Improving The Data Imbalance of Training Datasets for EEG-Based Authentication

Akintoye Oyedola
School of Computing
University of North Florida
Jacksonville, Florida, USA

***Abstract*—The unique behavioral attributes and physiological traits of end-users significantly contribute to the reliability of biometric authentication systems to authenticate or deny access to protected data. The difficulty in duplicating biometric features also protects these authentication systems from intruders. An electroencephalogram (EEG) is a test that measures the electrical activity on the brain surface. Every person's EEG signals are unique. While EEG-based biometric authentication systems perform even better than other biometric authentication systems like fingerprint and iris recognition, they remain unreliable and often deliver false negative results during authentication. These unreliable decisions have two causes: changes in the mental state of the end-user trying to use the authentication feature and imbalanced training datasets for the Machine Learning (ML) models used in EEG-based biometric authentication. This paper details guidelines to produce training datasets that offset the data imbalance and improve the reliability of EEG-based biometric authentication.EEG signals; biometric authentication; machine learning**

***Index Terms*—EEG signals, biometric authentication, machine learning, data imbalance**

I. Introduction

This research addresses the reliability issues of EEG-based authentication systems which limit scalability and application. Biometric authentication systems offer the most data security. They are employed in industries like travel and migration control, law enforcement, medical records, and the military [1]. Current EEG datasets for training ML models for authentication decision-making have high variance levels. Popular ML models for EEG datasets include Support Vector Machine (SVM), Recurrent Neural Networks (RNNs), and Multilayer Perceptron (MLP) deep learning models to analyze the data, train the models, and identify variance levels [2]. These imbalanced datasets significantly reduce the recognition accuracy of EEG-based biometric authentication systems and make them impractical for large-scale use in banking, airport security, and military data centers. Producing and analyzing an EEG dataset that performs better than current training datasets would help improve the performance of ML models and consequently minimize false negative authentication decisions. In this context, a False Negative decision denies access to an authorized user. The first objective of this research is to improve the accuracy of EEG signal classifiers for better reliability and consequently better data security. The second objective is to produce training datasets that perform better than publicly available EEG datasets for training EEG-based ML models. This study is important because of the practical need to make EEG-based authentication more reliable as it is a highly secure authentication method. The main contributions of this paper are:

- The production of a UNF-EEG dataset with a new technique for data pre-processing.
- Comparison of EEG datasets that show better recognition accuracy and lower variance in the UNF-EEG. The imbalance of a training dataset determines the recognition accuracy of ML models used for EEG-based biometric authentication.

II. Background

Authentication is verifying the identity and permissions of a user or entity before enabling access to resources or systems. In computing, authentication is crucial to access management, data security, and resource allocation [3]. Biometric authentication is regarded as the most secure user authentication method due to its unique behavioral attributes and physiological traits that are difficult to replicate [4]. Behavioral attributes include typing patterns and gait while iris data and fingerprint information are examples of physiological attributes. Also, biometric authentication is secure because user biometric credentials cannot be lost or forgotten compared to passwords or access cards [5]. Fingerprint, facial recognition, iris, and voice recognition are the most popular biometric authentication methods. Biometric authentication occurs in two stages: enrollment and verification. During enrollment, a biometric scanner records a user's behavioral or physical attributes for which feature extraction algorithms are used to produce a biometric modality [6]. During verification, a matcher module compares the biometric modality of a registered user with an encrypted biometric submission. The similarity of both submissions determines whether access is granted or denied. In addition to better data security and recognition accuracy, biometric authentication is more user-friendly than traditional authentication methods like passwords and PINs since it passively acquires a user's biometric data during the enrollment and verification stages. Users do not need to remember the answer to a security question or provide a complex account password.

Electroencephalogram (EEG) records the brain's electrical activities caused by synaptic activations of the brain's neurons. EEGs are unique to every individual [7]. EEG-based

authentication records EEG signals while a user performs a task, pre-processes the signal data to enhance their quality, and uses deep learning to train a machine learning model to perform authentication. In EEG-based biometric authentication, a brain-computer interface (BCI) creates a path between a brain and a computer to decode brain wave patterns, human intentions, and mental states [8]. EEG-based authentication is applied to multimodal authentication systems that extract more than one biometric feature, BCIs that form a direct path between a human brain and a computer, and cryptographic key generation [9]. Extensive research proves that EEG biometric signals are more resistant to security breaches like spoof attacks and biometric feature duplication compared to other biometric authentication methods [9]. Unlike fingerprint or iris recognition systems, EEG data is inaccessible to attackers and practically immune to replication [9]. In addition, EEG biometrics is resistant to force situations where an attacker pressurizes a user to approve authentication so the attacker can gain unauthorized access. EEGs rely on emotional state and EEG biometric authentication systems will deny access upon changes to the normal brain waves' pattern [9]. To compare, an attacker could access other biometric authentication systems by threatening the user. EEG-based biometric authentication is mainly applied in the medical field to safeguard access to medical records. While brain-computer interface devices like EMOTIV System and NeuroSky have become portable and commercial [10], [11], these devices still pose security risks [12].

Successful EEG-based authentication relies on the user's cognitive ability and mental state. High-stress situations significantly impact recognition accuracy. Studies show that EEG signals may not be stable enough to scale EEG-based authentication applications beyond the medical field [13]. Available EEG datasets show a wide variation in recognition accuracy even though most of EEG datasets often have small participant sizes. Aznan et al. proved that using synthetic EEG data could improve the convergence speed of machine learning models [14] and Piplani et al. improved the security of EEG-based authentication by adding adversarial data to their training datasets [15]. This study addresses the variability issue, provides a clean dataset for future research, and creates a guideline for scaling EEG-based authentication to achieve improved user buy-in.

III. Objectives

The instability of EEG signals poses a reliability issue for the verification stage of the authentication process. Reliability is one reason why EEG-based authentication systems are currently limited to the medical field [9]. A user's mental state varies each time they use an authentication system and current EEG datasets only contain raw, unfiltered EEG signals. This data is used to train Machine Learning (ML) models like SVM, RNNs, and MLP to either grant or deny access to protected data and systems [14]. Consequently, the high variance in EEG datasets for training ML models negatively affects the recognition accuracy of this authentication method. Scalability is another application issue because EEG-based authentication systems require complex and non-user-friendly technology [7]. The tools necessary for this authentication method are expensive and difficult for the average user to use. This research explores affordable and portable alternatives that improve the user-friendly rating of EEG-based authentication systems. This paper addresses the following research questions:

1) How do we improve the accuracy of EEG signal classifiers for better reliability and consequently better data security?
2) Can a produced training dataset perform better than publicly available EEG datasets for training EEG-based ML models?

IV. Methods

I will answer the first research question by imitating credible pre-processing techniques adopted in neuroscience research. Notch filters and MNE-Python are the main tools to generate a noiseless and evenly distributed training dataset. By producing a UNF-EEG dataset, I will filter it correctly and compare its variance levels with current training datasets. These results will determine the best way to pre-process noiseless datasets for EEG data. The second research question requires a thorough experiment to compare the performance levels of a produced UNF-EEG dataset with the publicly sourced datasets introduced in the next section. Thus, I will experiment to create an EEG dataset to answer this question.

V. Data

In this section, the data acquisition process is clearly defined and justified. Then, detailed information on the data analysis was presented.

A. Data Acquisition

For this experiment, I will collect three datasets and produce one dataset. The goal is to compare the results of the procured dataset with common EEG datasets available online to minimize bias and demonstrate repeatability. I will collect the following three datasets for this experiment:

1) An established EEG database was recorded via a Neuroscan EEG/ERP on a dedicated recording PC [16]. This database, curated by Ouyand et al., can be accessed here.
2) A public download of raw EEG signals from previous research studies published on the credible Human Electrophysiology, Anatomic Data, and Integrated Tools (HeadIT) resource software.
3) A Harvard University EEG Motor Movement/Imagery signals dataset.

I will also produce a UNF-EEG dataset by sharing an interest survey for 80 University of North Florida (UNF) community participants. The volunteering participants will get no tangible reward. Data acquisition will occur in the Virtual Learning Center (VLC) on the third floor of the University of North Florida's library. The VLC is well-lit and equipped with computers and virtual reality headsets. I will incorporate NeuroSky's NeuroView specialization software accessible via

NeuroSky Research Tools into the library's VR headsets. Researchers can use NeuroView to view and record EEG signals in a Comma-Separate-Values (CSV) file. I will record the electric pulses of the participants as their brains react to different stimuli during five tasks:

- Task 1 - Reading a paragraph excerpt from Hamlet.
- Task 2 – Mental multiplication calculations. For example, 56 multiplied by 16 while maintaining physical stillness.
- Task 3 - Signing into a UNF myWings account.
- Task 4 - Signing into a UNF myWings account after a sudden jump scare (to mimic an abnormal mental state).
- Task 5 – Participants in a relaxed posture, allowing their minds to rest without wandering thoughts. This task will serve as the baseline.

Tasks 1-3 occur in the mental state with physical or imaginary body movements. Task 4 introduces external stimuli to disrupt the participant's normal mental state. It mimics a scenario where an attacker uses coercion to convince an authorized user to grant them access to a protected system. Task 5 is a resting state. After the experiment, I will export the CSV file to an Excel spreadsheet.

B. Data Analysis

The Ouyand, HeadIT, Harvard, and UNF-EEG datasets will be preprocessed using MNE-Python and notch filters to remove noise and reveal true EEG signals. This process mimics the methodological EEG-preprocessing guidelines for neuroscience research as seen here. Then, I will calculate and compare the F1 score of each dataset. F1 score is the harmonic mean of the precision and recall of a classification model. 1.0 is a perfect F1 score which indicates perfect precision and recall. 0 is the lowest possible score. Precision is the proportion of positive predictions that are positive. Recall is the proportion of True Positive (TP) occurrences where positive predictions are correctly identified. These evaluation metrics account for the inaccurate categorization of minority class predictions in imbalanced datasets. The equations are shown here:

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

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

$$Recall = \frac{TP}{FP + FN} \tag{3}$$

$$F1Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

Where TP denotes True Positive, TN is True Negative, FP is False Positive, and FN is False Negative.

The UNF-EEG dataset would achieve a predicted F1 score that is 10% better than the other 3 datasets due to improved data acquisition technology like the NeuroView software and the novel pre-processing technique. Subsequently, I will use the supervised Support Vector Machine (SVM) machine learning model, the Recurrent Neural Networks (RNNs), and the Multilayer Perceptron (MLP) deep learning model to analyze the data, train the models, and identify variance levels. RNNs excel at sequence modeling and show improved results with added user history inputs [2]. The MLP model uses neuron layers to classify data correctly. I predict the UNF-EEG dataset produces a similar variance level as the other 3 publicly sourced datasets. I also anticipate that the MLP and RNN models will record a higher recognition accuracy probability when exposed to only the UNF-EEG training dataset. In contrast, I predict the SVM performs poorly in EEG signal classification compared to the other training models.

VI. Execution Plan

The research aims to determine the best training dataset for EEG-based authentication systems using SVM, RNNs, and MLP models to compare the recognition accuracy of four datasets. The research timeline is 7 months and it will occur in 6 stages.

- Stage 1 - *Data Collection*. This involves downloading the CSV files of the Ouyand, HeadIT, and Harvard EEG signals datasets as mentioned in the Data Analysis subsection. This process takes minutes as I will use computer resources at the UNF library.
- Stage 2 – *Recruit Participants*. I procure the requirement to produce the UNF-EEG dataset. I purchase the NeuroView software and NeuroSky Research Tools for $499.99 from this website [10], [11]. Then, I access the Qualtrics Survey Tool for free with my UNF account and create a survey for potential participants. This survey will include basic inquiries about participants' UNF affiliation, gender, knowledge of EEG, and brain-related medical history. It will also contain a detailed description of the experiment and ask participants to submit their UNF email addresses if the experiment interests them. This stage may take up to 2 months to reach the target size of 80 participants.
- Stage 3 – *Allocate Tasks*. Here, I contact two research assistants for the experiment. Research Assistant A acclimatizes to the procured software while Research Assistant B drafts a participant survey for the UNF-EEG dataset acquisition experiment. I reserve the Virtual Learning Center at the UNF library for multiple sessions to carry out the experiment to produce the UNF-EEG dataset. Stage 3 is completed in one month.

- Stage 4 – *Experiment*. Each interested survey respondent schedules one 45-minute session. Then, they sign a participation waiver exempting UNF and all three researchers from liability. Each participant sits in a chair, wears a NeuroView-equipped VR headset, and plays a simple trivia game.
  During this activity, Research Assistant A observes the participant's demeanor and decision-making. Research Assistant B employs NeuroSky's Research Tools to record EEG signals via the participant's headset. Simultaneously, I direct the participant to perform the 5 tasks in the Data Acquisition subsection. I will be responsible for scaring the participants with sudden movements and sounds to surprise them. This disrupts the normal mental state of each participant, allowing our UNF-EEG dataset to include a minority class of scenarios where adversaries use coercion to make users provide unauthorized access to protected data and systems. After each 45-minute session, participants answer a brief questionnaire about their experience before they leave the VLC. The experiment stage will be completed within two months.
- Stage 5 – *Data Handling*. After Stage 4, my Research Assistants and I will export the UNF-EEG dataset into an Excel spreadsheet and then encrypt it using AED. Next, we will perform our data analysis as highlighted in the Data Analysis subsection. Then, my Research Assistants and I will thoroughly pre-process and analyze the datasets before training the SVM, RNN, and MLP learning models to compare performance levels. This stage may take up to one month to complete.
- Stage 6 - *Present Results*. The final stage involves data visualization, discussion about the experiment and its results, and an outline for further analysis or future work in this area. This stage occurs within three weeks.

I will present LaTeX submissions to the industry standard Association for Computing Machinery (ACM) Journal of Machine Learning Research and the Institute of Electrical and Electronics Engineers (IEEE) International Conference on Neural Engineering (NER) within 1 year of completing the research.

A. Ethical Considerations

1) The publicly available datasets protect the identities of the observed participants. Anyone who accesses the databases cannot view the identifying information of any participant.
2) I ensure all participants are strangers to all three researchers to prevent affinity bias.
3) I will use the Advanced Encryption Standard (AES) algorithm to secure the UNF-EEG dataset before exporting it as a CSV file to prevent unauthorized access. The UNF-EEG dataset will also be password-protected. This ensures that all three researchers must be present to provide unique passwords to access the encrypted dataset. This prevents a scenario where one researcher covertly accesses the dataset to modify the data or influence the qualitative results.
4) All participants must demonstrate a detailed understanding of the experiment process. This minimizes the wariness of the research and holds all three researchers accountable. Participants can recognize if a researcher's actions contradict the experiment's objectives. Consequently, participants may elect to end the session and make a formal complaint.

VII. Expected Results

The minority class of the UNF-EEG dataset contains instances of unstable EEG signals that should fail authentication. The majority class refers to EEG data that are acceptable for authentication by trained ML models. The UNF-EEG dataset's preprocessing occurs in 3 stages after the initial MNE-Python filtering technique highlighted in the Data Analysis subsection. Firstly, I would use the oversampling model to populate the dataset. Then, I would generate synthetic data via multivariate Gaussian heuristic functions to improve the validity of synthetic data introduced to the UNF-EEG dataset. This step maximizes the performance of the oversampling model. Finally, I will employ Generative Adversarial Networks (GANs) to achieve class balance for the minority class data distribution. After the 3-stage data treatment, the UNF-EEG would record better performance results at recognition accuracy than the Ouyand, HeadIT, and Harvard datasets. The chart below shows the composition for the UNF-EEG dataset after the 3-stage data processing technique.

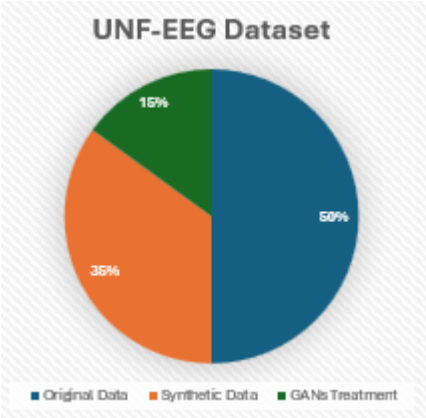

Fig. 1. Composition of the UNF-EEG dataset

Krishnan et al. determined that oversampling is the most effective sampling model for creating a balanced dataset [17]. Oversampling decreases the majority class bias by bolstering the minority class predictions and increasing training examples for improved feature classification. As such, I would apply the oversampling method to the UNF-EEG dataset in contrast to the Ouyand, HeadIT, and Harvard EEG University datasets. I expect the UNF-EEG dataset to be balanced 5% more than the other datasets as the oversampling method distributes the majority and minority classes more

evenly. The table and chart below reflect this comparison.

Table I: Performance Comparison of Datasets With The Oversampling Technique

Dataset	Accuracy	Precision	Recall	F1 Score
UNF-EEG	99.96%	98.03%	100.00%	98.98%
Ouyand	95.30%	96.59%	94.92%	94.99%
HeadIT	95.51%	95.68%	93.49%	93.98%
Harvard	97.46%	94.24%	91.63%	92.48%

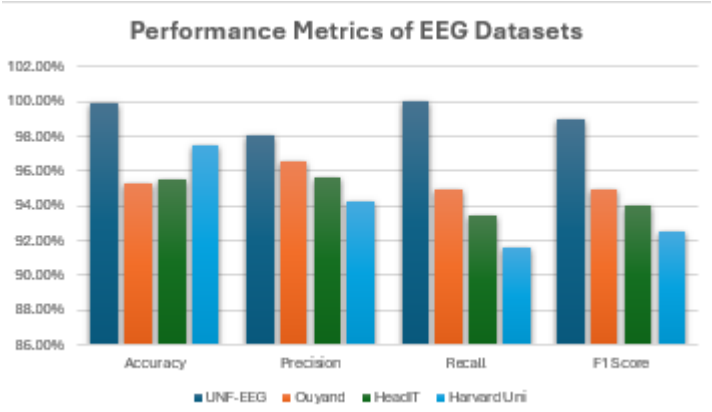

Fig. 2. Performance Comparison of EEG Datasets

Additionally, Thogarchety and Das proved that genetic algorithms with heuristic or vector functions provide the best statistics when measuring the precision, accuracy, recall, and F1 score of training datasets [18]. Synthetic data generation via multivariate Gaussian heuristic functions performs better than other options at training ML models for biometric authentication [18]. Heuristic functions provide better synthetic data points to buffer the original UNF-EEG dataset. This synthetic data generation technique solves the class imbalance problem for classification models and mimics real-world data points. Using Generative Adversarial Networks (GANs) to generate highly structured data is an understudied area due to reliability issues with feature classification [19]. Jia and Zhang developed a process to enhance the data of minority classes, guarantee class balance for training datasets, and prevent the reliability issues of GANs [19]. As such, the final stage of data pre-processing for the UNF-EEG dataset uses a multi-generator to learn the data distribution of the minority class and provide class balance for synthetic data generation.

VIII. Conclusion

The experiment in this paper produced a UNF-EEG dataset and extensively pre-processed it with innovative techniques to achieve class balance for the minority class. Comparison with publicly available EEG datasets shows the superiority of the UNF-EEG dataset for training machine learning models for EEG-based biometric authentication. The experiment results show that the UNF-EEG dataset performs better than current state-of-the-art training datasets. Yet, future research should focus on real-world applications of the specific combination of pre-processing techniques utilized in this research in EEG-based authentication systems.

References

[1] A. Hadid, N. Evans, S. Marcel, and J. Fierrez, "Biometrics systems under spoofing attack: An evaluation methodology and lessons learned," *IEEE Signal Processing Magazine*, vol. 32, no. 5, pp. 20–30, 2015.

[2] Z. Guan, J. Wang, X. Wang, W. Xin, J. Cui, and X. Jing, "A comparative study of rnn-based methods for web malicious code detection," in *2021 IEEE 6th International Conference on Computer and Communication Systems (ICCCS)*, 2021, pp. 769–773.

[3] M. Saqib and A. H. Moon, "A systematic security assessment and review of internet of things in the context of authentication," *Computers & Security*, vol. 125, p. 103053, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016740482200445X

[4] E. Baha, A. Fadhel, P. Buenaventura, C. Y. Yeun, J. Zemerly, and K. Eldelbi, "Multimodal biometric authentication systems: Exploring iris and eeg data," in *2024 2nd International Conference on Cyber Resilience (ICCR)*, 2024, pp. 1–4.

[5] R. Ryu, S. Yeom, S.-H. Kim, and D. Herbert, "Continuous multimodal biometric authentication schemes: A systematic review," *IEEE Access*, vol. 9, pp. 34 541–34 557, 2021.

[6] U. Sumalatha, K. K. Prakasha, S. Prabhu, and V. C. Nayak, "A comprehensive review of unimodal and multimodal fingerprint biometric authentication systems: Fusion, attacks, and template protection," *IEEE Access*, vol. 12, pp. 64 300–64 334, 2024.

[7] T. Nishimoto, H. Higashi, H. Morioka, and S. Ishii, "Eeg-based personal identification method using unsupervised feature extraction and its robustness against intra-subject variability," *Journal of Neural Engineering*, vol. 17, no. 2, p. 026007, 2020.

[8] M. A. Lebedev, A. J. Tate, T. L. Hanson, Z. Li, J. E. O'Doherty, J. A. Winans, P. J. Ifft, K. Z. Zhuang, N. A. Fitzsimmons, D. A. Schwarz, A. M. Fuller, J. H. An, and M. A. L. Nicolelis, "Future developments in brain-machine interface research," *Clinics*, vol. 66, pp. 25–32, 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1807593222015812

[9] A. Jalaly Bidgoly, H. Jalaly Bidgoly, and Z. Arezoumand, "A survey on methods and challenges in eeg based authentication," *Computers & Security*, vol. 93, p. 101788, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167404820300730

[10] K. Stytsenko, E. Jablonskis, and C. Prahm, "Evaluation of consumer eeg device emotiv epoc," in *MEi: CogSci Conference*, 2011, p. 99.

[11] J. Katona, I. Farkas, T. Ujbanyi, P. Dukan, and A. Kovari, "Evaluation of the neurosky mindflex eeg headset brain waves data," in *2014 IEEE 12th international symposium on applied machine intelligence and informatics (SAMI)*. IEEE, 2014, pp. 91–94.

[12] K. Xia, W. Duch, Y. Sun, K. Xu, W. Fang, H. Luo, Y. Zhang, D. Sang, X. Xu, F.-Y. Wang, and D. Wu, "Privacy-preserving brain–computer interfaces: A systematic review," *IEEE Transactions on Computational Social Systems*, vol. 10, no. 5, pp. 2312–2324, 2023.

[13] R. Das, E. Maiorana, and P. Campisi, "Motor imagery for eeg biometrics using convolutional neural network," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 2062–2066.

[14] N. K. Nik Aznan, A. Atapour-Abarghouei, S. Bonner, J. D. Connolly, N. Al Moubayed, and T. P. Breckon, "Simulating brain signals: Creating synthetic eeg data via neural-based generative models for improved ssvep classification," in *2019 International Joint Conference on Neural Networks (IJCNN)*, 2019, pp. 1–8.

[15] T. Piplani, N. Merill, and J. Chuang, "Faking it, making it: Fooling and improving brain-based authentication with generative adversarial networks," in *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 2018, pp. 1–7.

[16] R. Ouyang, X. Wu, and Z. Lv, "Personal identification and authentication in multi-task eeg database using eegnet and siamese network," in *2024 International Joint Conference on Neural Networks (IJCNN)*, 2024, pp. 1–8.

[17] L. P. Krishnan, I. Vakilinia, S. Reddivari, and S. Ahuja, "Handling imbalanced data for detecting scams in ethereum transactions using sampling techniques," in *2024 12th International Symposium on Digital Forensics and Security (ISDFS)*, 2024, pp. 1–6.

[18] P. Thogarchety and K. Das, "Synthetic data generation using genetic algorithm," in *2023 2nd International Conference for Innovation in Technology (INOCON)*, 2023, pp. 1–6.

[19] Y. Jia and X. Zhang, "An approach for generating high quality structured data," in *2023 5th International Conference on Communications, Information System and Computer Engineering (CISCE)*, 2023, pp. 406–410.