

Highlights

A Statistical Approach for Synthetic EEG Data Generation

Gideon Vos, Maryam Ebrahimpour, Liza van Eijk, Zoltan Sarnyai, Mostafa Rahimi Azghadi

- EEG data can play a key role in diagnosing mental health conditions, but collecting large datasets is resource and time-intensive.
- Synthetic data offers a means to reduce reliance on extensive real-world recordings, accelerate research, and safeguard patient privacy.
- Existing methods for generating synthetic EEG data often rely on advanced deep learning models that require substantial computational resources and specialized technologies.
- The proposed method employs standard correlation analysis and random sampling to generate synthetic EEG data while maintaining the integrity of key signals.
- In favor of reproducible research and to advance the field, all programming code used in this study is made publicly available.

A Statistical Approach for Synthetic EEG Data Generation

Gideon Vos^a, Maryam Ebrahimpour^a, Liza van Eijk^b, Zoltan Sarnyai^c,
Mostafa Rahimi Azghadi^a

^a*College of Science and Engineering, James Cook University, James Cook
Dr, Townsville, 4811, QLD, Australia*

^b*College of Health Care Sciences, James Cook University, James Cook
Dr, Townsville, 4811, QLD, Australia*

^c*College of Public Health, Medical, and Vet Sciences, James Cook University, James
Cook Dr, Townsville, 4811, QLD, Australia*

Abstract

Introduction. Electroencephalogram data plays a critical role in understanding and diagnosing mental health conditions. However, recording EEG data is both costly and time-consuming, particularly when aiming to build large datasets required for training machine learning models. This limitation has driven interest in synthetic data generation as a means to augment existing datasets. Synthetic data not only reduces the dependency on extensive real-world recordings but also accelerates research by providing readily available training samples. While previous studies have explored various methods for data augmentation, generating high-quality synthetic EEG data that preserves the integrity of emotional and mental health signals remains a challenge. This study addresses this gap by proposing a method to generate synthetic EEG data using correlation analysis and random sampling.

Methods. Correlation analysis was used to determine interdependencies between frequency bands in original EEG data. Next, synthetic EEG samples were generated by leveraging random sampling techniques, guided by the correlation structure of the original EEG data. Synthetic samples were next tested against the original dataset using correlation analysis to ensure fidelity, and samples with high correlation were retained. Finally, the generated synthetic EEG data was subjected to distribution analysis and machine learning classification models trained to distinguish between original and synthetic samples, serving as a benchmark for the quality of the synthetic data.

Results. The synthetic EEG data generated using our proposed method exhibits high fidelity to the original dataset, while preserving subject emotional and mental health state. Similar correlation coefficients between the synthetic and original data confirmed the preservation of the underlying structure, with the synthetic EEG data matching the distribution of the original EEG data, while PERMANOVA analysis showing no statistical difference. A Random Forest machine learning classification model trained to classify synthetic versus original samples performed no better than random guessing, indicating the inability to distinguish between the two datasets.

Conclusion. This study presents a robust method for generating synthetic EEG data for brain health studies. By leveraging correlation analysis and random sampling, the proposed method creates synthetic data that closely mimic the characteristics of real-world EEG data. The inability to distinguish between synthetic and original samples by machine learning models underscores the quality of the generated data. This approach offers a cost-effective and scalable solution for augmenting EEG datasets, enabling more efficient training of machine learning models for mental health applications, while enhancing patient privacy.

Keywords: Machine Learning, EEG, Synthetic Data

PACS: 07.05.Mh, 87.19.La

2000 MSC: 68T01, 92-08

1. Introduction

Electroencephalography (EEG) is a non-invasive technique that records electrical activity in the brain through electrodes placed on the scalp. It has become a cornerstone in diagnosing and monitoring a variety of psychiatric [1, 2] and neuropsychiatric disorders, including epilepsy [3–5], depression [6–8], schizophrenia [9, 10], and other neurological conditions. EEG is particularly valued for its ability to capture real-time brain activity, offering clinicians and researchers a window into the dynamic processes of the brain. Its affordability, portability, and non-invasive nature make it a practical tool in both clinical and research contexts.

A critical limitation in EEG-based research and clinical applications is the scarcity of high-quality, labeled datasets. The recording and collection of

EEG data is time-consuming, resource-intensive [11], and subject to stringent privacy regulations [12, 13]. Furthermore, the variability in EEG patterns across individuals [14], coupled with differences in recording protocols [15], often leads to challenges in creating large, standardized datasets. These limitations hinder the training and validation of machine learning (ML) models, which require diverse and representative data to achieve robust performance. Addressing these challenges is essential for advancing the use of ML in EEG analysis, particularly in developing diagnostic tools and personalized treatment strategies.

Synthetic data generation has emerged as a promising solution to these challenges [16–20]. Synthetic data refers to artificially generated datasets that mimic the statistical properties and structural patterns of real data without compromising individual privacy. This approach, first conceptualized over three decades ago [21, 22], has gained significant traction in recent years, driven by advancements in artificial intelligence (AI) and ML. Techniques such as generative adversarial networks (GANs) [6, 23], ensemble models [24], and latent diffusion models [25] have been employed to generate synthetic biomarker and EEG data that closely resembles real recordings. These synthetic datasets enable researchers to overcome data scarcity, enhance model generalizability, and facilitate data sharing across institutions without violating privacy regulations.

The potential applications of synthetic EEG data in ML are vast. By augmenting real datasets with synthetic data, researchers can address bias [19, 26], improve the performance of predictive models [3, 27–29] and democratize research through the publication of open data [20, 30]. Synthetic data also allows for the simulation of diverse patient populations, ensuring that ML models are trained on data that reflects the variability in real-world scenarios. This is particularly important in personalized medicine, where treatment recommendations must account for individual differences in brain activity and response patterns.

Despite its advantages, the use of synthetic data in medical research is not without challenges [13, 18, 30, 31]. Ensuring the fidelity and reliability of synthetic EEG data is critical [32], as any discrepancies between synthetic and real data could impact the performance of ML models. Additionally, the ethical implications of synthetic data generation, including potential misuse

and the need for transparency in data creation processes, must be carefully considered [13]. Addressing these challenges requires ongoing research and the development of standardized guidelines for the generation and use of synthetic data in healthcare.

2. Related Work

Rujas *et al.* [16] conducted a systematic review examining the application of synthetic data generation in healthcare, reporting that 36% of studies focused on its use in developing image classification machine learning models. Similarly, Pezoulas *et al.* [20] highlighted a significant rise in publications exploring synthetic data in healthcare. They identified cost and time efficiency as key drivers for synthetic data adoption, followed by the enhancement of privacy protections. Their review also revealed that deep learning-based synthetic data generators were employed in 72.6% of studies, with statistical approaches accounting for 15.1%. Among deep learning methods, GANs were the most frequently utilized. However, the synthetic generation of EEG data remains an under-explored area of research [6, 16–18, 20]

The use of deep learning techniques for synthetic data generation has been associated with a modest improvement in predictive accuracy, averaging 4%, compared to models trained on original data [3, 28, 29]. This improvement is likely due to the introduction of controlled noise into the training data. However, training GANs is computationally demanding, time-intensive, and technically complex [20]. Alternative methods, including Variational Autoencoders (VAEs) [33], diffusion models [28, 33], nonparametric tree-based techniques [31], and Bayesian networks [26, 31], have also been utilized in a number of prior studies with some success. Notably, Rankin *et al.* [31] investigated the reliability of supervised machine learning models trained on synthetic data. Their findings indicated that tree-based classifiers, compared to deep-learning models, are particularly sensitive to synthetic data, with 92% of models tested demonstrating reduced predictive accuracy, compared to those trained on original, non-synthetic data.

In this study, we present a computationally efficient and scalable method for generating synthetic EEG data using standard statistical approaches, including random sampling and Spearman correlation analysis. This method

aims to enhance the predictive performance of EEG-driven machine learning models in healthcare applications, offering a practical solution for addressing data scarcity while maintaining computational feasibility.

3. Methods

3.1. Datasets

Three EEG datasets were utilized in this study (Table 1). The EEG During Mental Arithmetic Tasks (Stress) dataset [34] is provided pre-labeled for a relaxed state and an acute stress state during an arithmetic task. This dataset was additionally used to build a regression stress prediction model using XGBoost [35] with data split 70%/30% for training and testing, achieving an Area Under the Curve (AUC) of 1.0. The resulting model was utilized to predict stress on the additional datasets included in this study, with the aim of testing whether external emotion prediction can be reliably synthesized across datasets.

For the SAM40 dataset (SAM) [36], subjects were recorded for 25-second intervals while performing four different tasks: the stroop color-word test (SCWT), solving arithmetic questions, identification of symmetric mirror images, and a state of relaxation, with 3 trials recorded for each. The stress prediction model built using XGBoost and the Stress dataset was used to predict acute stress during each trial of the 4 tasks.

The third dataset, Mental Workload (Workload) [37] was collected while subjects performed low, medium and high levels of two different complex tasks which included the N-back test game [38] to enforce the short term memory, and a flight simulation. Performance scores were attributed during each task to measure each subject’s ability to perform under each difficulty level. The stress prediction model built using XGBoost and the Stress dataset was again used to predict acute stress during each of the 3 difficulty levels for both tasks. For this dataset, the two tasks were separated into Mental Workload Dataset 1 and Mental Workload Dataset 2, resulting in a total of 4 experimental datasets for this paper. This separation was performed to evaluate data synthesis across tasks for the same subject group.

Table 1: Datasets utilized in this study.

Dataset	EEG Device	Channels	Subjects
SAM40 [36]	Emotiv Epoc Flex	32	40 (14F, 26M, mean age: 21.5 years)
EEG During Mental Arithmetic Tasks [34]	Neurocom	23	35 (26F, 9M, mean age: 18.25 years)
Mental Workload [37]	Emotiv Epoc X	14	15 (age: 20-60)

3.2. Pre-processing

To ensure a consistent and standardized pre-processing pipeline, artefact removal was uniformly applied to all four datasets, regardless of whether they were reported as artefact-free. Pre-processing steps included average referencing across all EEG electrodes, followed by the application of a band-pass filter with a frequency range of 1 Hz to 45 Hz. Independent Component Analysis (ICA) was subsequently performed using the MNE library [39], to identify and remove ocular, muscular and other potential artefacts. Finally, all datasets were resampled to 250 Hz to maintain frequency uniformity.

The artefact-free data of each dataset were then transformed into the frequency domain to extract power in the alpha, beta, delta, theta, and gamma bands across the frontal, central, parietal, occipital, and temporal regions, if available. This step was designed to mitigate any potential impact of sensor mismatch arising from the use of different recording devices (see Table 1) across the original 3 datasets. Finally, Spearman correlation analysis was conducted for each dataset, serving as a baseline threshold for the synthetic data generation phase. The pre-processing pipeline and correlation analysis workflows are illustrated in Figure 1, with the final output data denoted as (A).

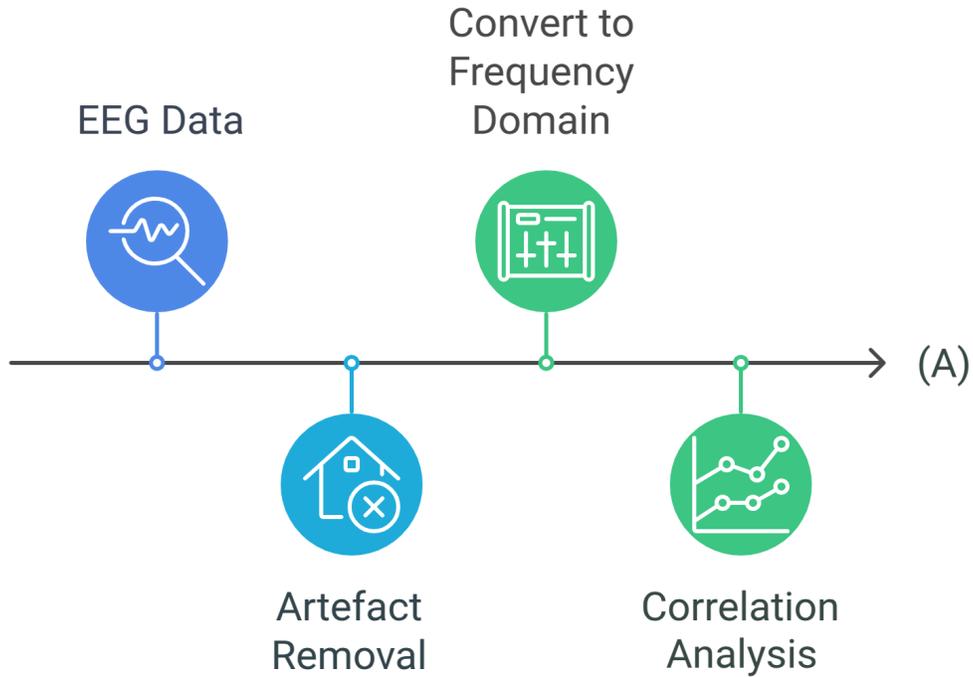


Figure 1: Standard EEG preprocessing pipeline with final correlation analysis.

3.3. Synthesis

Generation of N synthetic samples for each of the 4 datasets involves the following step by step process:

1. N Random samples are selected from the pre-processed dataset (Figure 1, denoted as A).
2. Spearman correlation is performed between the selected N samples and the rest of the pre-processed data (A).
3. Samples exhibiting less correlation than a specified threshold (0.20 in the study experiments) are discarded, with the remaining retained as (B).
4. Steps 2 to 3 are repeated until sufficient (B) samples are retained to equal the input parameter N .

In this study, 10-second epochs were generated for each dataset prior to conversion to the frequency domain (Figure 1). Therefore, if 20 10-second

epochs of synthetic data is required, N should be set to 20. Epoch durations of 10 seconds were selected due to the short recording length of subject samples within the original datasets, however this is not mandatory and can be adjusted as required based on input data recording duration. The synthetic data generation process flow is detailed in Figure 2, with final synthetic samples denoted as (B).

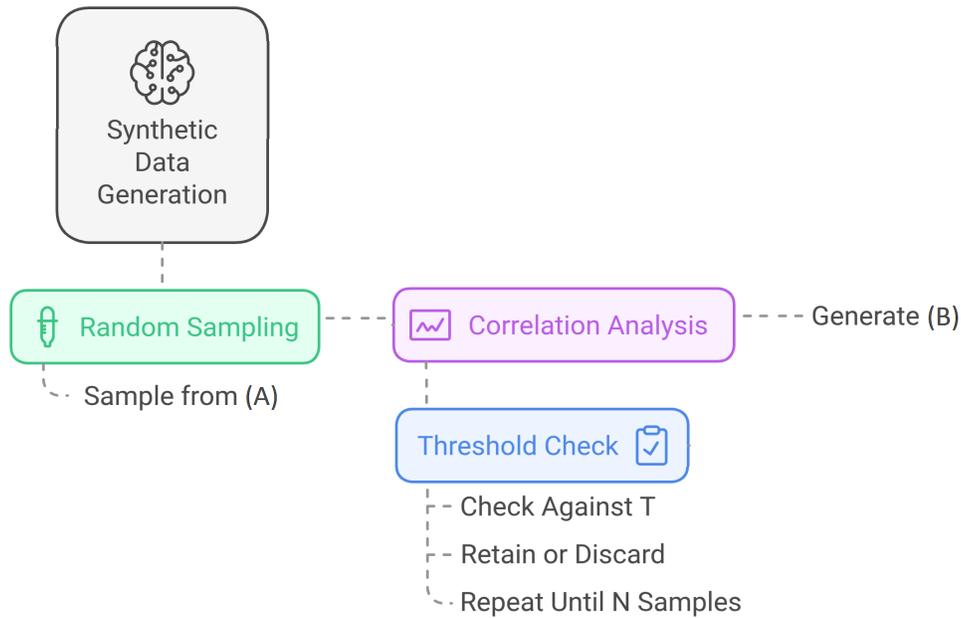


Figure 2: Synthetic EEG data generation process flow.

3.4. Validation

In order to validate the quality of the generated synthetic samples, the following statistical and machine learning approaches were subsequently applied:

- Distributions of original (A) and synthetic samples (B) were plotted and visually compared.
- A non-parametric PERMANOVA test was performed on (B) after testing for multivariate normality using the Shapiro-Wilk test.

- A Random Forest (RF) classification model was built after labeling and merging both original (A) with synthetic samples (B), and trained to classify input data as either original (class 0) or synthetic (class 1).
- An RF model was trained on the original dataset (A) to predict the assigned stress label, and evaluated against the stress label of the synthetically generated samples (B).
- The prior step was repeated, by training on the synthetic samples (B) to predict the stress label, and evaluated against the original data (A).

Synthetic samples constituting 70 10-second epochs were generated and evaluated for each of the four datasets (Table 1 with workload dataset split into two datasets). This number ($N=70$) of samples were selected due to the relatively small sample sizes of the original datasets utilized in this study.

In order to compare the proposed method to existing synthesis methods including the use of Generative Adversarial Networks (GAN) and Variational Autoencoders (VAE), two experiments were performed to synthesize EEG data using the EEG During Mental Arithmetic Tasks (Stress) dataset [34]. Distributions of original and synthetic samples were plotted and visually compared before training an RF model to predict the assigned stress label, and evaluated against the stress label of the synthetically generated samples.

4. Results

Following the validation process described in Section 3.4, distribution plots for each dataset were generated, as shown in Figures 3 and 4. Analyzing these distributions is a crucial step in evaluating the quality of synthetic EEG data [6]. By comparing the statistical properties of synthetic and real data, distribution analysis helps assess whether the synthetic samples capture the underlying patterns and variability of the original dataset. In this study, the distributions of the synthetic data for all four datasets closely resemble those of the original data, with minor deviations observed in Mental Workload Dataset 1 across four of the five recorded brain regions.

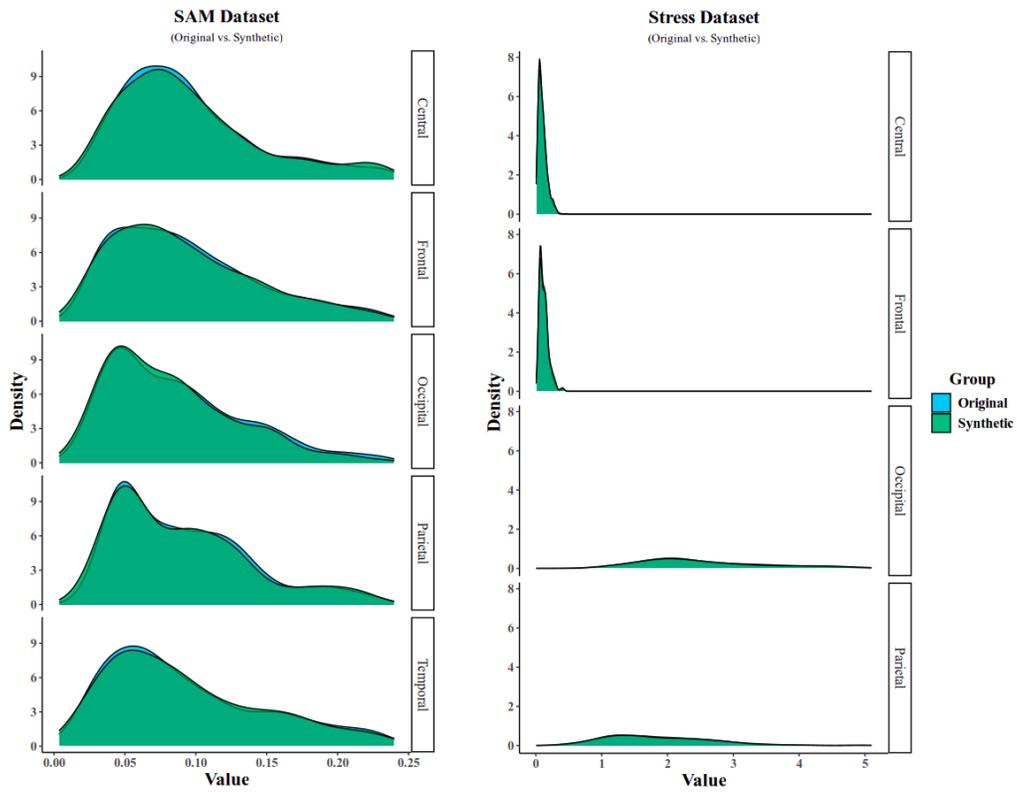


Figure 3: Distribution analysis of SAM and Stress datasets.

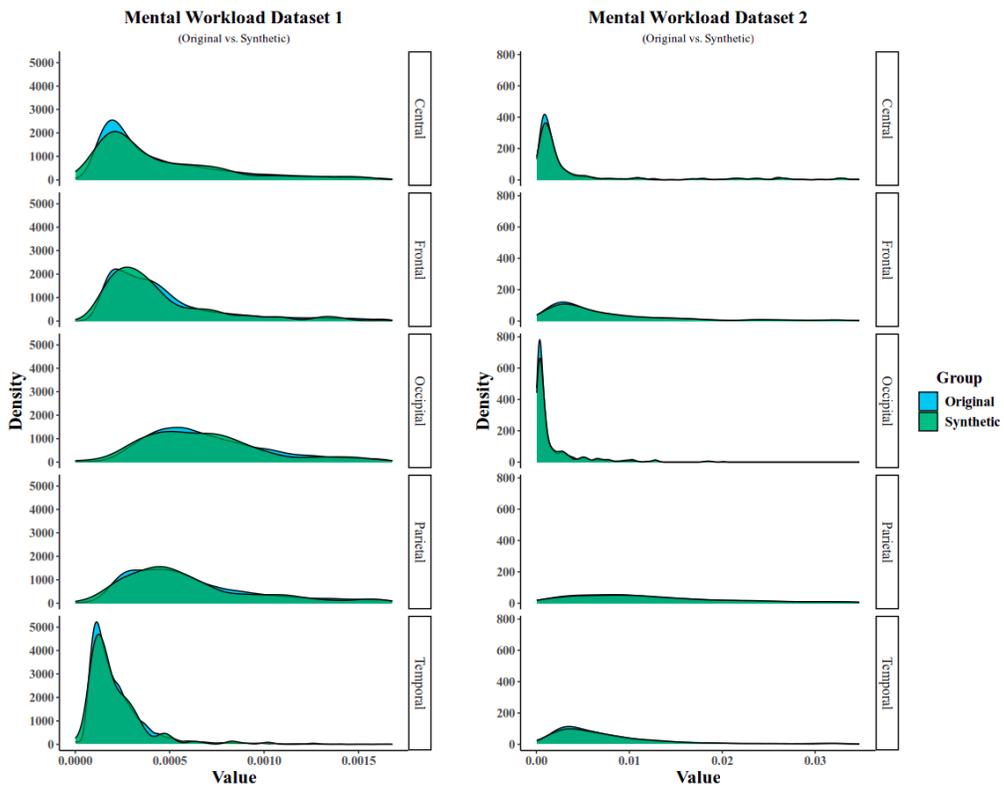


Figure 4: Distribution analysis of Workload datasets 1 and 2.

The Spearman correlation analysis demonstrates consistent coefficients between the original and synthetic data across all four datasets. Notably, the stress labels in the Stress dataset are well-preserved in the synthetic data, as illustrated in Figure 5. Figure 6 highlights the Spearman correlation between the original and synthetic versions of the SAM dataset, revealing minimal signal degradation for the arithmetic task, particularly in gamma frontal and alpha central features.

The acute stress measure, predicted using the XGBoost model trained on the original SAM dataset, is effectively maintained in the synthetic data. Additionally, the relaxation state shows a significant correlation with the stress state, underscoring robust stress prediction performance. For the Stroop test, correlations are observed primarily in frontal and temporal regions, while the arithmetic task shows correlations in central and parietal regions, with weaker associations in the occipital region. Stress-related signals are evident across frontal regions, delta central, parietal, and occipital regions, as well as theta parietal and occipital regions.

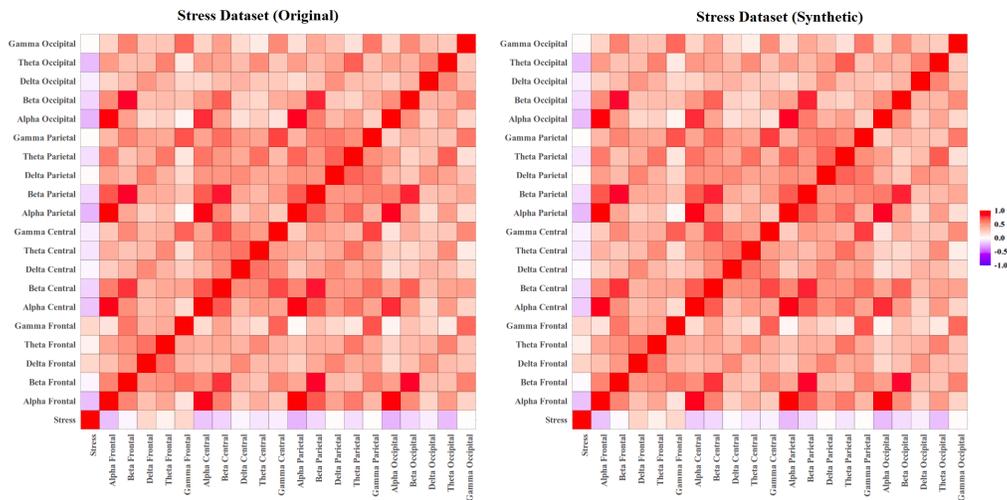


Figure 5: Spearman correlation of original Stress dataset (left) and synthetic dataset (right).

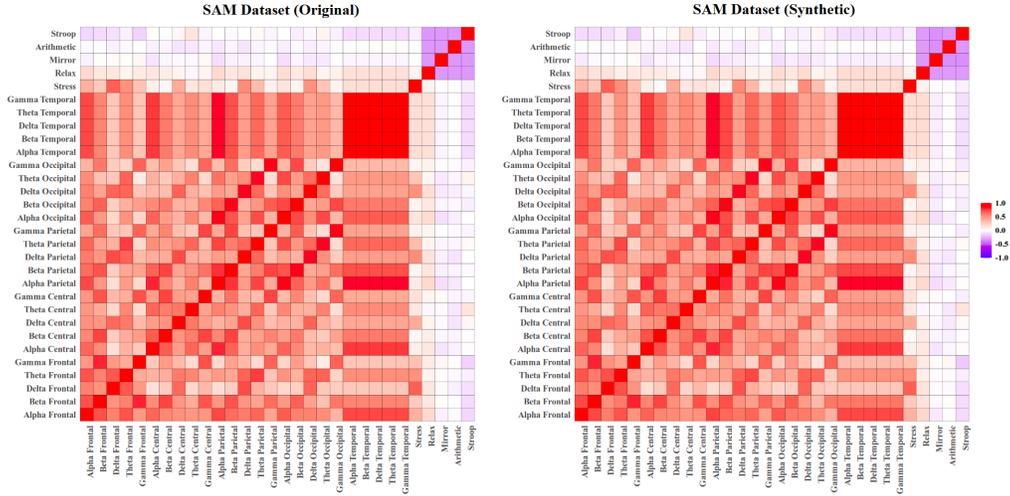


Figure 6: Spearman correlation of original SAM dataset (left) and synthetic dataset (right).

Workload Datasets 1 and 2 exhibit stronger correlations in the synthesized data compared to the original data (Figures 7 and 8). Key biomarkers present in the original datasets, such as heart rate (HR) and heart rate variability (HRV), are well-preserved in the synthetic data, alongside the stress measures predicted using the XGBoost model. Notably, Figures 7 and 8 reveal significant differences in correlation patterns between Workload Datasets 1 and 2, emphasizing the influence of varying experimental tests and protocols on EEG data.

In Workload Dataset 1, stress levels are generally low. However, when present, stress-related activity is observed in frontal, central, and occipital regions, with no notable activity in the temporal region. Specific frequency bands show distinct patterns, with beta and theta activity in the occipital region, gamma activity peaking in the central region, and theta activity in the frontal region. High-task difficulty is associated with the strongest correlations in frontal and occipital regions, while low-task difficulty shows similar patterns but in the opposite direction. Medium-task difficulty predominantly correlates with activity in temporal and central regions. These observations were preserved in the synthetically generated samples.

Workload Dataset 2, in contrast, evokes a markedly higher stress response. Stress-related activity is distributed across frontal, central, and occipital regions, with high-, medium-, and low-task difficulty levels showing widespread brain activity. The strongest correlations are observed in frontal, central, parietal, and occipital regions, with slightly lower correlations in the temporal region, indicating whole-brain involvement. Correlation with stress is most pronounced during high-task difficulty, with lower correlations for medium and low-task difficulty, aligning with the expected cognitive demands. Temporal regions show the highest activity overall, reflecting the significant temporal demands of the task. These observations were again preserved in the synthetically generated samples.

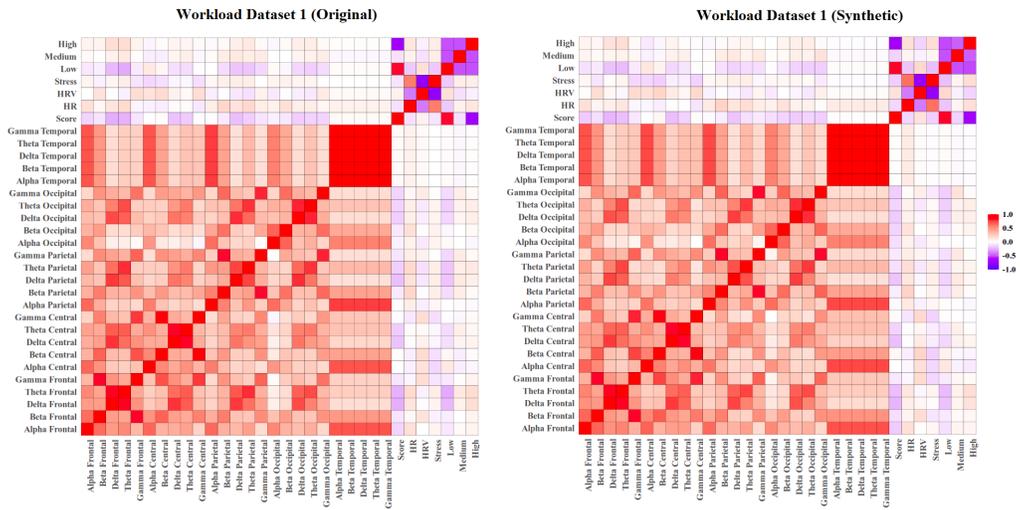


Figure 7: Spearman correlation of original Workload 1 dataset (left) and synthetic dataset (right).

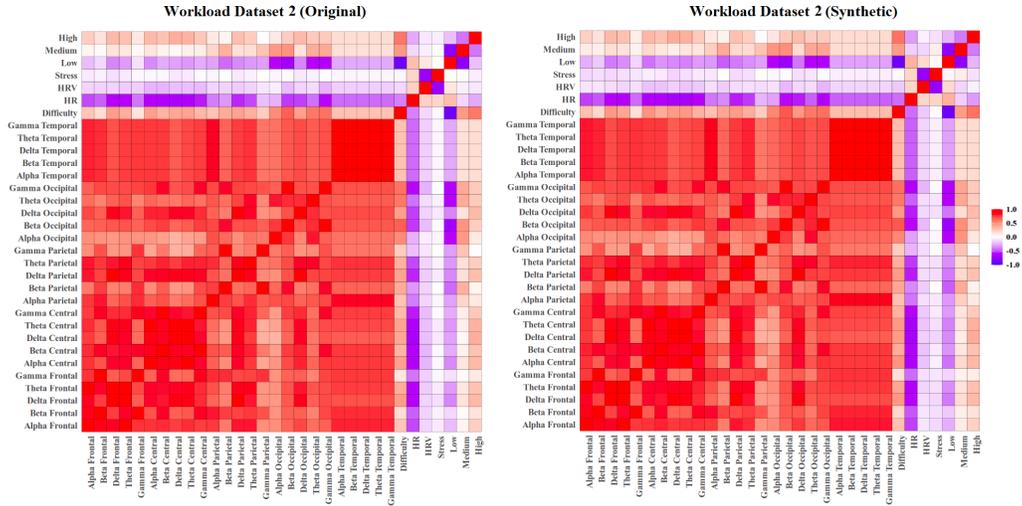


Figure 8: Spearman correlation of original Workload 2 dataset (left) and synthetic dataset (right).

PERMANOVA analyses were conducted on the SAM and Stress datasets, revealing no statistically significant differences between the original and synthetic data ($p=0.598$ and $p=0.556$, respectively).

Additionally, four RF classification models were trained on the original EEG data to predict stress in the synthetic counterparts, and vice versa. This approach serves as a robust validation technique as a well-trained machine learning model should be able to differentiate between two datasets if they contain distinct statistical properties or structural differences. If the model struggles to distinguish between original and synthetic data, it suggests that the synthetic data effectively replicates the key features of the real dataset.

In this study, the RF models exhibited near-random classification performance, with error rates of 47.62% for the SAM dataset and 52.04% for the Stress dataset values close to the 50% mark expected for indistinguishable distributions. These results indicate a high degree of similarity between the original and synthetic EEG data, reinforcing the validity of the synthetic samples for downstream machine learning applications. This approach provides an empirical, performance-based evaluation, complementing traditional statistical comparisons such as distribution analysis by offering strong evi-

dence that the synthetic EEG data retains meaningful patterns and variability present in the original data.

Figures 9 and 10 show the distribution analysis when comparing the original EEG During Mental Arithmetic Tasks (Stress) data [34] to its synthetic counterpart generated using a GAN (Figure 9) and VAE (Figure 10). The GAN consisted of a feedforward neural network with a 16-dimensional latent input mapped through a 64-unit hidden layer with ReLU activation, followed by a linear transformation to the 5-dimensional feature space (alpha, beta, delta, gamma, theta) using a Sigmoid output layer. The discriminator mirrored this structure, accepting the 5-dimensional input, processing it through a 64-unit hidden layer with ReLU, and outputting a scalar probability via a Sigmoid activation to distinguish real from synthetic samples. This model was trained over 50 epochs using the Binary Cross-Entropy loss and the Adam optimizer with a learning rate of 0.001 for both networks. A batch size of 32 was employed during training. For the VAE, the encoder was mapped to the 5-dimensional input into a 64-unit hidden layer with ReLU activation, followed by two parallel linear layers that predict the latent mean and log-variance vectors, each of dimensionality 16. The decoder layer then reconstructs the input via a mirrored structure, projecting the latent vector through a 64-unit hidden layer, and ultimately to the original input space using a Sigmoid activation function to ensure output values remain within the $[0, 1]$ range. This model was similarly trained for 50 epochs using the Adam optimizer with a learning rate of 0.001 and a batch size of 32.

The resulting distributions of both the GAN and VAE show substantial variation, while the RF model validation (refer section 3.4) managed to perform near-perfect classification separation with accuracy scores of 99% and 97%, respectively. These results further highlight the importance of validating the synthetic data using a multi-step process as described in section 3.4.

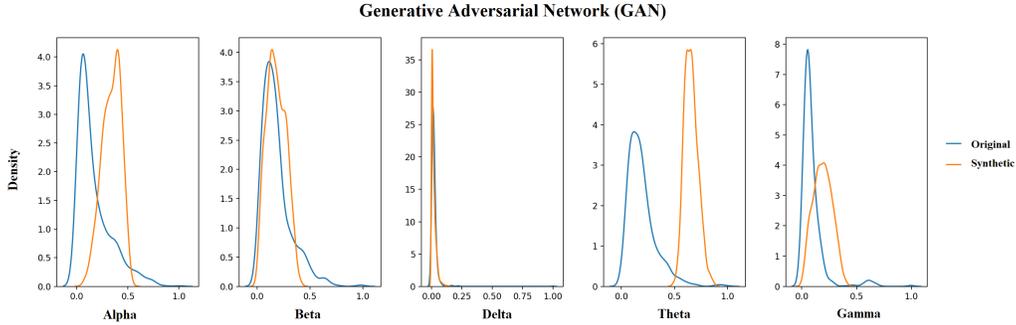


Figure 9: Distribution analysis of original vs. synthetic data generated using a GAN.

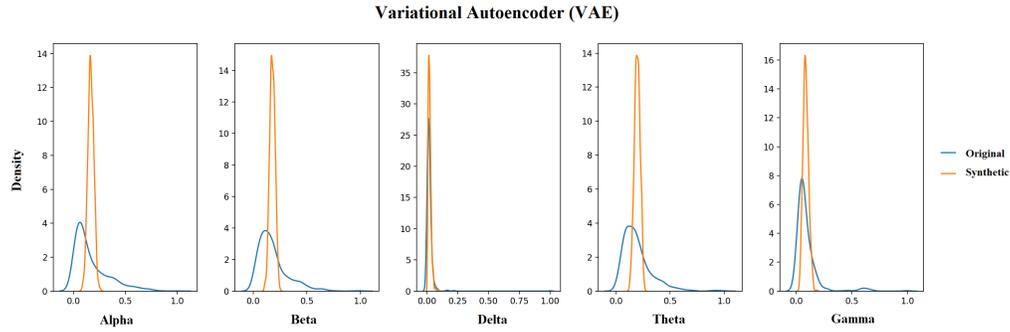


Figure 10: Distribution analysis of original vs. synthetic data generated using a VAE.

5. Discussion

The primary objective of this study was to develop and validate a statistical approach capable of producing high-quality synthetic EEG data. To achieve this, we utilized four original datasets that incorporated additional biomarkers, emotion scores, and stress levels recorded during complex task performance. These diverse datasets provided a robust foundation for experimentation and model development. Our synthesis approach employed a hybrid methodology [17] combining random sampling from the original datasets with Spearman correlation-based filtering to ensure the desired EEG duration and feature consistency.

Validation of the generated synthetic EEG data was conducted using a

comprehensive 5-step quality evaluation framework. This framework encompassed distribution analysis, statistical testing, and the application of machine learning classifiers to rigorously assess the fidelity of the synthetic data in representing the original datasets. The results demonstrated that the synthetic EEG data preserved the key characteristics and variability of the original data along with non-EEG biomarkers, underlying emotion and acute stress response, making it suitable for downstream applications such as research, algorithm development, and training of machine learning models.

A notable advantage of our method is its simplicity and efficiency, requiring significantly lower computational resources compared to more complex approaches such as GANs or diffusion models. This efficiency is particularly advantageous in scenarios where computational resources are limited or rapid data generation is required. Furthermore, the cost-effectiveness of synthetic EEG data is underscored by the high expense of acquiring real EEG recordings, which can exceed \$800 per session [11]. By utilizing synthetic data, researchers can substantially reduce research costs while maintaining access to diverse and representative datasets.

In addition to cost savings, the use of synthetic EEG data offers significant benefits in terms of patient privacy. Real EEG data is inherently sensitive, and its use in research and machine learning carries potential risks related to data security and ethical concerns. Synthetic data mitigates these risks by eliminating direct links to individual patients, thereby enhancing privacy while enabling wide-ranging applications. The proposed method further highlights the potential of hybrid generative approaches to address the growing demand for accessible, high-quality EEG data. By advancing synthetic data methodologies, this research contributes to the broader goal of enabling ethical, cost-effective, and scalable solutions for neuroscience and related disciplines.

6. Study Limitations

The study’s findings are limited by the use of a small number of EEG datasets, each with relatively modest sample sizes and a focus primarily on mental health and emotional states. While synthetic EEG data provides a viable alternative in scenarios with resource constraints, its current state does not fully replicate the complexity and authenticity of real EEG data, and

remains insufficient as a complete substitute for original EEG recordings in research and clinical applications. Future work could explore the integration of additional physiological and contextual data to further enhance the utility and realism of synthetic datasets. Additionally, extending the validation framework to include domain-specific performance metrics and real-world applications, could provide deeper insights into the applicability of synthetic EEG data across diverse fields.

7. Conclusion

This study introduces a scalable and cost-efficient method for generating synthetic EEG data, addressing critical challenges such as the high costs of traditional EEG acquisition and the limited availability of open-access datasets. Synthetic EEG data offers a transformative opportunity to overcome these barriers, enabling the development of robust and privacy-preserving datasets that facilitate the training of machine learning models for healthcare applications. By employing correlation analysis and random sampling, the proposed approach produces synthetic datasets that closely replicate the statistical and structural properties of real-world EEG data. The inability of machine learning models to differentiate synthetic samples from the original, underscores the high fidelity of the generated data. To encourage further exploration and application of this method, the complete source code used in this study is publicly available on GitHub at <https://github.com/xalentis/SyntheticEEG>.

References

- [1] S. M. Park, B. Jeong, D. Y. Oh, C.-H. Choi, H. Y. Jung, J.-Y. Lee, D. Lee, J.-S. Choi, Identification of major psychiatric disorders from resting-state electroencephalography using a machine learning approach, *Frontiers in Psychiatry* 12 (Aug. 2021). doi:10.3389/fpsy.2021.707581.
- [2] J. J. Newson, T. C. Thiagarajan, Eeg frequency bands in psychiatric disorders: A review of resting state studies, *Frontiers in Human Neuroscience* 12 (Jan. 2019). doi:10.3389/fnhum.2018.00521.
- [3] K. Rasheed, J. Qadir, T. J. OBrien, L. Kuhlmann, A. Razi, A generative model to synthesize eeg data for epileptic seizure prediction,

IEEE Transactions on Neural Systems and Rehabilitation Engineering
29 (2021) 2322–2332. doi:10.1109/tnsre.2021.3125023.

- [4] S. Dash, D. K. Dash, R. K. Tripathy, R. B. Pachori, Timefrequency domain machine learning for detection of epilepsy using wearable eeg sensor signals recorded during physical activities, *Biomedical Signal Processing and Control* 100 (2025) 107041. doi:10.1016/j.bspc.2024.107041.
- [5] Y. Li, Y. Zhao, Y. Chen, M. Meng, Z. Ren, Z. Zhao, N. Wang, T. Zhao, B. Cui, M. Li, J. Liu, Q. Wang, J. Han, B. Wang, X. Han, Effects of anti-seizure medications on resting-state functional networks in juvenile myoclonic epilepsy: An eeg microstate analysis, *Seizure: European Journal of Epilepsy* 124 (2025) 48–56. doi:10.1016/j.seizure.2024.12.004.
- [6] F. P. Carrle, Y. Hollenbenders, A. Reichenbach, Generation of synthetic eeg data for training algorithms supporting the diagnosis of major depressive disorder, *Frontiers in Neuroscience* 17 (Oct. 2023). doi:10.3389/fnins.2023.1219133.
- [7] Y. Xi, Y. Chen, T. Meng, Z. Lan, L. Zhang, Depression detection based on the temporal-spatial-frequency feature fusion of eeg, *Biomedical Signal Processing and Control* 100 (2025) 106930. doi:10.1016/j.bspc.2024.106930.
- [8] K. Boby, S. Veerasingam, Depression diagnosis: Eeg-based cognitive biomarkers and machine learning, *Behavioural Brain Research* 478 (2025) 115325. doi:10.1016/j.bbr.2024.115325.
- [9] J. Ruiz de Miras, A. J. Ibez-Molina, M. F. Soriano, S. Iglesias-Parro, Schizophrenia classification using machine learning on resting state EEG signal, *Biomedical Signal Processing and Control* 79 (2023) 104233. doi:10.1016/j.bspc.2022.104233.
URL <https://www.sciencedirect.com/science/article/pii/S1746809422006875>
- [10] C. A. Ellis, A. Sattiraju, R. Miller, V. Calhoun, Examining reproducibility of eeg schizophrenia biomarkers across explainable machine learning models (Aug. 2022). doi:10.1101/2022.08.16.504159.

- [11] J. P. Ney, M. R. Nuwer, L. J. Hirsch, M. Burdelle, K. Trice, J. Parvizi, The cost of after-hour electroencephalography, *Neurology Clinical Practice* 14 (2) (Apr. 2024). doi:10.1212/cpj.0000000000200264.
- [12] M. A. Boudewyn, M. A. Erickson, K. Winsler, J. D. Ragland, A. Yonelinas, M. Frank, S. M. Silverstein, J. Gold, A. W. MacDonald, C. S. Carter, D. M. Barch, S. J. Luck, Managing eeg studies: How to prepare and what to do once data collection has begun, *Psychophysiology* 60 (11) (Jun. 2023). doi:10.1111/psyp.14365.
- [13] A. Arora, S. K. Wagner, R. Carpenter, R. Jena, P. A. Keane, The urgent need to accelerate synthetic data privacy frameworks for medical research, *The Lancet Digital Health* (Nov. 2024). doi:10.1016/s2589-7500(24)00196-1.
- [14] A. Kamrud, B. Borghetti, C. Schubert Kabban, The effects of individual differences, non-stationarity, and the importance of data partitioning decisions for training and testing of eeg cross-participant models, *Sensors* 21 (9) (2021) 3225. doi:10.3390/s21093225.
- [15] G. Vos, M. Ebrahimpour, L. van Eijk, Z. Sarnyai, M. R. Azghadi, Stress monitoring using low-cost electroencephalogram devices: A systematic literature review (2024). doi:10.48550/ARXIV.2403.05577.
- [16] M. Rujas, R. Martn Gmez del Moral Herranz, G. Fico, B. Merino-Barbancho, Synthetic data generation in healthcare: A scoping review of reviews on domains, motivations, and future applications, *International Journal of Medical Informatics* 195 (2025) 105763. doi:10.1016/j.ijmedinf.2024.105763.
- [17] H. Murtaza, M. Ahmed, N. F. Khan, G. Murtaza, S. Zafar, A. Bano, Synthetic data generation: State of the art in health care domain, *Computer Science Review* 48 (2023) 100546. doi:10.1016/j.cosrev.2023.100546.
- [18] J. Pantanowitz, C. D. Manko, L. Pantanowitz, H. H. Rashidi, Synthetic data and its utility in pathology and laboratory medicine, *Laboratory Investigation* 104 (8) (2024) 102095. doi:10.1016/j.labinv.2024.102095.

- [19] L. Juwara, A. El-Hussuna, K. El Emam, An evaluation of synthetic data augmentation for mitigating covariate bias in health data, *Patterns* 5 (4) (2024) 100946. doi:10.1016/j.patter.2024.100946.
- [20] V. C. Pezoulas, D. I. Zaridis, E. Mylona, C. Androutsos, K. Apostolidis, N. S. Tachos, D. I. Fotiadis, Synthetic data generation methods in healthcare: A review on open-source tools and methods, *Computational and Structural Biotechnology Journal* 23 (2024) 2892–2910. doi:10.1016/j.csbj.2024.07.005.
- [21] D. B. Rubin, Statistical disclosure limitation, *Journal of official Statistics* 9 (2) (1993) 461–468.
- [22] R. J. Little, Statistical analysis of masked data, *Journal of Official statistics* 9 (2) (1993) 407.
- [23] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: *Advances in Neural Information Processing Systems*, Vol. 27, 2014, p. 1.
URL <https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf>
- [24] G. Vos, K. Trinh, Z. Sarnyai, M. Rahimi Azghadi, Ensemble machine learning model trained on a new synthesized dataset generalizes well for stress prediction using wearable devices, *Journal of Biomedical Informatics* 148 (2023) 104556. doi:10.1016/j.jbi.2023.104556.
- [25] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, *arXiv preprint arXiv:2112.10752* (2021).
URL <https://arxiv.org/abs/2112.10752>
- [26] B. Draghi, Z. Wang, P. Myles, A. Tucker, Identifying and handling data bias within primary healthcare data using synthetic data generators, *Heliyon* 10 (2) (2024) e24164. doi:10.1016/j.heliyon.2024.e24164.
- [27] A. Hernandez-Matamoros, H. Fujita, H. Perez-Meana, A novel approach to create synthetic biomedical signals using birnn, *Information Sciences* 541 (2020) 218–241. doi:10.1016/j.ins.2020.06.019.

- [28] G. Siddhad, M. Iwamura, P. P. Roy, Enhancing eeg signal-based emotion recognition with synthetic data: Diffusion model approach (2024). doi:10.48550/ARXIV.2401.16878.
- [29] B. Khosravi, F. Li, T. Dapamede, P. Rouzrokh, C. U. Gamble, H. M. Trivedi, C. C. Wyles, A. B. Selligren, S. Purkayastha, B. J. Erickson, J. W. Gichoya, Synthetically enhanced: unveiling synthetic datas potential in medical imaging research, *eBioMedicine* 104 (2024) 105174. doi:10.1016/j.ebiom.2024.105174.
- [30] E.-J. van Kesteren, To democratize research with sensitive data, we should make synthetic data more accessible, *Patterns* 5 (9) (2024) 101049. doi:10.1016/j.patter.2024.101049.
- [31] D. Rankin, M. Black, R. Bond, J. Wallace, M. Mulvenna, G. Epelde, Reliability of supervised machine learning using synthetic data in health care: Model to preserve privacy for data sharing, *JMIR Medical Informatics* 8 (7) (2020) e18910. doi:10.2196/18910.
- [32] V. B. Vallevik, A. Babic, S. E. Marshall, S. Elvatun, H. M. Brgger, S. Alagaratnam, B. Edwin, N. R. Veeraragavan, A. K. Befring, J. F. Nygrd, Can i trust my fake data a comprehensive quality assessment framework for synthetic tabular data in healthcare, *International Journal of Medical Informatics* 185 (2024) 105413. doi:10.1016/j.ijmedinf.2024.105413.
- [33] O. Ozdenizci, D. Erdogmus, On the use of generative deep neural networks to synthesize artificial multichannel eeg signals, in: 2021 10th International IEEE/EMBS Conference on Neural Engineering (NER), IEEE, 2021, pp. 427–430. doi:10.1109/ner49283.2021.9441381.
- [34] I. Zyma, S. Tukaev, I. Seleznev, Eeg during mental arithmetic tasks (2018). doi:10.13026/C2JQ1P.
- [35] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system (2016). doi:10.1145/2939672.2939785.
URL <https://doi.org/10.1145/2939672.2939785>
- [36] R. Ghosh, N. Deb, K. Sengupta, A. Phukan, N. Choudhury, S. Kashyap, S. Phadikar, R. Saha, P. Das, N. Sinha, P. Dutta, Sam 40: Dataset of

- 40 subject eeg recordings to monitor the induced-stress while performing stroop color-word test, arithmetic task, and mirror image recognition task, *Data in Brief* 40 (2022) 107772. doi:10.1016/j.dib.2021.107772.
- [37] J. Yauri, P. Folch, D. lvarez, D. Gil Resina, A. Hernndez-Sabat, Dataset to predict mental workload based on physiological data (2022). doi:10.5565/ddd.uab.cat/259591.
- [38] W. K. Kirchner, Age differences in short-term retention of rapidly changing information., *Journal of Experimental Psychology* 55 (4) (1958) 352–358. doi:10.1037/h0043688.
- [39] A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen, M. S. Hämäläinen, MEG and EEG data analysis with MNE-Python, *Frontiers in Neuroscience* 7 (267) (2013) 1–13. doi:10.3389/fnins.2013.00267.