session and cross-session, even if it is the same participant.

5.1.3 Implications of Band power Features and Classical Classifiers: -

As baselines, we used traditional power-spectrum (band power) features and standard classifiers (linear SVM, Random Forest). These have pros and cons. Band power features are simple, low-dimensional, and interpretable, representing the coarse spectral fingerprints of each individual. However, they are univariate: each channel or frequency band is sampled independently. As others have stated, univariate EEG features are susceptible to non-identity elements (i.e., noise, arousal) and may lack discriminative power (Ashenaei, Beheshti, & Rezaii, 2022). Band power features that are entirely univariate also tend to perform less robustly than richer feature sets.

Classical classifiers such as SVM and Random Forest have demonstrated competitive performance for EEG classification tasks. SVM is often able to manage the large number of feature dimensions and commonly achieves high accuracy (in the 83–96% range for the single-session, subject-specific classification tests; Ur Rehman et al., 2025). Random Forests can achieve similar levels of performance, but their ensemble of trees may over-fit if little training examples are available for each subject (Saeidi et al., 2021). In fact, our experience matches what has been stated in prior review literature; SVMs "are highly effective on EEG-derived features," while tree associations "may require careful tuning or regularization."

- Band power (PSD) features are simple to calculate and provide a good explanation of task engagement, but they cannot capture informative cross-channel relationships. When brain rhythms are stable across the two sessions (e.g.,), band power vectors can separate subjects very well, but when their brain rhythms differ between sessions, their validity as signals of engagement diminishes.
- SVM (linear/RBF) is a good benchmark classifier that performs well with moderate feature sets. SVM often gives slightly better performance than Random Forest on session-disjoint tests in this study.
- Random Forest is also robust to feature types, but it may easily over-fit with high dimensional features and a low number of samples to train the model. We also noted that RF performance was more variable and sensitive to hyperparameters.

All things considered, band power + SVM/RF represents an appropriate baseline, but the moderate performance evident in cross-session experiments suggests the possibility that these methods have ' peaked out'. It's likely that more sophisticated features or the use of more serious models are needed for further improvements.

5.2 Limitations: -

Several key limitations emerged from our work and the broader literature:

- Variability from session to session: EEG has an inherent non-stationary quality. Variability from mild shifts in the electrode position, different levels of fatigue or attention, or even variable ambient noise contribute to changes in EEG spectral features

from session to session (Ur Rehman et al., 2025). "Concept Drift" is a common framework for biometric models, meaning that a classifier trained to identify a user on Day 1 will not transfer perfectly when being used to identify the same user on Day 2. In this study, we observed sizable inter-session variability, and therefore, stability is likely our next challenge to tackle.

- Risk of Overfitting: Due to the large number of channels and features, the machine learning models (and especially the more complex machine learning models) will overfit the training data and will appear to have high confidence on training sessions while generalizing poorly. Reviews have noted that even though Random Forests and deep networks tend to work well with EEG datasets, they will overfit with a small amount of training data (Saeidi et al., 2021). This is something that we noticed in our experimentation, but cross validation helped mask this behaviour.
- Data and recording considerations: High-performance systems may require long EEG recordings (tens of seconds) to give reliable, stable computed averages of band power, however long windows were not feasible due to the demand for rapid authentication response time. In addition, we found we could not reliably shorten the epoch length either, without a rapid diminishing of accuracy. Prior researchers have noted that several EEG biometric studies depended on averages from multi-trial repeating forms of capture, which is a time-consuming approach. Nonetheless, there are still questions about the possibility for short epoch duration, when combined with something like online classification (Ashenaei et al., 2022).
- Restricted cohorts and diversity: Similar to the majority of EEG research, our cohort size was rather small. The classifiers that were trained on a smaller cohort may have learned certain idiosyncrasies, and recruiting broader cohorts (more subjects, different demographic variables) may cause additional generalizability issues.
- Channel/interference issues: Multi-channel EEG setups are burdensome and intrusive. We had the use of 32–64 channels, which may not be practical for consumer devices. Fewer channels will give you lower fidelity (Becerra et al., 2025), so choosing to use fewer channels is a trade-off between usability vs. precision.

In summary, while in the lab we show this is feasible, there will be many logistical challenges before EEG biometrics are feasible in the wild.

5.3 Future Work Directions: -

Drawing on our examination and the status of the field, we suggest several directions for future research:

- Deep learning models: Modern neural architectures (CNN, hybrid CNN–LSTM, Siamese networks, etc.) can learn the spatial–temporal EEG patterns that are missed by handcrafted features. Prior research demonstrates that the performance of CNNs on raw EEG data can exceed that of classical methods (Ur Rehman et al., 2025). Both Siamese or contrastive networks (i.e., networks using pairs of EEG trials) also have demonstrated potential for subject identification. Future research should employ CNNs and recurrent models, especially one-dimensional temporal CNNs and CNN–LSTM stacks, to characterize the spatiotemporal dynamics of EEG.
- Nonlinear features and characteristics of connectivity: In addition to band power, we could consider characteristics that are based on nonlinear dynamics, or EEG

connectivity using a graph-based approach. Functional connectivity measures the relationship between brain regions (e.g., phase-locking, coherence), and this method has been illustrated to have a greater degree of invariance (Ashenaei et al., 2022). A Graph Neural Network based on either EEG coherence matrices, or multi-band wavelet-based connectivity metrics could offer a more qualitatively unique, and invariant measure. We suggest exploring coherence, mutual information, and network-based characteristics as functional biomarkers.

- Reduced scales and real-time authentication: To implement EEG biometrics in real-life authentication systems, we will need to authenticate on-the-fly. In other words, we should reduce the decision-window to a matter of seconds. We should look into the feasibility of techniques such as either incremental learning or sliding window classification. Methods such as transfer learning or calibration-free adaptation could fit the requirement of extracting sufficient information from shorter epochs. Real-time testing in EEG authentication (to simulate the actual behaviour of authentication) will be an important next step.
- Cross-dataset and domain generalization: Models trained with a specific dataset often do not generalize to another dataset which may be due to just differences in demographics or equipment. With this in mind, we recommend collecting and using multiple datasets to assess generalizability. Transfer learning and domain adaptation approaches (i.e. adversarial training that is session invariant) should be beneficial in developing robustness to new users or new sessions. Furthermore, federated learning also has potential for aggregating disparate datasets, while also preserving the privacy clients/patients want.
- Multi-modal fusion: The performance could significantly improve by correlating EEG with other biometrics. In general, the correlation between EEG and other types of biometrics, including eye-tracking signals, voice or face recognition, ECG, or even behavioural biometrics will provide complementary cues regarding identity (Becerra et al., 2025; Ur Rehman et al., 2025). Recent studies have demonstrated that multi-modal approaches (e.g. EEG + facial recognition) increase accuracy and liveness detection, and we envision systems where EEG is only one component of a multi-modal biometric suite, and can be intelligently fused at the feature or decision level.
- New methods: Finally, we need to consider utilizing transformer-based architectures or meta-learning for EEG and tackle the issue of data erosion and ethical issues in detail. Also, innovations in mobile EEG and dry electrode technology may also ease feasibility of data collection for our learning and enable us to work on a moderately larger scale.

5.4 Conclusion: -

In conclusion, EEG-based biometric methods have significant promise, but their actual application must consider the variability and scalability of EEG-based biometric testing. Future systems can begin to consider both high accuracy and reliability going forward by implementing deep models, richer features, and multi-modal data. Any trade-offs we have observed (accuracy vs. generalizability) can also help guide the development of methods, where it is far better to have robust, but perhaps not quite as high, accuracy across sessions than to be overly optimistic with accuracy in a single session. In the long term, the development of machine learning, and EEG technology, will ideally help mitigate these reliability limitations

so that EEG can be used, alone or in conjunction with other systems, to complement existing biometric systems.

REFERENCES

Akbarnia, Y., & Daliri, M. R. (2024). EEG-based identification system using deep neural networks with frequency features. Heliyon, 10(4), e25999.

Albaiati, A. E., Akbar, M. F., Hssayeni, M. D., Khalil, A., Ab Wahab, M. N., Weli, S. S., & Raheema, E. A. (2025). Deep Learning Approaches for EEG-Based Biometrics: A Systematic Review. *IEEE Access*.

Al-Janabi, R. A., Al-Qaysi, Z. T., & Suzani, M. S. (2024*).* Deep transfer learning model for EEG biometric decoding. Applied Data Science and Analysis, 2024, 4–16.

Apicella, A., Arpaia, P., D'Errico, G., Marocco, D., Mastrati, G., Moccaldi, N., & Prevete, R. (2024). Toward cross-subject and cross-session generalization in EEG-based emotion recognition: Systematic review, taxonomy, and methods. *Neurocomputing*, *604*, 128354.

Arnau-González, P., Arevalillo-Herráez, M., Katsigiannis, S., & Ramzan, N. (2018). On the influence of affect in EEG-based subject identification. *IEEE Transactions on Affective Computing*, *12*(2), 391-401.

Ashenaei, R., Beheshti, A. A., & Rezaii, T. Y. (2022). Stable EEG-based biometric system using functional connectivity based on time-frequency features with optimal channels. *Biomedical Signal Processing and Control*, *77*, 103790.

Balci, F. (2023). *DM-EEGID: EEG-based biometric authentication system using hybrid attention-based LSTM and MLP algorithm*. Traitement du Signal, 40(1), 65–79.

Becerra, A., Daza, R., Cobos, R., Morales, A., & Fierrez, J. (2025). M2LADS Demo: A System for Generating Multimodal Learning Analytics Dashboards. *arXiv preprint arXiv:2502.15363*.

Bidgoly, A. J., Bidgoly, H. J., & Arezoumand, Z. (2020). A survey on methods and challenges in EEG based authentication. *Computers & Security*, *93*, 101788.

Bidgoly, A. J., Bidgoly, H. J., & Arezoumand, Z. (2022). Towards a universal and privacy preserving EEG-based authentication system. *Scientific Reports*, *12*(1), 2531.

Chan, H. L., Kuo, P. C., Cheng, C. Y., & Chen, Y. S. (2018). Challenges and future perspectives on electroencephalogram-based biometrics in person recognition. *Frontiers in neuroinformatics*, *12*, 66.

Chen, J. X., Mao, Z. J., Yao, W. X., & Huang, Y. F. (2020). *EEG-based biometric identification with convolutional neural network*. Multimedia Tools and Applications, 79(15–16), 10655–10675.

Del Pup, G., Pasini, A., & Vassanelli, S. (2025). Brain states and decoding performance: How experimental design biases the evaluation of EEG-based machine learning classifiers. *Frontiers in Neuroscience, 19*, 1511308.

Di, Y., An, X., Zhong, W., Liu, S., & Ming, D. (2021). The time-robustness analysis of individual identification based on resting-state EEG. *Frontiers in Human Neuroscience*, *15*, 672946.

Didaci, L., Pani, S. M., Frongia, C., & Fraschini, M. (2024). How Time Window Influences Biometrics Performance: An EEG-Based Fingerprint Connectivity Study. *Signals*, *5*(3), 597-604.

Fan, Y., Shi, X., & Li, Q. (2021). *CNN-based personal identification system using resting state electroencephalography*. Computational Intelligence and Neuroscience, 2021, Article ID 1160454.

Ghasemi, E., Ebrahimi, M., & Ebrahimie, E. (2022). Machine learning models effectively distinguish attention-deficit/hyperactivity disorder using event-related potentials. *Cognitive Neurodynamics*, *16*(6), 1335-1349.

Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. e215–e220. RRID:SCR_007345.

Gong, Y., Wang, M., Zhang, Y., Zhang, W., & Pang, S. (2024, December). A Unified Deep Learning-Based EEG Biometric Authentication System for Cross-Session Scenarios. In *International Conference on Advanced Data Mining and Applications* (pp. 48-62). Singapore: Springer Nature Singapore.

Kaewwit, C., Lursinsap, C., & Sophatsathit, P. (2017). High accuracy EEG biometrics identification using ICA and AR model. Journal of ICT, 16(2), 354–373.

Kang, J.-H., Jo, Y. C., & Kim, S.-P. (2018). Electroencephalographic feature evaluation for improving personal authentication performance. Neurocomputing, 287, 93–101.

Khan, H. A., Ul Ain, R., Kamboh, A. M., Butt, H. T., Shafait, S., Alamgir, W., ... & Shafait, F. (2022). The NMT scalp EEG dataset: An open-source annotated dataset of healthy and pathological EEG recordings for predictive modeling. *Frontiers in neuroscience*, *15*, 755817.

Kim, D., & Kim, K. (2019). Resting state EEG-based biometric system using concatenation of quadrantal functional networks. IEEE Access, 7, 65745–65756.

Kumar, M. G., Narayanan, S., Sur, M., & Murthy, H. A. (2021). Evidence of task-independent person-specific signatures in EEG using subspace techniques. *IEEE Transactions on Information Forensics and Security*, *16*, 2856-2871.

Lopes, F., Leal, A., Medeiros, J., Pinto, M. F., Dourado, A., Dümpelmann, M., & Teixeira, C. (2022). EPIC: Annotated epileptic EEG independent components for artifact reduction. *Scientific data*, *9*(1), 512.

Lotte, F., Bougrain, L., Cichocki, A., Clerc, M., Congedo, M., Rakotomamonjy, A., & Yger, F. (2018). A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update. *Journal of neural engineering*, *15*(3), 031005.

Lu, Y., Wang, W., Lian, B., & He, C. (2024). Feature extraction and classification of motor imagery EEG signals in motor imagery for sustainable brain–computer interfaces. *Sustainability*, *16*(15), 6627.

Lyu, S., & Cheung, R. C. (2023). Efficient Multiple Channels EEG Signal Classification Based on Hierarchical Extreme Learning Machine. *Sensors*, *23*(21), 8976.

Ma, J., Yang, B., Qiu, W., Li, Y., Gao, S., & Xia, X. (2022). A large EEG dataset for studying cross-session variability in motor imagery brain-computer interface. *Scientific Data*, *9*(1), 531.

Maiorana, E. (2021). Learning deep features for task-independent EEG-based biometric verification. *Pattern Recognition Letters*, *143*, 122-129.

Maiorana, E., & Campisi, P. (2018). Longitudinal evaluation of EEG-based biometric recognition. IEEE Transactions on Information Forensics and Security, 13(5), 1123–1138.

Maiorana, E., La Rocca, D., & Campisi, P. (2015). On the permanence of EEG signals for biometric recognition. *IEEE Transactions on Information Forensics and Security*, *11*(1), 163-175.

Mao, Z., Yao, W., & Huang, Y. (2017). *EEG-based biometric identification with deep learning*. In Proceedings of the 8th International IEEE/EMBS Conference on Neural Engineering (NER) (pp. 609–612). IEEE.

Monsy, J. C., & Vinod, A. P. (2020). EEG-based biometric identification using frequency-weighted power feature. IET Biometrics, 9(6), 251–258.

Mota, M. R., Silva, P. H., Luz, E. J., Moreira, G. J., Schons, T., Moraes, L. A., & Menotti, D. (2021). A deep descriptor for cross-tasking EEG-based recognition. *PeerJ Computer Science*, *7*, e549.

Nakamura, T., Goverdovsky, V., & Mandic, D. P. (2018). In-ear EEG biometrics for feasible and readily collectable real-world person authentication. IEEE Transactions on Information Forensics and Security, 13(3), 648–661.

Ozdenizci, O., Wang, Y., Koike-Akino, T., & Erdogmus, D. (2019). *Adversarial deep learning in EEG biometrics*. IEEE Signal Processing Letters, 26(5), 710–714.

Plucińska, R., Jędrzejewski, K., Malinowska, U., & Rogala, J. (2023). Leveraging multiple distinct EEG training sessions for improvement of spectral-based biometric verification results. *Sensors*, *23*(4), 2057.

Rahman, A., Chowdhury, M. E., Khandakar, A., Tahir, A. M., Ibtehaz, N., Hossain, M. S., ... & Kadir, M. A. (2022). Robust biometric system using session invariant multimodal EEG and keystroke dynamics by the ensemble of self-ONNs. *Computers in Biology and Medicine*, *142*, 105238.

Rehman, T. U., Alruwaili, M., Siddiqi, M. H., Alhwaiti, Y., Anwar, S., Halim, Z., & Alam, M. (2025). Advancing EEG-based biometric identification through multi-modal data fusion and deep learning techniques. *Complex & Intelligent Systems*, *11*(9), 398.

Ruiz-Blondet, M. V., Jin, Z., & Laszlo, S. (2016). CEREBRE: A novel method for very high accuracy event-related potential biometric identification. *IEEE Transactions on Information Forensics and Security*, *11*(7), 1618-1629.

Saeidi, M., Karwowski, W., Farahani, F. V., Fiok, K., Taiar, R., Hancock, P. A., & Al-Juaid, A. (2021). Neural decoding of EEG signals with machine learning: A systematic review. *Sensors, 21*(24), 8237.

Sun, Y., Lo, F. P. W., & Lo, B. (2019). EEG-based user identification system using 1D-convolutional long short-term memory neural networks. *Expert Systems with Applications*, *125*, 259-267.

Thomas, K. P., & Vinod, A. P. (2018). EEG-based biometric authentication using gamma band power during rest state. Circuits, Systems, and Signal Processing, 37(1), 277–289.

Waili, T., Johar, M. G. M., Sidek, K. A., Mohd Nor, N. S. H., Yaacob, H., & Othman, M. (2019). EEG-based biometric identification using correlation and MLPNN models. International Journal of Online and Biomedical Engineering, 15(10), 19–30.

White, J., & Power, S. D. (2023). k-fold cross-validation can significantly over-estimate true classification accuracy in common EEG-based passive BCI experimental designs: an empirical investigation. *Sensors*, *23*(13), 6077.

Wilaiprasitporn, T., Ditthaporn, A., Matchaparn, K., Tongbuasirilai, T., Banluesombatkul, N., & Chuangsuwanich, E. (2020). Affective EEG-based person identification using the deep learning approach. *IEEE Transactions on Cognitive and Developmental Systems, 12*(3), 486–496.

Yang, J., Gao, S., & Shen, T. (2022). A two-branch CNN fusing temporal and frequency features for motor imagery EEG decoding. *Entropy*, *24*(3), 376.

Yang, S., Deravi, F., & Hoque, S. (2018). Task sensitivity in EEG biometric recognition. *Pattern Analysis and Applications*, *21*(1), 105-117..

Yang, Y., Hou, Z., Wang, Y., Ma, H., Sun, P., Ma, Z., ... & Li, X. (2022). HCRNet: high-throughput circRNA-binding event identification from CLIP-seq data using deep temporal convolutional network. *Briefings in Bioinformatics*, *23*(2), bbac027.

Zhang, R., Zeng, Y., Tong, L., Shu, J., Lu, R., Li, Z., ... & Yan, B. (2022). EEG identity authentication in multi-domain features: a multi-scale 3D-CNN approach. *Frontiers in Neurorobotics*, *16*, 901765.

Zhang, Y., Pezeshki, M., Brakel, P., Zhang, S., Bengio, C. L. Y., & Courville, A. (2017). Towards end-to-end speech recognition with deep convolutional neural networks. *arXiv preprint arXiv:1701.02720*.

APPENDICES

Appendix 1:-

```
# === Imports & Config ===

import glob

import os

import re

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt


from scipy.signal import welch

from sklearn.model_selection import StratifiedKFold, GridSearchCV, train_test_split

from sklearn.preprocessing import StandardScaler, label_binarize

from sklearn.pipeline import Pipeline

from sklearn.metrics import (accuracy_score, confusion_matrix, classification_report,

                             roc_curve, auc)

from sklearn.ensemble import RandomForestClassifier

from sklearn.svm import SVC

import joblib



# ---- Set where your EDF subject folders live ----

# Example structure:

#   DATA_ROOT/

#     S001/S001R01.edf, S001R02.edf, ...

#     S002/S002R01.edf, ...

DATA_ROOT  = r"C:\Users\Admin\Desktop\ALVIN\eeg-motor-movementimagery-dataset-1.0.0"  # <-- change to your EDF root

OUTPUT_DIR = r"C:\Users\Admin\Desktop\ALVIN\outputs_final_loso"  # separate from your CSV project
```

```
# ---- Analysis knobs ----
WINDOW_SEC = 5.0        # per-window length (seconds)
OVERLAP    = 0        # 50% overlap
DURATION_S = None       # None = use full file; or set (e.g., 60.0) to trim
RANDOM_SEED = 42
np.random.seed(RANDOM_SEED)

# 19-channel target order (legacy 10/20 names you used before)
CHANNELS_19 = ['Fp1','Fp2','F3','F4','F7','F8','T3','T4','C3','C4',
            'T5','T6','P3','P4','O1','O2','Fz','Cz','Pz']

# Map modern MNE names to your legacy labels (so features align with your old pipeline)
CHANNEL_ALIAS = {'T7':'T3', 'T8':'T4', 'P7':'T5', 'P8':'T6'}

# EEG bands (same as before, stopping at 45 Hz to avoid 50 Hz powerline)
BANDS = {
   "Delta": (0.5, 4.0),
   "Theta": (4.0, 7.0),
   "Alpha": (8.0, 13.0),
   "Beta":  (13.0, 30.0),
   "Gamma": (30.0, 45.0),
}

# ---- Organized output folders (mirrors your preferred layout) ----
DIRS = {
   "corrected_csv": os.path.join(OUTPUT_DIR, "corrected_csv"),  # mostly unused for EDF, kept for
parity
   "features":     os.path.join(OUTPUT_DIR, "features"),
   "models":       os.path.join(OUTPUT_DIR, "models"),
   "plots":        os.path.join(OUTPUT_DIR, "plots"),
   "similarity":   os.path.join(OUTPUT_DIR, "similarity"),
```

```
    # features subfolders
    "feat_sessions": os.path.join(OUTPUT_DIR, "features", "sessions"),  # per-session window features
    "feat_master":   os.path.join(OUTPUT_DIR, "features", "master"),    # concatenated master tables
    "feat_splits":   os.path.join(OUTPUT_DIR, "features", "splits"),    # train/val/test CSVs

    # plots subfolders
    "plots_psd":     os.path.join(OUTPUT_DIR, "plots", "psd"),
    "plots_band":    os.path.join(OUTPUT_DIR, "plots", "band"),
    "plots_cm":      os.path.join(OUTPUT_DIR, "plots", "cm"),
    "plots_roc":     os.path.join(OUTPUT_DIR, "plots", "roc"),

    # models subfolders
    "models_ckpt":   os.path.join(OUTPUT_DIR, "models", "checkpoints"),
    "models_final":  os.path.join(OUTPUT_DIR, "models", "final"),
}
for d in DIRS.values():
    os.makedirs(d, exist_ok=True)

print("Output tree:")
for k,v in DIRS.items():
    print(f" - {k}: {v}")
edf_paths = sorted(glob.glob(os.path.join(DATA_ROOT, "S???", "S???R??.edf")))
if not edf_paths:
    raise FileNotFoundError(f"No EDF files found under {DATA_ROOT}. Check the path and pattern.")

rows = []
pat = re.compile(r"(S\d{3})[\\/](S\d{3})R(\d{2})\.edf$", re.IGNORECASE)

for p in edf_paths:
    m = pat.search(p.replace("/", os.sep))
```

```
    if not m:
        # fallback: try filename only
        base = os.path.basename(p)
        m2 = re.match(r"(S\d{3})R(\d{2})\.edf$", base, re.IGNORECASE)
        if not m2:
            continue
        subj = base[:4]   # e.g., S109
        sess = m2.group(1)  # R##
    else:
        subj = m.group(2)     # S109 (second capture)
        sess = f"R{m.group(3)}"  # R01, R02, ...

    event_path = os.path.splitext(p)[0] + ".edf.event"
    rows.append({"Subject": subj.upper(), "Session": sess.upper(),
                 "edf_path": p, "event_path": event_path if os.path.exists(event_path) else None})

index_df = pd.DataFrame(rows).sort_values(["Subject","Session"]).reset_index(drop=True)

# Save the index for traceability
index_csv = os.path.join(DIRS["feat_master"], "session_index.csv")
index_df.to_csv(index_csv, index=False)

# Quick summaries
counts = index_df.groupby("Subject")["Session"].count().rename("n_sessions")
summary = counts.reset_index().sort_values("n_sessions", ascending=False)

print(f"Found {len(index_df)} sessions across {summary.shape[0]} subjects.")
display(index_df.head(10))
display(summary.head(10))
# === Cell 3A: Inspect raw channel names & normalization ===
import re
```

```
import mne

# Pick first EDF file
row = index_df.iloc[0]
edf_path = row["edf_path"]
raw = mne.io.read_raw_edf(edf_path, preload=False, verbose="ERROR")
orig = raw.ch_names

def rough_norm(name: str) -> str:
    n = name.upper().strip()
    n = n.replace("EEG ", "").replace("EEG_", "").replace("EEG", "")
    n = re.sub(r"[ \-_()]", "", n)
    for suf in ("REF","LE","RE","A1","M1","A2","M2","AVG","AVERAGE","AVGREF"):
        if n.endswith(suf):
            n = n[: -len(suf)]
    n = {"T7":"T3","T8":"T4","P7":"T5","P8":"T6"}.get(n, n)
    pretty = {
        "FP1":"Fp1","FP2":"Fp2","F3":"F3","F4":"F4","F7":"F7","F8":"F8",
        "T3":"T3","T4":"T4","C3":"C3","C4":"C4","T5":"T5","T6":"T6",
        "P3":"P3","P4":"P4","O1":"O1","O2":"O2","FZ":"Fz","CZ":"Cz","PZ":"Pz"
    }
    return pretty.get(n, n)

norm = [rough_norm(c) for c in orig]
df_names = pd.DataFrame({"original": orig, "normalized": norm})
print("First 40 channels:")
display(df_names.head(40))

targets = set(CHANNELS_19)
present = sorted(list(set(norm).intersection(targets)))
missing = sorted(list(targets.difference(present)))
```

```
print("Present targets:", present)
print("Missing targets:", missing)
# === Cell 3B: Strict EDF loader & smoke test ===
import re
import numpy as np

def _normalize_eeg_name_strict(name: str) -> str:
    n = name.upper().strip()
    n = n.replace("EEG ", "").replace("EEG_", "").replace("EEG", "")
    n = re.sub(r"[^A-Z0-9]", "", n)
    for suf in ("REF","LE","RE","A1","M1","A2","M2","AVG","AVERAGE","AVGREF"):
        if n.endswith(suf):
            n = n[: -len(suf)]
    n = {"T7":"T3","T8":"T4","P7":"T5","P8":"T6"}.get(n, n)
    pretty = {
        "FP1":"Fp1","FP2":"Fp2","F3":"F3","F4":"F4","F7":"F7","F8":"F8",
        "T3":"T3","T4":"T4","C3":"C3","C4":"C4","T5":"T5","T6":"T6",
        "P3":"P3","P4":"P4","O1":"O1","O2":"O2","FZ":"Fz","CZ":"Cz","PZ":"Pz"
    }
    return pretty.get(n, n)

def load_edf_19(edf_path, duration_s=DURATION_S, target_channels=CHANNELS_19):
    raw = mne.io.read_raw_edf(edf_path, preload=True, verbose="ERROR")
    fs = float(raw.info["sfreq"])

    # Rename channels
    rename = {}
    for ch in raw.ch_names:
        nn = _normalize_eeg_name_strict(ch)
        if nn != ch:
            rename[ch] = nn
```

```
    if rename:
        raw.rename_channels(rename)

    # Crop to fixed duration
    if duration_s is not None:
        n_samp = int(duration_s * fs)
        raw.crop(tmax=(n_samp - 1) / fs)

    # Select available target channels
    present = [ch for ch in target_channels if ch in raw.ch_names]
    if not present:
        print("Normalized names (first 40):", raw.ch_names[:40])
        raise ValueError("No appropriate channels found after strict normalization.")

    data = raw.get_data(picks=present).T
    df = pd.DataFrame(index=np.arange(data.shape[0]), columns=target_channels, dtype=float)
    df[:] = np.nan
    present_df = pd.DataFrame(data, columns=present)
    for ch in present:
        df[ch] = present_df[ch].values

    print(f"Loaded {os.path.basename(edf_path)} | fs={fs:.2f} Hz | samples={df.shape[0]} | present
{len(present)}/19")
    miss = [ch for ch in target_channels if ch not in present]
    if miss:
        print("Missing targets:", miss)
    return df, fs

# --- Smoke test on first file ---
row = index_df.iloc[0]
df_test, fs_test = load_edf_19(row["edf_path"])
```

```
print(row["Subject"], row["Session"], "fs=", fs_test, "shape=", df_test.shape)
display(df_test.head())
# === Cell 4: Windowing + Welch bandpowers (abs/rel) for ALL sessions ===
import os
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from scipy.signal import welch

# Assumes these already exist from earlier cells:
# - index_df (Cell 2)
# - load_edf_19 (Cell 3B)
# - DIRS, WINDOW_SEC, OVERLAP, DURATION_S, CHANNELS_19, BANDS (Cell 1 / your setup)

# ---- Helpers ----
def segment_windows(n_samples, fs, win_sec=WINDOW_SEC, overlap=OVERLAP):
    win_len = int(win_sec * fs)
    hop = max(1, int(win_len * (1 - overlap)))
    return [(s, s + win_len) for s in range(0, n_samples - win_len + 1, hop)]

def welch_psd(x, fs, nperseg, noverlap):
    return welch(x, fs=fs, nperseg=int(nperseg), noverlap=int(noverlap), detrend='constant')

def bandpower_integrate(f, Pxx, band):
    fmin, fmax = band
    idx = (f >= fmin) & (f < fmax)
    if not np.any(idx):
        return 0.0
    return float(np.trapz(Pxx[idx], f[idx]))

def compute_window_features_df(df_win, fs, channels=CHANNELS_19, bands=BANDS):
```

```
    """
    Given a windowed DataFrame (time x channels), compute abs & rel bandpower per channel.
    Returns: dict of features.
    """
    nperseg = max(4, int(2 * fs))       # conservative for 5 s windows
    noverlap = int(0.5 * nperseg)
    feats = {}
    for ch in channels:
        x = pd.to_numeric(df_win[ch], errors='coerce').fillna(0.0).values
        f, Pxx = welch_psd(x, fs, nperseg, noverlap)
        total = float(np.trapz(Pxx, f))
        for b, rng in bands.items():
            abs_p = bandpower_integrate(f, Pxx, rng)
            rel_p = abs_p / total if total > 0 else np.nan
            feats[f"{ch}_{b}"]     = abs_p
            feats[f"{ch}_{b}_Rel"] = rel_p
    return feats


def plot_psd_overlay(df_full, fs, channels, out_png, title):
    nperseg = int(4 * fs)
    noverlap = int(0.5 * nperseg)
    plt.figure(figsize=(12, 8))
    for ch in channels:
        x = pd.to_numeric(df_full[ch], errors='coerce').fillna(0.0).values
        f, Pxx = welch_psd(x, fs, nperseg, noverlap)
        plt.plot(f, 10*np.log10(Pxx), label=ch)
    plt.xlim(0, 50)
    plt.xlabel("Frequency (Hz)"); plt.ylabel("PSD (dB/Hz)")
    plt.title(title)
    plt.legend(bbox_to_anchor=(1.05, 1), loc="upper left", fontsize=8)
    plt.tight_layout()
```

```
    plt.savefig(out_png, dpi=150); plt.close()

def plot_band_bars(wide_feat_row, channels, bands, out_abs, out_rel, title):
    xs = np.arange(len(channels)); width = 0.15; band_list = list(bands.keys())
    # Absolute
    plt.figure(figsize=(14, 6))
    for i, b in enumerate(band_list):
        vals = [wide_feat_row.get(f"{ch}_{b}", np.nan) for ch in channels]
        plt.bar(xs + i*width, vals, width, label=b)
    plt.xticks(xs + width*2, channels, rotation=45)
    plt.ylabel("Absolute Band Power")
    plt.title(title + " (Absolute)")
    plt.legend()
    plt.tight_layout()
    plt.savefig(out_abs, dpi=150); plt.close()
    # Relative
    plt.figure(figsize=(14, 6))
    for i, b in enumerate(band_list):
        vals = [wide_feat_row.get(f"{ch}_{b}_Rel", np.nan) for ch in channels]
        plt.bar(xs + i*width, vals, width, label=b)
    plt.xticks(xs + width*2, channels, rotation=45)
    plt.ylabel("Relative Band Power")
    plt.title(title + " (Relative)")
    plt.legend()
    plt.tight_layout()
    plt.savefig(out_rel, dpi=150); plt.close()

# Ensure required output subfolders exist (based on your DIRS)
for key in ["feat_sessions", "feat_master", "plots_psd", "plots_band"]:
    os.makedirs(DIRS[key], exist_ok=True)
```

```
# ---- Process all sessions in index_df ----
all_rows = []
per_session_counts = []

for i, row in index_df.iterrows():
    subj, sess, edf_path = row["Subject"], row["Session"], row["edf_path"]

    # Load 19-channel frame (uses your strict loader from Cell 3B)
    df_full, fs = load_edf_19(edf_path, duration_s=DURATION_S, target_channels=CHANNELS_19)

    # Windowing
    wins = segment_windows(len(df_full), fs, WINDOW_SEC, OVERLAP)
    sess_rows = []
    for (s, e) in wins:
        dfw = df_full.iloc[s:e]
        feats = compute_window_features_df(dfw, fs, channels=CHANNELS_19, bands=BANDS)
        feats["Subject"] = subj
        feats["Session"] = sess
        feats["StartSample"] = int(s)
        feats["Fs"] = float(fs)
        sess_rows.append(feats)
        all_rows.append(feats)

    # Save per-session features
    sess_df = pd.DataFrame(sess_rows)
    out_csv = os.path.join(DIRS["feat_sessions"], f"{subj}_{sess}_winfeat.csv")
    sess_df.to_csv(out_csv, index=False)

    # Wide summary over full recording for plots
    wide_feats = compute_window_features_df(df_full, fs, channels=CHANNELS_19, bands=BANDS)
    psd_png   = os.path.join(DIRS["plots_psd"],  f"psd_{subj}_{sess}.png")
```

```
    bandA_png = os.path.join(DIRS["plots_band"], f"band_abs_{subj}_{sess}.png")
    bandR_png = os.path.join(DIRS["plots_band"], f"band_rel_{subj}_{sess}.png")
    plot_psd_overlay(df_full, fs, CHANNELS_19, psd_png,  f"EEG PSD — {subj} {sess}")
    plot_band_bars(wide_feats, CHANNELS_19, BANDS, bandA_png, bandR_png, f"Band Powers — {subj} {sess}")

    per_session_counts.append({"Subject": subj, "Session": sess,
                               "n_windows": len(sess_rows), "fs": fs})

# Save master table and counts
master_df = pd.DataFrame(all_rows)
master_csv = os.path.join(DIRS["feat_master"], f"eeg_master_windows_{int(WINDOW_SEC)}s.csv")
master_df.to_csv(master_csv, index=False)

counts_df = pd.DataFrame(per_session_counts).sort_values(["Subject","Session"])
counts_csv = os.path.join(DIRS["feat_master"], "session_window_counts.csv")
counts_df.to_csv(counts_csv, index=False)

print(f"Saved per-session CSVs to: {DIRS['feat_sessions']}")
print(f"Saved master feature table: {master_csv}  (shape={master_df.shape})")
print(f"Saved session window counts: {counts_csv}")
display(counts_df.head())
# === Cell 5: Quick preview & sanity checks ===
print("Master features shape:", master_df.shape)
print("Columns (first 10):", master_df.columns[:10].tolist())
print("Subjects:", sorted(master_df['Subject'].unique().tolist())[:10], '...')

# Windows per subject
w_per_subj = master_df.groupby("Subject")["StartSample"].count().rename("n_windows").reset_index()
display(w_per_subj.sort_values("n_windows", ascending=False).head(10))
```

```
# Windows per (Subject, Session)
w_per_sess =
master_df.groupby(["Subject","Session"])["StartSample"].count().rename("n_windows").reset_index
()
display(w_per_sess.head(10))
# === Cell 6: Build SESSION-DISJOINT 60/20/20 splits per subject ===
import os
import pandas as pd

# Assumes you already ran Cell 4 and have the master feature CSV
FEAT_MASTER = os.path.join(OUTPUT_DIR, "features", "master")
master_csv  = os.path.join(FEAT_MASTER, "eeg_master_windows_5s.csv")

# Read just Subject/Session to build a split plan
df_index = pd.read_csv(master_csv, usecols=["Subject", "Session"]).drop_duplicates()

def sort_sessions(sess_list):
    """
    Sort session labels like ['R01','R02','R10'] correctly.
    If your labels are different, adjust this parser.
    """
    def key(s):
        # Expect 'R##' → numeric; fallback to string
        try:
            return int(str(s).strip().lstrip("R"))
        except:
            return s
    return sorted(sess_list, key=key)

def split_60_20_20(sess_list):
    sess = sort_sessions(sess_list)
    n = len(sess)
```

```
    if n == 1:
        return sess, [], []              # all train if only 1
    if n == 2:
        return [sess[0]], [], [sess[1]]     # train, test
    # general case
    n_train = max(1, int(round(0.6 * n)))
    n_val   = max(1, int(round(0.2 * n)))
    # ensure at least 1 test
    if n_train + n_val >= n:
        n_val = max(1, n_val - 1)
    train = sess[:n_train]
    val   = sess[n_train:n_train+n_val]
    test  = sess[n_train+n_val:]
    if len(test) == 0:
        test = [sess[-1]]
        if val and val[-1] == test[0]:
            val = val[:-1]
    return train, val, test

plan_rows = []
for subj, g in df_index.groupby("Subject"):
    sess_list = g["Session"].unique().tolist()
    tr, va, te = split_60_20_20(sess_list)
    plan_rows.append({
        "Subject": subj,
        "Train": ",".join(tr),
        "Val":   ",".join(va),
        "Test":  ",".join(te),
        "n_sessions": len(sess_list)
    })
```

```
split_plan_df = pd.DataFrame(plan_rows).sort_values("Subject")

# Save plan
SPLIT_DIR = os.path.join(OUTPUT_DIR, "features", "splits")
os.makedirs(SPLIT_DIR, exist_ok=True)
split_csv = os.path.join(SPLIT_DIR, "session_split_plan_60_20_20.csv")
split_plan_df.to_csv(split_csv, index=False)

print("Saved session-wise split plan →", split_csv)
display(split_plan_df.head(12))
# === Cell 7 (LOSO): Train on TRAIN sessions, tune with LOSO on TRAIN, pick by VAL, report on TEST ===
import os
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.model_selection import LeaveOneGroupOut  # <-- new
# (Assumes all other sklearn imports were done in Cell 1)

# Paths
FEAT_MASTER = os.path.join(OUTPUT_DIR, "features", "master")
master_csv  = os.path.join(FEAT_MASTER, "eeg_master_windows_5s.csv")
SPLIT_DIR   = os.path.join(OUTPUT_DIR, "features", "splits")
split_csv   = os.path.join(SPLIT_DIR, "session_split_plan_60_20_20.csv")

# Load master features & split plan
dfm   = pd.read_csv(master_csv)
plan  = pd.read_csv(split_csv)

# Build lookup: subject → set(s) of sessions for each split
def plan_to_sets(plan_df):
```

```
    def to_set(x):
        s = str(x).strip()
        return set([t for t in s.split(",") if t]) if s and s.lower() != "nan" else set()
    train_map, val_map, test_map = {}, {}, {}
    for _, r in plan_df.iterrows():
        subj = r["Subject"]
        train_map[subj] = to_set(r["Train"])
        val_map[subj]   = to_set(r["Val"])
        test_map[subj]  = to_set(r["Test"])
    return train_map, val_map, test_map

train_map, val_map, test_map = plan_to_sets(plan)

def mask_for(split_map):
    return dfm.apply(lambda r: r["Session"] in split_map.get(r["Subject"], set()), axis=1)

m_train = mask_for(train_map)
m_val   = mask_for(val_map)
m_test  = mask_for(test_map)

df_train = dfm[m_train].copy()
df_val   = dfm[m_val].copy()
df_test  = dfm[m_test].copy()

print("Window counts (session-disjoint):")
print("  Train:", len(df_train), " Val:", len(df_val), " Test:", len(df_test))

# Features and labels
feat_cols = [c for c in dfm.columns if c not in ["Subject","Session","StartSample","Fs"]]
X_train, y_train = df_train[feat_cols].values, df_train["Subject"].values
X_val,   y_val   = df_val[feat_cols].values,   df_val["Subject"].values
```

```
X_test,  y_test  = df_test[feat_cols].values,  df_test["Subject"].values
classes = sorted(dfm["Subject"].unique().tolist())

# --- NEW: groups for LOSO over TRAIN sessions (one group per Subject+Session) ---
groups_train = (df_train["Subject"] + "_" + df_train["Session"]).values

# Models & grids (reuse imports from Cell 1)
rf_pipe = Pipeline([
    ("scaler", StandardScaler(with_mean=True, with_std=True)),
    ("clf", RandomForestClassifier(random_state=RANDOM_SEED))
])
rf_grid = {
    "clf__n_estimators": [300, 500],
    "clf__max_depth": [None, 12, 18],
    "clf__min_samples_leaf": [1, 2],
}

svm_pipe = Pipeline([
    ("scaler", StandardScaler(with_mean=True, with_std=True)),
    ("clf", SVC(kernel="rbf", probability=True, random_state=RANDOM_SEED))
])
svm_grid = {
    "clf__C": [1, 5, 10],
    "clf__gamma": ["scale", 0.01, 0.001],
}

def train_and_select(name, pipe, grid, Xtr, ytr, Xva, yva, groups):
    """
    Hyperparameter tuning with Leave-One-Session-Out CV on TRAIN sessions.
    Each fold leaves out one entire TRAIN session (group) for validation.
    """
```

```
    logo = LeaveOneGroupOut()
    gs = GridSearchCV(
        pipe, grid,
        cv=logo.split(Xtr, ytr, groups),
        n_jobs=-1, scoring="accuracy", verbose=0
    )
    gs.fit(Xtr, ytr)
    best    = gs.best_estimator_
    cv_acc  = gs.best_score_            # LOSO-CV accuracy over TRAIN sessions
    val_acc = accuracy_score(yva, best.predict(Xva))  # held-out VAL sessions
    print(f"[{name}] LOSO-CV acc={cv_acc:.4f} | VAL acc={val_acc:.4f} | best={gs.best_params_}")
    return best, cv_acc, val_acc, gs.best_params_

# Train both models, pick by VAL accuracy
models = []
for name, base, grid in [
    ("RandomForest", rf_pipe, rf_grid),
    ("SVM_RBF",     svm_pipe, svm_grid),
]:
    best, cv_acc, val_acc, params = train_and_select(
        name, base, grid, X_train, y_train, X_val, y_val, groups_train
    )
    models.append((name, best, cv_acc, val_acc, params))

models.sort(key=lambda t: t[3], reverse=True)
sel_name, sel_model, sel_cv, sel_val, sel_params = models[0]
print(f"\nSelected model: {sel_name} | VAL acc={sel_val:.4f} | LOSO-CV={sel_cv:.4f} |
params={sel_params}")

# Final TEST evaluation (session-disjoint)
yt_pred  = sel_model.predict(X_test)
```

```
test_acc = accuracy_score(y_test, yt_pred)
cm       = confusion_matrix(y_test, yt_pred, labels=classes)

# Output dirs
PLOTS_CM   = os.path.join(OUTPUT_DIR, "plots", "cm");    os.makedirs(PLOTS_CM, exist_ok=True)
PLOTS_ROC  = os.path.join(OUTPUT_DIR, "plots", "roc");   os.makedirs(PLOTS_ROC, exist_ok=True)
MODELS_OUT = os.path.join(OUTPUT_DIR, "models", "final"); os.makedirs(MODELS_OUT,
exist_ok=True)

# Confusion matrix (TEST)
cm_png = os.path.join(PLOTS_CM, f"cm_test_{sel_name}_LOSO_sessiondisjoint.png")
plt.figure(figsize=(10, 9))
plt.imshow(cm, cmap="Blues")
plt.title(f"Confusion Matrix (TEST, LOSO) — {sel_name} | acc={test_acc:.4f}")
plt.xticks(np.arange(len(classes)), classes, rotation=90, fontsize=6)
plt.yticks(np.arange(len(classes)), classes, fontsize=6)
for i in range(cm.shape[0]):
    for j in range(cm.shape[1]):
        plt.text(j, i, str(cm[i, j]), ha='center', va='center', fontsize=5)
plt.colorbar(); plt.tight_layout(); plt.savefig(cm_png, dpi=150); plt.close()

# One-vs-rest ROC on TEST
roc_png = os.path.join(PLOTS_ROC, f"roc_test_{sel_name}_LOSO_sessiondisjoint.png")
y_prob  = sel_model.predict_proba(X_test)
Ybin    = label_binarize(y_test, classes=classes)
plt.figure(figsize=(8, 7))
macro_aucs = []
for i, cname in enumerate(classes):
    fpr, tpr, thr = roc_curve(Ybin[:, i], y_prob[:, i])
    roc_auc = auc(fpr, tpr)
    macro_aucs.append(roc_auc)
```

```
plt.plot([0,1],[0,1],'--', lw=1)
plt.xlabel("False Positive Rate"); plt.ylabel("True Positive Rate")
plt.title(f"OvR ROC (TEST, LOSO) — {sel_name} | Macro AUC={np.mean(macro_aucs):.3f}")
plt.tight_layout(); plt.savefig(roc_png, dpi=150); plt.close()

# Report & save model
rep     = classification_report(y_test, yt_pred, labels=classes, output_dict=True)
rep_df  = pd.DataFrame(rep).transpose()
rep_csv = os.path.join(MODELS_OUT, f"class_report_test_{sel_name}_LOSO_sessiondisjoint.csv")
rep_df.to_csv(rep_csv)

model_path = os.path.join(MODELS_OUT,
f"chosen_{sel_name}_trainval_LOSO_sessiondisjoint.joblib")
joblib.dump(sel_model, model_path)

print(f"\nTEST acc (LOSO, session-disjoint) = {test_acc:.4f}")
print("Saved:",
    "\n - Confusion matrix:", cm_png,
    "\n - ROC (OvR):", roc_png,
    "\n - Test classification report:", rep_csv,
    "\n - Final model:", model_path)
```

Appendix 2:-

```
# === Cell 1: Paths & Output Folders ===
import os

# Point these to the folders that CONTAIN your CSVs
MASTER_DIR = r"C:\Users\Admin\Desktop\ALVIN\outputs_final_loso\features\master"
SPLIT_DIR  = r"C:\Users\Admin\Desktop\ALVIN\outputs_final_loso\features\splits"

# Your requested output directory
OUTPUT_DIR = r"C:\Users\Admin\Desktop\ALVIN\output_spilt"
```

```
# Subfolders (auto-created)
PLOTS_CM    = os.path.join(OUTPUT_DIR, "plots", "cm")
PLOTS_ROC   = os.path.join(OUTPUT_DIR, "plots", "roc")
MODELS_OUT  = os.path.join(OUTPUT_DIR, "models")
REPORTS_OUT = os.path.join(OUTPUT_DIR, "reports")
SESSION_OUT = os.path.join(OUTPUT_DIR, "session_eval")
for d in [OUTPUT_DIR, PLOTS_CM, PLOTS_ROC, MODELS_OUT, REPORTS_OUT, SESSION_OUT]:
    os.makedirs(d, exist_ok=True)

print("OUTPUT_DIR:", OUTPUT_DIR)
# === Cell 2: Imports & Config ===
import glob, gc
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import os

from sklearn.model_selection import GridSearchCV
from sklearn.preprocessing import StandardScaler, label_binarize
from sklearn.pipeline import Pipeline
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report, roc_curve, auc
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import SVC
import joblib

# Keep runs stable on CPU-only Windows
os.environ["OMP_NUM_THREADS"] = "1"
os.environ["OPENBLAS_NUM_THREADS"] = "1"
os.environ["MKL_NUM_THREADS"] = "1"
os.environ["VECLIB_MAXIMUM_THREADS"] = "1"
```

```
os.environ["NUMEXPR_NUM_THREADS"] = "1"


RANDOM_SEED = 42
np.random.seed(RANDOM_SEED)
# === Cell 3: Resolve MASTER_CSV and SPLIT_CSV from folders ===
def pick_file(base_dir, prefer_names, fallback="*.csv"):
    if not os.path.isdir(base_dir):
        raise FileNotFoundError(f"Dir not found: {base_dir}")
    for name in prefer_names:
        hits = glob.glob(os.path.join(base_dir, name))
        if hits:
            return hits[0]
    hits = glob.glob(os.path.join(base_dir, fallback))
    if not hits:
        raise FileNotFoundError(f"No CSV found in {base_dir}")
    return hits[0]

MASTER_CSV = pick_file(MASTER_DIR,
                       ["eeg_master_windows_5s.csv", "*master*.csv"])
SPLIT_CSV  = pick_file(SPLIT_DIR,
                       ["session_split_plan_60_20_20.csv", "*split*plan*.csv"])

print("MASTER_CSV:", MASTER_CSV)
print("SPLIT_CSV :", SPLIT_CSV)

# quick preview
print("\nMaster head:")
print(pd.read_csv(MASTER_CSV, nrows=2))
print("\nSplit head:")
print(pd.read_csv(SPLIT_CSV, nrows=2))
# === Cell 4: Load master + plan; build session-disjoint Train/Val/Test ===
```

```
dfm = pd.read_csv(MASTER_CSV)
plan = pd.read_csv(SPLIT_CSV)

need = {"Subject","Session","StartSample","Fs"}
miss = need - set(dfm.columns)
if miss:
    raise ValueError(f"Master missing columns: {miss}")

# Support two split plan styles:
# A) row-aligned plan with "Split" column
# B) per-subject lists Train/Val/Test (comma-separated session IDs)
if ("Split" in plan.columns) and ("Subject" in plan.columns) and len(plan) == len(dfm):
    dfm["Split"] = plan["Split"].values
else:
    def _to_set(s):
        s = str(s).strip()
        return set(t for t in s.split(",") if t) if s and s.lower()!="nan" else set()
    tr, va, te = {}, {}, {}
    for _, r in plan.iterrows():
        subj = str(r["Subject"])
        tr[subj] = _to_set(r.get("Train",""))
        va[subj] = _to_set(r.get("Val",""))
        te[subj] = _to_set(r.get("Test",""))
    def _assign(row):
        s, sess = row["Subject"], row["Session"]
        if sess in tr.get(s,set()): return "Train"
        if sess in va.get(s,set()): return "Val"
        if sess in te.get(s,set()): return "Test"
        return "Unassigned"
    dfm["Split"] = dfm.apply(_assign, axis=1)
```

```
# Guard against unassigned rows
if (dfm["Split"]=="Unassigned").any():
    bad = dfm[dfm["Split"]=="Unassigned"][["Subject","Session"]].drop_duplicates().head(20)
    print("Unassigned examples:\n", bad)
    raise AssertionError("Some windows not assigned; fix your split CSV.")

print(dfm["Split"].value_counts())

df_train = dfm[dfm["Split"]=="Train"].copy()
df_val   = dfm[dfm["Split"]=="Val"].copy()
df_test  = dfm[dfm["Split"]=="Test"].copy()

print("\nWindow counts — Train:", len(df_train), " Val:", len(df_val), " Test:", len(df_test))

feat_cols = [c for c in dfm.columns if c not in ["Subject","Session","StartSample","Fs","Split"]]
X_train, y_train = df_train[feat_cols].values, df_train["Subject"].values
X_val,   y_val   = df_val[feat_cols].values,   df_val["Subject"].values
X_test,  y_test  = df_test[feat_cols].values,  df_test["Subject"].values
classes = sorted(dfm["Subject"].unique().tolist())

print("Shapes -> X_train", X_train.shape, "| X_val", X_val.shape, "| X_test", X_test.shape)
print("Subjects:", len(classes))
# === Cell 5: Define models & small grids (fast) ===
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import SVC
from sklearn.preprocessing import StandardScaler
from sklearn.pipeline import Pipeline

rf_pipe = Pipeline([
    ("scaler", StandardScaler(with_mean=True, with_std=True)),
    ("clf", RandomForestClassifier(random_state=RANDOM_SEED))
```

```
])
rf_grid = {
    "clf__n_estimators": [300],    # single strong value (fast)
    "clf__max_depth": [None, 18],  # small choice
    "clf__min_samples_leaf": [1],  # keep simple
}

svm_pipe = Pipeline([
    ("scaler", StandardScaler(with_mean=True, with_std=True)),
    ("clf", SVC(kernel="rbf", probability=True, random_state=RANDOM_SEED))
])
svm_grid = {
    "clf__C": [1, 10],
    "clf__gamma": ["scale"],       # robust default
}
# === Cell 6: Train on TRAIN; choose best by VAL ===
results = []
for name, pipe, grid in [
    ("RandomForest", rf_pipe, rf_grid),
    ("SVM_RBF",      svm_pipe, svm_grid),
]:
    gs = GridSearchCV(pipe, grid, cv=3, n_jobs=-1, scoring="accuracy", verbose=1)
    gs.fit(X_train, y_train)
    best = gs.best_estimator_
    val_acc = accuracy_score(y_val, best.predict(X_val))
    print(f"[{name}] VAL acc={val_acc:.4f} | best={gs.best_params_}")
    results.append((name, val_acc, gs.best_params_, best))

# pick best by validation accuracy
results.sort(key=lambda t: t[1], reverse=True)
best_name, best_val_acc, best_params, best_model = results[0]
```

```
print(f"\nSelected model: {best_name} | VAL acc={best_val_acc:.4f} | params={best_params}")


# Save chosen model
model_path = os.path.join(OUTPUT_DIR, "models",
f"chosen_{best_name}_trainval_sessiondisjoint.joblib")
joblib.dump(best_model, model_path)
print("Saved model:", model_path)
# === Cell 7: Evaluate on TEST (strict, unseen) ===
y_pred_test = best_model.predict(X_test)
test_acc = accuracy_score(y_test, y_pred_test)
print(f"\nTEST acc (session-disjoint) = {test_acc:.4f}")


cm = confusion_matrix(y_test, y_pred_test, labels=classes)
print("Confusion matrix shape:", cm.shape)
# === Cell 8: Save plots & reports (TEST) ===
def plot_confusion_matrix(cm, classes, out_path, title):
    plt.figure(figsize=(10,9))
    plt.imshow(cm, cmap="Blues")
    plt.title(title)
    plt.xticks(np.arange(len(classes)), classes, rotation=90, fontsize=6)
    plt.yticks(np.arange(len(classes)), classes, fontsize=6)
    for i in range(cm.shape[0]):
        for j in range(cm.shape[1]):
            plt.text(j, i, str(cm[i, j]), ha='center', va='center', fontsize=5)
    plt.colorbar()
    plt.tight_layout()
    plt.savefig(out_path, dpi=150)
    plt.close()


def compute_ovr_roc(proba, y_true, class_names, out_png, out_csv):
    Y = label_binarize(y_true, classes=class_names)
```

```
    aucs = []
    plt.figure(figsize=(8,7))
    for i, cname in enumerate(class_names):
        fpr, tpr, thr = roc_curve(Y[:, i], proba[:, i])
        roc_auc = auc(fpr, tpr); aucs.append(roc_auc)
    plt.plot([0,1],[0,1],'--', lw=1)
    plt.xlabel("False Positive Rate"); plt.ylabel("True Positive Rate")
    plt.title(f"OvR ROC (TEST) — Macro AUC={np.mean(aucs):.3f}")
    plt.tight_layout(); plt.savefig(out_png, dpi=150); plt.close()
    pd.DataFrame({"Class": class_names, "AUC": aucs}).to_csv(out_csv, index=False)

# Save CM
cm_png = os.path.join(PLOTS_CM, f"cm_test_{best_name}_sessiondisjoint.png")
plot_confusion_matrix(cm, classes, cm_png, f"Confusion Matrix — {best_name} (TEST) |
acc={test_acc:.4f}")
print("Saved:", cm_png)

# Save ROC (if available)
try:
    y_prob_test = best_model.predict_proba(X_test)
    roc_png = os.path.join(PLOTS_ROC, f"roc_test_{best_name}_sessiondisjoint.png")
    roc_csv = os.path.join(REPORTS_OUT, f"roc_test_{best_name}_sessiondisjoint.csv")
    compute_ovr_roc(y_prob_test, y_test, classes, roc_png, roc_csv)
    print("Saved:", roc_png)
    print("Saved:", roc_csv)
except Exception as e:
    print("ROC not available (predict_proba missing):", e)

# Classification report
rep = classification_report(y_test, y_pred_test, labels=classes, output_dict=True)
rep_df = pd.DataFrame(rep).transpose()
```

```
rep_csv = os.path.join(REPORTS_OUT, f"class_report_test_{best_name}_sessiondisjoint.csv")
rep_df.to_csv(rep_csv)
print("Saved:", rep_csv)
# === Cell 9 (optional): Per-session majority vote on TEST ===
def majority_vote(series):
    s = pd.Series(series)
    return s.mode().iloc[0]

df_test = dfm[dfm["Split"]=="Test"].copy()
df_test["y_pred"] = y_pred_test

session_votes = (
    df_test
    .groupby(["Subject","Session"])
    .agg(true_subject=("Subject","first"),
         pred_subject=("y_pred", majority_vote),
         n_windows=("y_pred","size"))
    .reset_index(drop=True)
)
session_votes["correct"] = (session_votes["true_subject"] ==
session_votes["pred_subject"]).astype(int)
session_acc = session_votes["correct"].mean()
session_csv = os.path.join(SESSION_OUT, "test_session_majority_vote.csv")
session_votes.to_csv(session_csv, index=False)
print(f"Session-level accuracy (TEST, majority vote) = {session_acc:.4f}")
print("Saved:", session_csv)
session_votes.head(10)
# Define a subdirectory for session-level reports inside your OUTPUT_DIR
SESSION_OUT = os.path.join(OUTPUT_DIR, "session_reports")
os.makedirs(SESSION_OUT, exist_ok=True)
print("Session report outputs will be saved in:", SESSION_OUT)
```

```
# === Per-session report (TEST) — window accuracy + majority vote ===
import os
import numpy as np
import pandas as pd

os.makedirs(SESSION_OUT, exist_ok=True)

# 1) Ensure Test rows and predictions exist
try:
    df_test
except NameError:
    df_test = dfm[dfm["Split"] == "Test"].copy()

try:
    y_pred_test
except NameError:
    y_pred_test = best_model.predict(X_test)

# 2) Attach predictions to Test windows
df_test_pred = df_test.copy()
df_test_pred["y_true"]  = df_test_pred["Subject"]
df_test_pred["y_pred"]  = y_pred_test
df_test_pred["correct"] = (df_test_pred["y_true"] == df_test_pred["y_pred"]).astype(int)

# Optional: a single column like S001R09 for readability
df_test_pred["SessionID"] = df_test_pred["Subject"].astype(str) + df_test_pred["Session"].astype(str)

# 3) Per-session window accuracy + majority vote
def majority_vote(series):
    s = pd.Series(series)
    return s.mode().iloc[0]
```

```
per_session = (
    df_test_pred
    .groupby(["Subject","Session"], as_index=False)
    .agg(
        n_windows   = ("correct","size"),
        n_correct   = ("correct","sum"),
        win_acc     = ("correct","mean"),
        majority_pred = ("y_pred", majority_vote),
    )
)
per_session["majority_correct"] = (per_session["majority_pred"] ==
per_session["Subject"]).astype(int)
per_session["SessionID"] = per_session["Subject"].astype(str) + per_session["Session"].astype(str)

# 4) Summary metrics
overall_window_acc   = df_test_pred["correct"].mean()
overall_session_acc  = per_session["majority_correct"].mean()

# 5) Save CSVs
win_csv   = os.path.join(SESSION_OUT, "test_window_predictions.csv")              # every window
sess_csv  = os.path.join(SESSION_OUT, "test_session_report.csv")                  # one row per session
summ_csv  = os.path.join(SESSION_OUT, "test_window_vs_session_summary.csv")       # one-line
summary

df_test_pred.to_csv(win_csv, index=False)
per_session[["Subject","Session","SessionID","n_windows","n_correct","win_acc","majority_pred","
majority_correct"]]\
    .sort_values(["Subject","Session"]).to_csv(sess_csv, index=False)

pd.DataFrame([{
    "PerWindow_Accuracy": float(overall_window_acc),
```

```
    "PerSession_MajorityVote_Accuracy": float(overall_session_acc),
    "Num_Test_Windows": int(len(df_test_pred)),
    "Num_Test_Sessions": int(len(per_session))
}]).to_csv(summ_csv, index=False)

print(f"Per-window Test accuracy (overall)       : {overall_window_acc:.4f}")
print(f"Per-session majority-vote Test accuracy   : {overall_session_acc:.4f}")
print("Saved:")
print(" -", win_csv)
print(" -", sess_csv)
print(" -", summ_csv)

# Preview a few rows
per_session.sort_values(["Subject","Session"]).head(10)
```