ADHD Subtype Diagnosis for Multiclass Imbalanced Augmented EEG Data

Vandana Joshi[1,2*] and Nirali Nanavati[2]
[1] Gujarat Technological University, Ahmedabad, Gujarat, India
[2] Sarvajanik College of Engineering & Technology, Surat, Gujarat, India
e-mail: vandana.joshi@scet.ac.in

Abstract. Attention-Deficit/Hyperactivity Disorder (ADHD) is a neurodevelopmental condition that impacts an individual's emotional state, actions, and ability to acquire knowledge. The diagnosis is challenging due to the different characteristics of its distinct subtypes: inattention, hyperactivity, and impulsivity. ADHD has been diagnosed using a range of techniques, such as fMRI and behavioural evaluations. Electroencephalography (EEG) has garnered interest due to its capacity to uncover atypical brain activity patterns. Nevertheless, using EEG data in ADHD research poses difficulties due to the limited annotated datasets and substantial inter-individual variations. To address these difficulties, we have investigated various approaches to improve EEG data: Synthetic Minority Over-sampling Technique (SMOTE) variants, Non-deterministic Conditional Tabular Generative Adversarial Network (ND-CTGAN), and EEG-GAN variants. SMOTE variants have generated synthetic samples to address class imbalances, enhancing classification performance. ND-CTGAN has generated synthetic EEG data of superior quality, preserved inter-channel correlations and spectral properties, and improved the classification model's accuracy, recall, and precision. Statistical analyses have verified that synthetic data closely reflects the distributions and connections seen in real data. This similarity allows for strong comparisons and improves the ability of models. These findings not only enhance our understanding of ADHD but also inspire hope for future studies in the field, demonstrating the potential of advanced augmentation techniques to significantly enhance classification accuracy in EEG research.

Keywords: ADHD subtypes diagnosis; EEG data augmentation; SMOTE; CTGAN; EEG-GAN variants.

Paper #9203 received 6 Jan 2025; revised manuscript received 29 Jan 2025; accepted for publication 4 Feb 2025; published online 14 Mar 2025. doi: 10.18287/JBPE25.11.010302.

1 Introduction

Attention-Deficit/Hyperactivity Disorder (ADHD) is a prevalent neurodevelopmental condition characterized by patterns of inattention, hyperactivity, and impulsivity that interfere with functioning or development [1]. The inattentive subtype is defined by trouble maintaining focus and frequent careless mistakes. Restlessness, impulsive behaviour, and fidgeting characterize the hyperactive-impulsive subtype. The mixed subtype combines symptoms of both the hyperactive-impulsive and inattentive subtypes [1].

In recent years, various methods have been developed for the diagnosis of ADHD by using different modalities. For instance, functional magnetic resonance imaging (FMRI) [2], diagnostic and statistical manual (DSM) [3, 4], behavioural rating, i.e., child behavior checklist (CBCL), neuropsychological test with integrated visual and auditory (IVA), continuous performance test (CPT) [5]. The subtypes of ADHD present a significant challenge in clinical settings due to their pronounced behavioural manifestations and associated neurophysiological patterns. Electroencephalography (EEG) has emerged as a valuable tool for studying these neurophysiological markers, offering insights into the atypical brain activity patterns of individuals with ADHD [6−9].

Despite its potential, the effective utilization of EEG data in ADHD research faces critical challenges. One significant issue is the limited availability of labeled EEG dataset [9, 10], which impedes the development of robust machine-learning models for diagnosis and prediction. The insufficiency of data is impaired by the variable nature of EEG signals, which require substantial training data to capture the complex dynamics associated with ADHD subtypes accurately. To address this limitation, various data augmentation techniques have been proposed as a solution to artificially expand EEG datasets, thereby enhancing the diversity and representativeness of training samples [10, 11].

The limitations associated with EEG augmentation include challenges in evaluation metrics, such as Inception Score (IS) and Fréchet Inception Distance (FID), which are prone to overfitting issues. Furthermore, Wasserstein distance (WD) introduces additional complexity, while Euclidean distance (ED) may deliver results that lack precision. It is also worth noting that the adversarial training process inherent in GANs is complex and time-consuming. Additionally, EEG augmentation is extensively utilized across various applications, including Brain-Computer Interfaces (BCI), motor imagery, the study of epilepsy, and emotion recognition. However, its potential application for ADHD subtypes EEG data has not been thoroughly investigated [11].

The usefulness of EEG applications is also constrained by significant inter-subject variability. This variability is primarily attributed to physiological distinctions among individuals, which can significantly impact the efficiency of models designed for generalization across different subjects. Generalizing from one group of individuals to another is crucial for various practical uses of EEG technology, prompting significant research efforts toward developing strategies capable of managing inter-subject variability [12]. Traditional EEG argumentation methods frequently depend on EEG image data, which can result in challenges related to the curse of dimensionality. Additionally, these methods often struggle to preserve inter-channel correlations effectively. Understanding the correlation between EEG channels is valuable in ADHD analysis [13].

Another key challenge in EEG-based ADHD research is the issue of class imbalance [14, 15]. In clinical datasets, the proportion of samples from individuals with ADHD often significantly differs from those without the disorder. This imbalance leads to biased model training, where the learning algorithms are predisposed to the majority class, resulting in poor generalization and reduced diagnostic accuracy for the minority class. Addressing class imbalance is crucial to ensure that models are equally adept at identifying both positive and negative instances of ADHD, thereby improving their reliability and clinical applicability.

This manuscript explores advanced techniques for oversampling and augmenting data designed explicitly for EEG datasets to address limited data and imbalanced class distribution issues. The objective is to enhance the variety of training data and rectify class imbalances to improve the efficacy of machine learning models in identifying ADHD subtypes from EEG signals.

This novel approach contributes to the methodological progress in EEG analysis and shows potential for more precise and fair diagnostic procedures in ADHD research. Traditional classification methods often struggle to capture the intricate, multidimensional nature of ADHD, making it challenging to identify subtypes precisely. In this context, SMOTE and Generative Adversarial Networks (GANs) have emerged as potential tools for augmenting data and generating synthetic data, thereby improving the capability of machine learning models to handle diverse and sparse datasets.

We propose a Non-Deterministic Conditional Tabular GAN (ND-CTGAN) to improve ADHD subtype classification. By generating diverse and representative synthetic data, this approach enhances the accuracy and robustness of machine learning models in handling ADHD variability, showing promising results. We also compared our method with the EEG-GAN variant (TTS-WGAN-GP) [16], and it showed better results than it.

Our manuscript is organized into four main sections, the first of which is the introduction. The second section initiates the discussion by outlining the theoretical framework of augmentation methods. The third section delves into the methodology employed in conducting the study. The fourth section details the research's outcomes and findings, followed by the conclusion and future work.

2 Related Work

This section reviews various methodologies for generating synthetic EEG data, focusing on three principal approaches: Synthetic Minority Over-sampling Technique (SMOTE), Generative Adversarial Networks (GANs), and traditional signal processing techniques.

2.1 SMOTE Variants for EEG Data Augmentation

Traditional EEG datasets often suffer from imbalances and limited sample sizes, challenging achieving robust classification performance. To mitigate these issues, various Synthetic Minority Over-sampling Technique (SMOTE) variants have been developed. By interpolating between existing instances of the minority class, SMOTE creates synthetic samples that successfully increase the representation of the underrepresented and improve the robustness of machine learning models [17]. This technique can be helpful in ADHD studies where data might be skewed due to a smaller number of diagnosed cases compared to control groups. In ADHD, research has been applied SMOTE to various types of ADHD datasets, including phenotypic information [18], electronic medical records (EMR) [19], FMRI [20], and EEG data [21]. This involves generating synthetic samples to balance the dataset, enhancing the machine learning model's ability to detect ADHD accurately by leveraging behavioral and neurological features. SMOTE variants have shown promise in improving training datasets in EEG data analysis in various applications. The study [22] highlights the effectiveness of the adaptive synthetic sampling (ADASYN) SMOTE variant in handling multi-class imbalances in EEG data, making it a viable technique for improving the performance of EEG signal classification systems to detect epilepsy events in the EEG signals.

In Ref. [23] the researcher introduced an effective fusion algorithm that significantly improves seizure detection in imbalanced EEG datasets. By integrating SMOTE and TomekLink (STL) synthetic sampling methods with ensemble learning, the algorithm enhances the accuracy and reliability of seizure classification. Moreover, a novel data augmentation technique, BNNSMOTE [24], generates synthetic samples by focusing on the boundary and neighborhood properties of the minority class (seizure events). It aims to improve the distribution of synthetic samples to represent the minority class better and enhance classifier training. Likewise, the Borderline-SMOTE and Radius-SMOTE [25] algorithm demonstrates their edge in efficient, emotional identification of EEG signals [26], While Borderline-SMOTE [27], support vector machine (SVM) SMOTE, and adaptive synthetic sampling [28] were used to increase the number of samples for the class of target stimuli in P300-based brain-computer interfaces (BCIs), whereas safe-level SMOTE has enhanced classification performance for different anxiety levels from the DASPS dataset.

2.2 GAN Approaches for EEG Data Augmentation

Generative Adversarial Networks (GANs) have become a cornerstone in advanced EEG data augmentation, enabling the generation of realistic synthetic data that mirrors the complexity of real EEG signals. Variants of GAN, for instance, the Deep Convolutional GAN (DCGAN) applied on spectrogram-generated images from STFT (Short-Time Fourier Transform) of EEG, with transfer learning models, have generated high-fidelity EEG signals by minimizing divergence between real and synthetic data distributions for epileptic seizure prediction [29]. Auxiliary Classifier GAN (AC-GAN) also demonstrated that generated synthetic spikes samples improve machine learning classification tasks in the diagnosis of epilepsy [30]. Conditional Wasserstein GAN (CWGAN) enhances the performance of EEG-based emotion recognition models by conditioning the generation process on specific labels, producing more relevant synthetic data [31, 32]. Balanced Wasserstein GAN with gradient penalty (BWGAN-GP) [33] and Common Spatial GAN (CS-GAN) [34] address specific needs in EEG data, such as balancing datasets and generating spatially consistent signals, respectively, in BCI applications. Likewise, Event-Related Potentials GAN (ERP-GAN) effectively improves the stability of the model for single-trial P300 detection [35].

Time Series GAN [36], is designed to handle temporal dependencies and generate synthetic EEG sequences that capture the dynamic nature of neural activity. These GAN-based approaches significantly enrich EEG datasets, providing diverse and high-quality synthetic data that enhances model training and performance in Seizure detection [37]. Another approach in Ref. [16] the study derives a hybrid architecture, TTS-WGAN-GP, that contains a transformer-encoder-based generator and discriminator for time series (TTS) and conditional GANs. This architecture focuses on temporal features, conditional data, and stability while generating EEG synthetic data.

2.3 Other Methods for EEG Data Augmentation

Besides SMOTE and GAN-based techniques, other EEG data augmentation methods are crucial for improving the diversity and resilience of EEG training datasets. Time domain augmentations, like smooth time masking and time reversal, enhance temporal variability. Frequency domain augmentations, such as frequency shifting and band-stop filtering, can improve training data by altering spectral properties using various materials. The spatial variety of EEG data is improved via spatial domain augmentations such as channel symmetry, channel dropout, and channel shuffling [10]. A comparative research study is presented for motor imagery EEG data in Ref. [38]. These include simple approaches such as averaging, time slice recombination, and cropping, as well as more sophisticated ones such as variational autoencoder (VAE).

3 Materials and Methods

In addressing EEG data augmentation challenges, we aim to generate synthetic EEG data specifically for the Hyperactive ADHD subtype. A major issue with existing methodologies is class imbalance, which can lead to biased outcomes and hamper the accuracy of classification models. Traditional approaches often rely on EEG image data, which leads to the curse of dimensionality and fails to maintain inter-channel correlations, thus diminishing the model's performance. Moreover, image-based techniques do not adequately capture subject-wise variability despite their effectiveness, limiting their applicability across diverse subjects. By focusing on maintaining spectral characteristics and enhancing the inter-channel correlations of synthetic EEG data, we seek to improve classification accuracy while enabling a more comprehensive exploration of the feature space.

The following subsections outline this study's materials, experimental design, and methodological approaches. We employed two methods for data augmentation and classification. Firstly, we compared augmented data using various Synthetic Minority Over-sampling Technique (SMOTE) variants. Secondly, we introduced a novel approach, the non-deterministic Conditional Tabular Generative Adversarial Network (ND-CTGAN), to improve classification performance.

3.1 Dataset

In our study, we examined EEG data across different ADHD subtypes [15], including 17 subjects with inattentive ADHD, 17 with combined ADHD, and 3 with hyperactive ADHD. The EEG recordings were captured using 64 channels, initially sampled at 2048 Hz and subsequently down-sampled to 256 Hz to facilitate efficient processing. To ensure data quality, a bandpass filter was applied to remove slow drifts below 1 Hz and high-frequency noise above 20 Hz, including line noise. The recordings were made while subjects were working and responding to auditory stimuli. The tasks involved listening to human speech streams composed of /ba/, /da/, and /ga/ syllables. Each trial had a central" Target" stream with three syllables and a "Distractor" stream with five syllables to the right. This dataset is available openly; the link is in the data availability statement.

3.2 Data Augmentation and Classification with SMOTE Variants

To address the issue of class imbalance in our EEG dataset, we applied data augmentation using various Synthetic Minority Over-sampling Technique (SMOTE) variants. These variants included SMOTE, Borderline-SMOTE, ADASYN, SMOTE and Tomek, and SVM-SMOTE. Each is designed to generate synthetic samples for the hyperactive AHDD minority class. The augmented data were then subjected to classification using a range of machine learning algorithms. This comparative analysis aimed to evaluate the effectiveness of different SMOTE variants in enhancing classification performance by creating a more balanced dataset.

Fig. 1 shows the flow diagram of data augmentation and classification with SMOTE variants. After preprocessing raw EEG data, six non-linear features are extracted: approximate entropy, singular value decomposition entropy (SVD), spectral entropy, Petrosian Fractal dimension (FD), Katz FD, and Higuchi FD. In our previous work [38], these features yield promising results for EEG analysis.

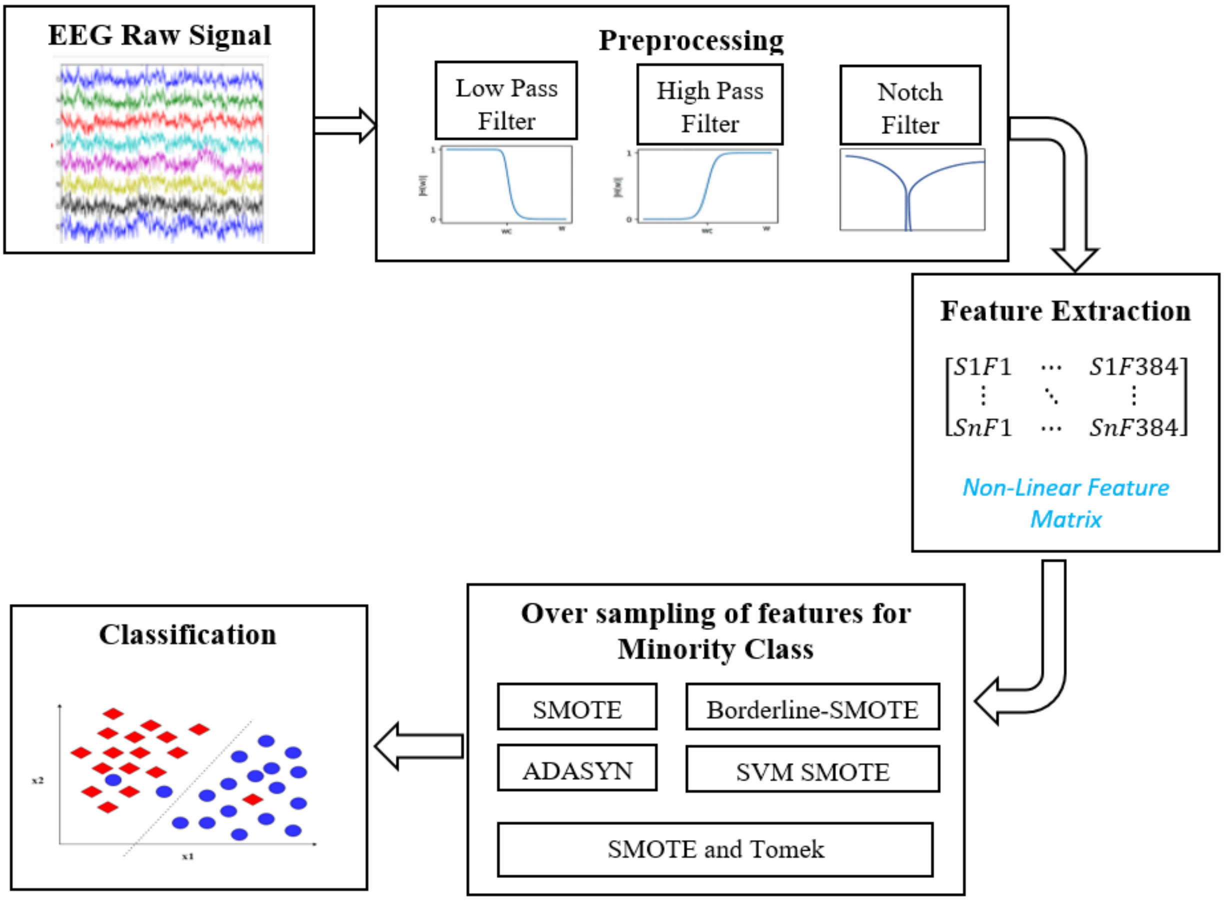
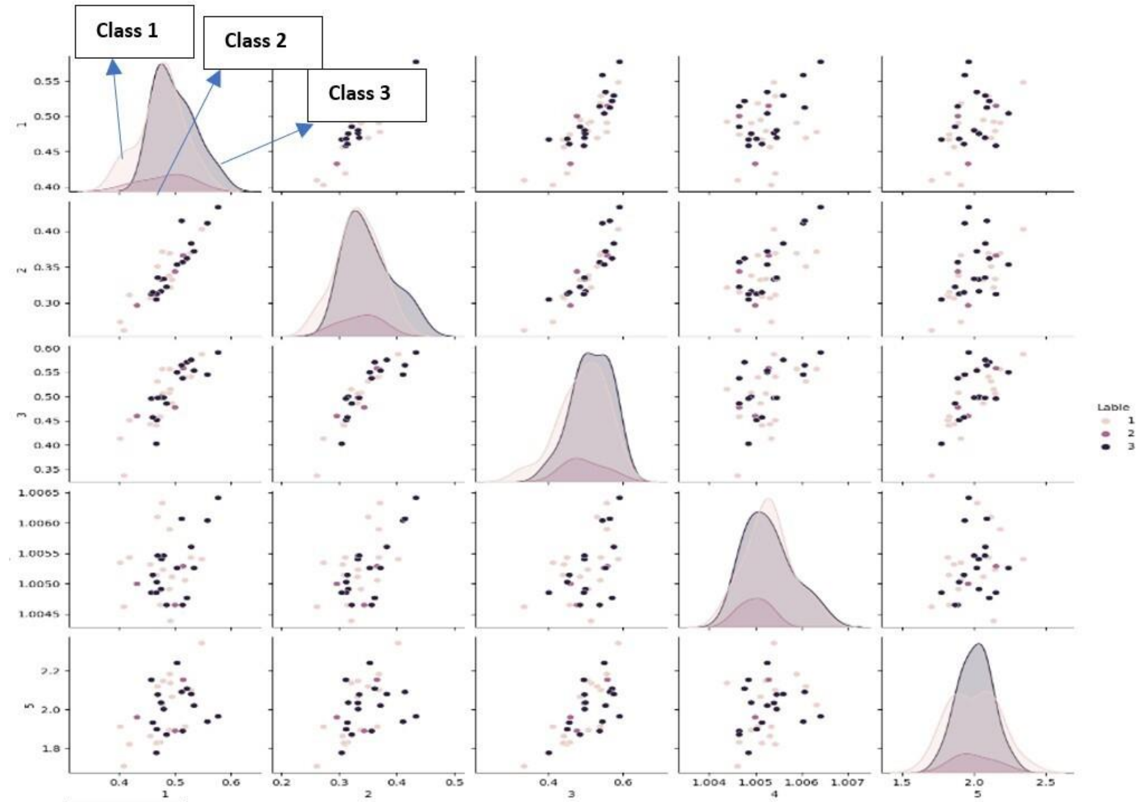
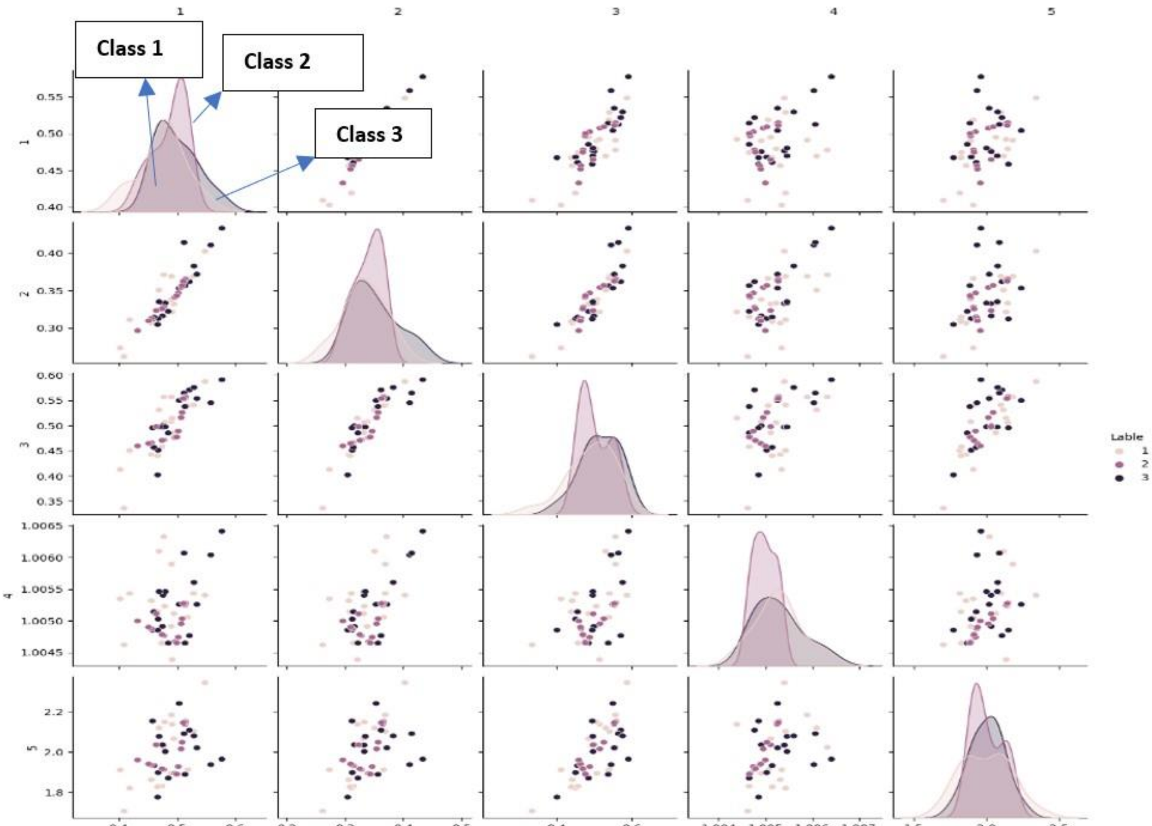


Fig. 1 Block diagram of data augmentation and classification with SMOTE variants.



(a)



(b)

Fig. 2 Pair plot for first five features of class 1, 2, and 3 (on the X-axis and Y-axis): (a) before ADASYN, (b) after ADASYN.

A total of 37 EEG samples were initially available for oversampling, including 17 from the combined ADHD type, three from the inattentive ADHD type, and three from the hyperactive ADHD type. The dataset was balanced using oversampling techniques, resulting in an equal distribution of 51 samples across all ADHD types. These samples were analyzed using various classification methods. Fig. 2 presents a pair plot for the first five features of the feature vector, both before and after applying ADASYN. This plot illustrates the oversampling of hyperactive samples and demonstrates that the synthetic data maintains a distribution similar to the real data.

Further, five SMOTE variants including standard SMOTE, Borderline, ADASYN, SMOTE SVM, and SMOTE Tomek, were proposed by us for multiclass classification. Three classes were considered for classification: Combined (Class 1), Hyperactive (Class 2), and Inattentive (Class 3). Seven different classifier models have been applied and compared with each other. Random Forest, MLP (Multilayer Perceptron), SVM (Support Vector Machine), KNN (K-Nearest Neighbours), Logistic Regression, Decision Tree (DT), and Gaussian Naive Bayes (GaussianNB), before and after the application of several data augmentation techniques. The models were evaluated on their classification accuracy on the original dataset (indicated as "before Augmentation") and after applying different augmentation methods. Table 1 illustrates the performance of all classifiers. The ten-cross validation (10-CV) method was used for training and testing. In the SVM and logistic regression method, three different classes were applied to the one-vs-rest (OVR) approach to evaluate the performance. The results indicate that data augmentation generally leads to an improvement in model accuracy. For instance, Logistic Regression achieves its highest accuracy of 0.70 when Borderline SMOTE is applied.

We found that augmented EEG data using SMOTE variants cannot trace which synthetic data points are generated from which subject data, making it challenging to directly compare real and generated data for each subject. To further improve the classification of EEG data, we introduced a nondeterministic Conditional Tabular Generative Adversarial Network (CTGAN) for data augmentation (Explained in the following section).

3.3 Data Augmentation and Classification with ND-CTGAN

Unlike traditional methods, this approach generates synthetic EEG data that captures complex, non-linear relationships and retains essential inter-channel correlations and spectral characteristics. By augmenting our dataset with this advanced technique, we aimed to enhance the model's generalization capability and robustness. In contrast, the nondeterministic CTGAN allows for the generation of data that preserves individual subject characteristics, facilitating more meaningful comparisons between real and synthetic datasets. The detailed algorithm is illustrated as Algorithm 1.

This method was applied to pre-processed EEG data of individual subjects and then used for feature extraction with the real data. The synthetic data produced by the nondeterministic CTGAN was subsequently used to extract the nonlinear features and to train various machine learning classifiers. Six nonlinear features are extracted, as stated in the 3.2 Section. Fig. 3 illustrates all the phases of the processes.

Algorithm 1 Generation of synthetic samples using CTGAN for hyperactive subject data.

Input:

- $O_{subject}$: Original EEG data for the hyperactive subject.
- $N_{desired}$: The desired number of synthetic samples

Output:

- $S_{synthetic}$: Synthetic EEG samples for the hyperactive subject.

Procedure:

a) Load the original EEG dataset $O_{subject}$ for the hyperactive subject.
b) Set the desired number of synthetic samples $N_{desired}$.
c) Utilize the CTGAN model to generate an initial set of synthetic samples $S_{initial}$ based on $D_{subject}$.
d) Compute the number of synthetic samples $N_{generated}$ in $S_{initial}$.
e) While $N_{generated} \neq N_{desired}$:
   i. apply CTGAN to generate additional synthetic samples $S_{additional}$;
   ii. Combine $S_{initial}$ with $S_{additional}$ to update the synthetic dataset $S_{initial}$;
   iii. Update $N_{generated}$ with the new count of samples in $S_{initial}$.
f) Collect synthetic samples: Once $N_{generated} = N_{desired}$, assign $S_{synthetic} = S_{initial}$.
g) Return the synthetic EEG samples $S_{synthetic}$ for the hyperactive subject.

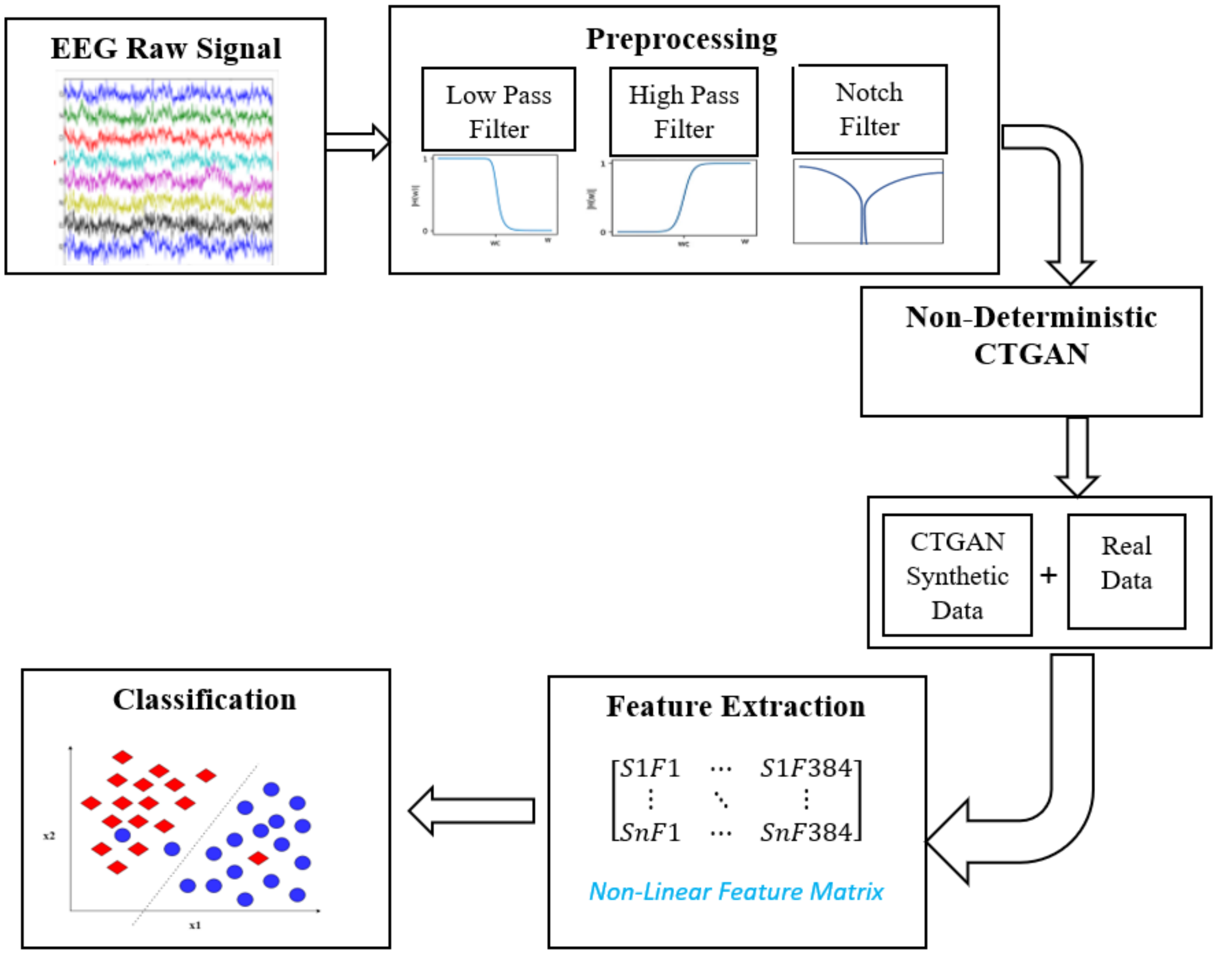


Fig. 3 Block diagram data augmentation and classification with ND-CTGAN.

CTGAN Architecture

CTGAN is a GAN-based method for modeling tabular data distribution and sampling rows. Fig. 4 illustrates the architecture. The mode-specific normalization N (0, 1) is used for a continuous variable with floating point values. CTGAN employs a variational Gaussian mixture model [39] As a unit distribution model, it estimates the number of distributions for each variable and normalizes the variable values based on the estimated distributions. During training, CTGAN uses these encoded values instead of the original values. After training, CTGAN transforms the generated data back to the original scale when generating synthetic data.

For each continuous column Ci, use the Variational Gaussian Mixture Model (VGM) to estimate the number of modes m_i and fit a Gaussian mixture. The learned Gaussian mixture is

$$PC_i(c_{i,j}) = \sum_{K=1}^{3} \mu_k N(c_{i,j}; \eta_k, \Phi_k), \qquad (1)$$

where μ_k and Φ_k are the weight and standard deviation of a mode, respectively.

The CTGAN model's training process is comprehensive. It does not depend on the local structure within individual columns; instead, it uses fully connected networks in both the generator and discriminator to capture all possible correlations between columns. With its two completely connected hidden layers, the generator employs batch normalization and the ReLU activation function. The synthetic row representation blends activation functions, generating scalar values with the hyperbolic tangent function (tanh) and discrete values with Gumbel Softmax. Using the leaky rectified linear unit (ReLU) function and applying dropout to every hidden layer, the discriminator operates within the PacGAN framework, using 10 samples in each pac to avoid mode collapse effectively. Additionally, the model is trained via the Wasserstein GAN (WGAN) loss function, which incorporates a gradient penalty.

The training process employs an Adam optimizer with a learning rate of $2 \cdot 10^4$. Researchers in Ref. [40] It has been proven that synthetic data from the CTGAN is more suitable for EEG data. We used pre-processed raw tabular data with 1,792 samples and 64 channel characteristics per subject. We used the CTGAN algorithm to generate synthetic data for hyperactive subjects. The steps in this procedure are outlined in the following algorithm.

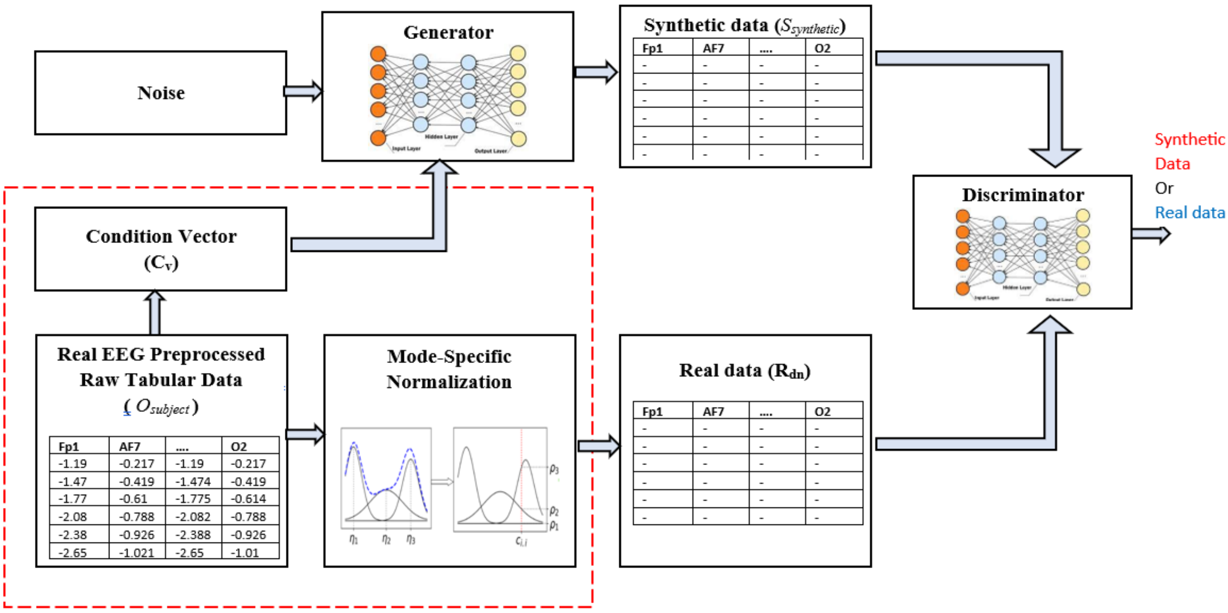


Fig. 4 The architecture of conditional tabular GAN (CTGAN).

Evaluation of Synthetic Data

Several statistical and quality metrics are used to evaluate the performance of our proposed CTGAN before using the synthetic data for classification tasks. These metrics help ensure that the generated data closely resembles the real data in terms of distribution and relationships between variables. The statistical and quality metrics comprise mean and standard deviation (STD), cumulative distribution function (CDF), quality score: Kolmogorov-Smirnov statistic (KS statistic), and correlation similarity. The CDF represents the likelihood that a random variable will have a value less than or equal to a given value. Comparing the CDFs of each characteristic in the synthetic and actual data may determine how well the distributions match throughout the whole range of values.

The KS statistic measures the maximum difference between the empirical CDFs of the real and synthetic data for each feature. It provides a non-parametric test for comparing distributions without making assumptions about their exact form. The KS statistic can be used to evaluate the similarity between actual and synthetic data distributions. A smaller KS statistic suggests a closer match between the two distributions. Correlation similarity assesses how the associations between pairs of attributes in the synthetic data look like those in the real data. It ensures that the synthetic data maintains the interconnections and associations among variables, which is essential for upholding the integrity of the fundamental structure of the data.

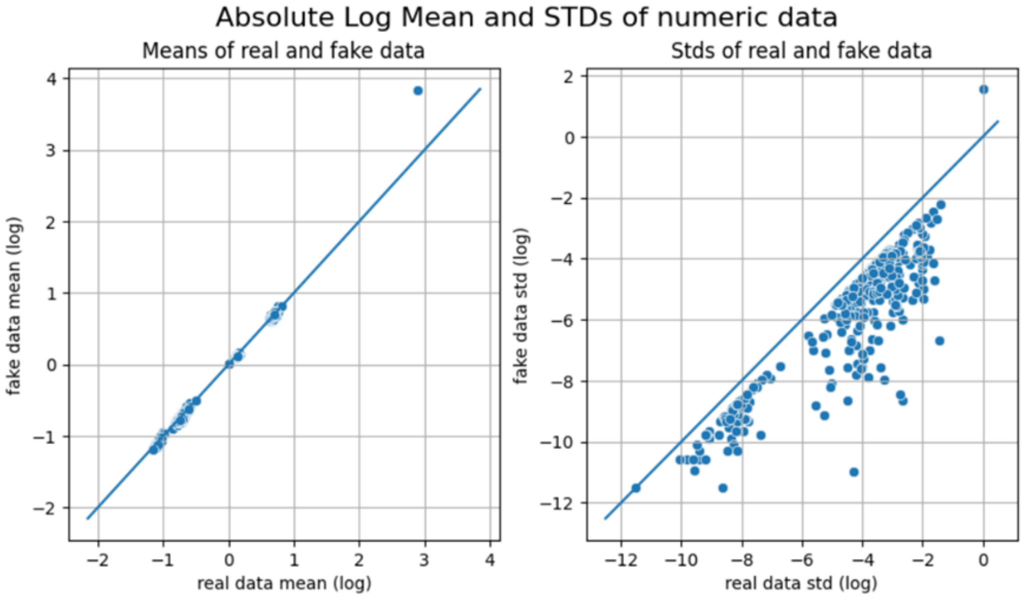
A total of 1793 × 67 samples were utilized for the KS Complement analysis. Based on the sample size, the critical values for the Kolmogorov-Smirnov test were determined as 0.169 for a significance level of $\alpha = 0.05$ and 0.203 for $\alpha = 0.01$. The observed average KS values were recorded at 0.220 (1−0.78) and 0.31 (1−0.69). Given that both 0.220 and 0.31 exceed 0.203, the results indicate statistical significance at the 5% and 1% levels.

Analysis of Synthetic Data

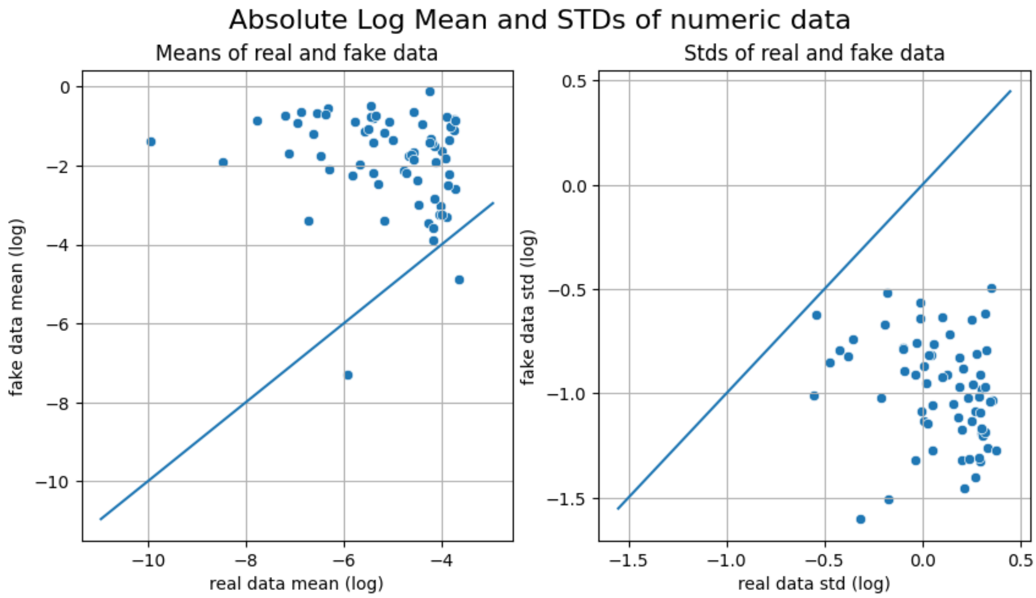
We have analyzed the EEG synthetic data of ND-CTGAN and EEG-GAN variants. The visualizations in Fig. 5 show similarities between the synthetic and actual data. The mean aligns well, and the standard deviation shows a small change for ND-CTGAN, but it shows a difference for EEG-GAN variants.

Fig. 6 shows the distribution of channels for real and synthetic data of channels like FP1, PC1, and PC4. This indicates that the distribution of synthetic data is well matched with the real for ND-CTGAN compared to EEG-GAN. Likewise, by considering the cumulative sums per channel, as in Fig. 7, most of the channel data in the synthetic data match closely with the actual data for ND-CTGAN but not for EEG-GAN.

Fig. 8 illustrates the “Column Shapes” property of the quality score for each EEG channel computed using the KS-metric. The average KSComplement values obtained were 0.78 for ND-CTGAN and 0.69 for EEG GAN. In the SDMetrics library, the KSComplement is derived by inverting the KS statistic using the formula “1 − (KS statistic)”. A higher score signifies greater data quality, with a score of 1.0 representing the optimal scenario in which the synthetic data is indistinguishable from the real data, while a score of 0.0 indicates complete dissimilarity between the two. In our analysis, a score of 0.7 approaches 1.0, suggesting a reasonable similarity between the real and synthetic data.



(a)



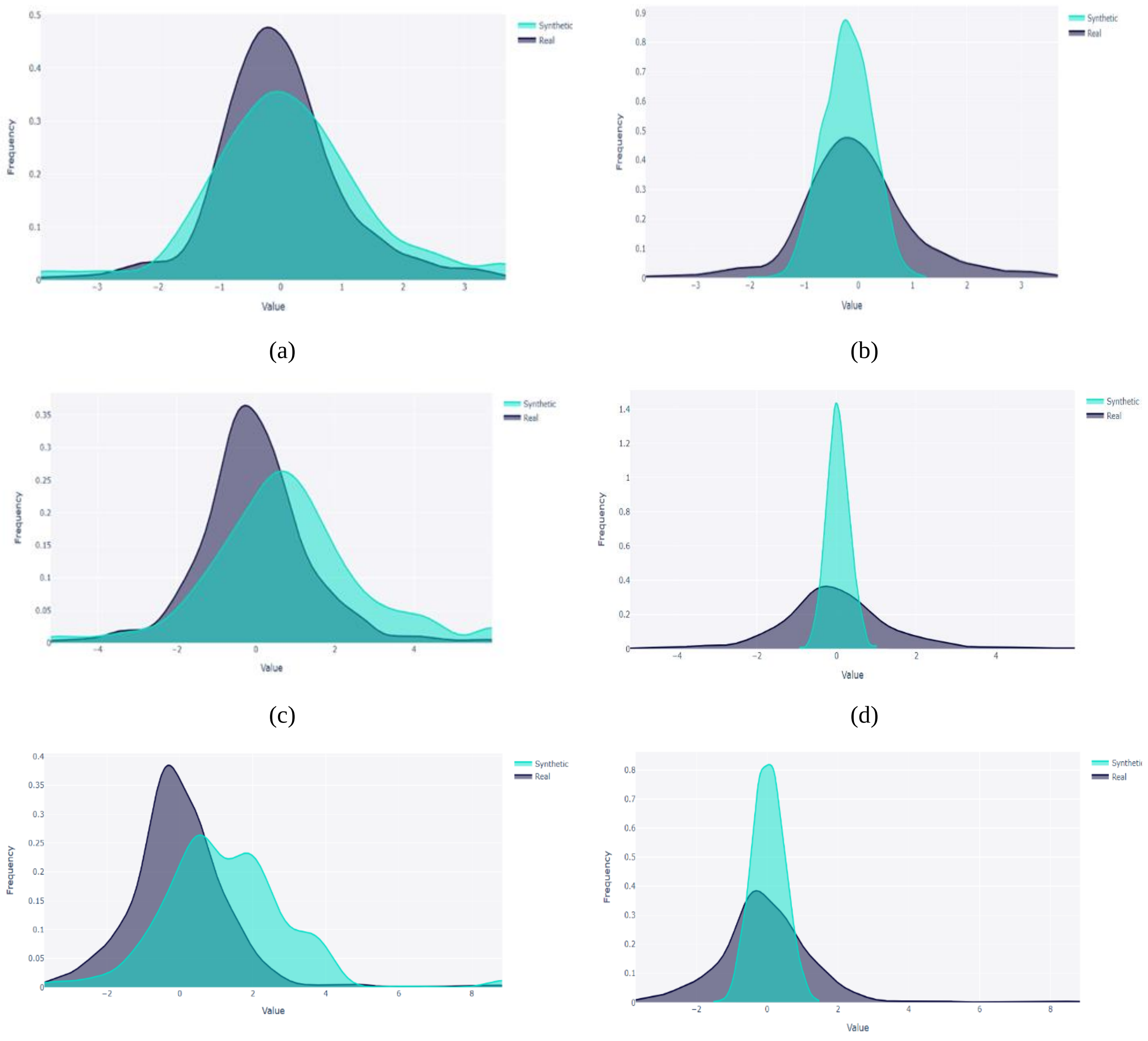
(b)

Fig. 5 Mean and STD (Real vs Synthetic) of (a) ND-CTGAN, (b) EEG-GAN.

The correlation similarity plots of samples in Fig. 9 illustrate a comparison of 64 channel properties using both real and generated data. This analysis helps us understand the relationships between channels in real and synthetic datasets. For instance, the similarity score between the FT8 and PO3 pair is 0.95. The correlation coefficient between the Fz and O1 channels is 0.38 for real data and 0.3 for synthetic data, suggesting a strong resemblance. These findings indicate that the synthetic data generated possess qualities that closely resemble real data.

Performance Evaluation of Classifiers

Following a thorough evaluation of synthetic data generated by ND-CTGAN and EEG-GAN, both generated and real datasets are employed for classification purposes. We examine seven distinct classifiers: Random Forest, Multi-Layer Perceptron (MLP), K-Nearest Neighbors (KNN), Logistic Regression, Decision Trees (DT), and Gaussian Naive Bayes (NB), which are widely recognized in ADHD classification research [41]. The classifiers are assessed using key performance metrics, including accuracy, recall, precision, and the Receiver Operating Characteristic (ROC) curve. To ensure a reliable evaluation of the models, we implemented a cross-validation strategy with 10 folds utilizing stratified K-Fold. This approach enables a robust assessment of model performance by testing it on various subsets of data, resulting in an averaged measure of effectiveness across multiple data splits.

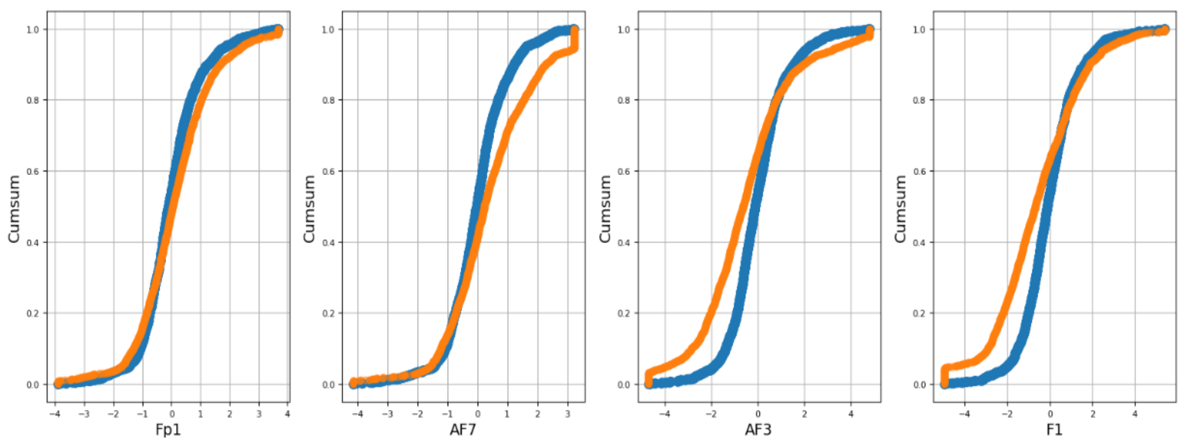


(a) (b)

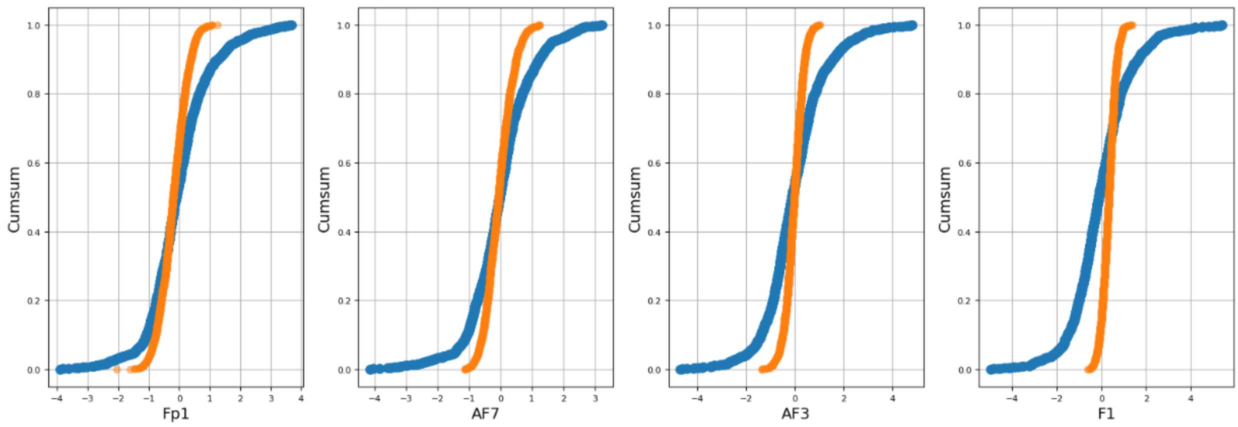
(c) (d)

(e) (f)

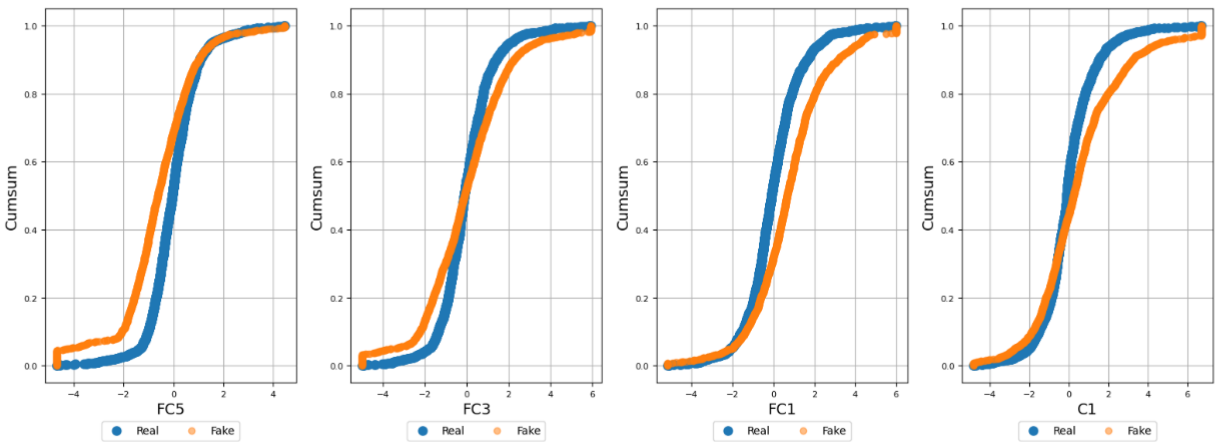
Fig. 6 Comparison of the distribution of channel data (Real vs Synthetic) between (a, c, e) ND-CTGAN and (b, d, f) EEG-GAN: FP1 channel of (a) ND-CTGAN and (b) EEG-GAN, FC1 channel of (c) ND-CTGAN and (d) EEG-GAN, PO4 channel of (e) ND-CTGAN and (f) EEG-GAN.



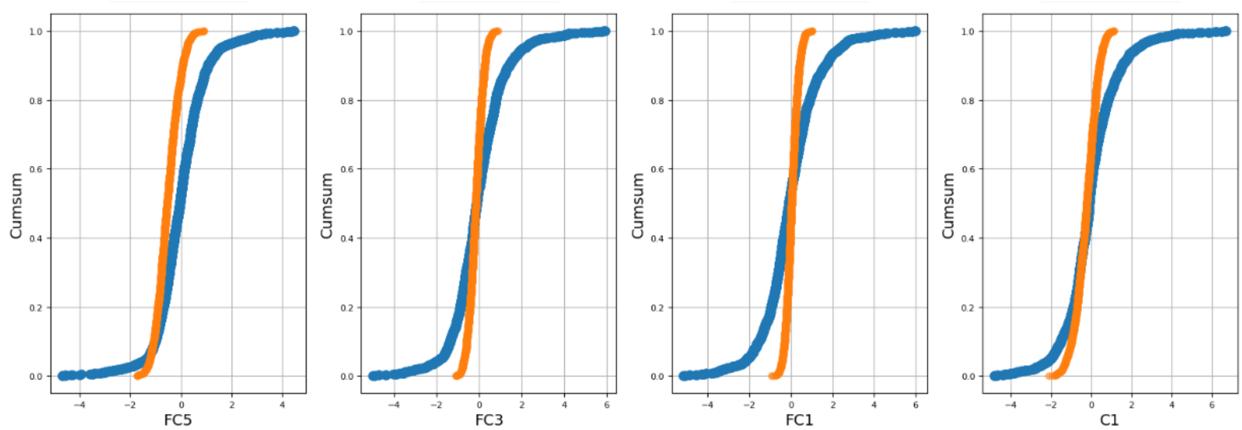
(a)



(b)



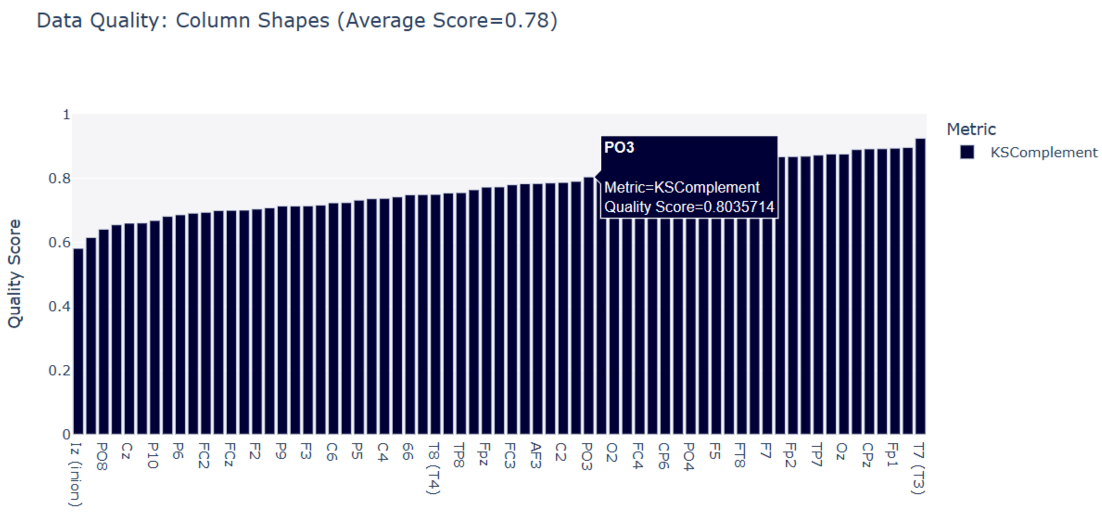
(c)



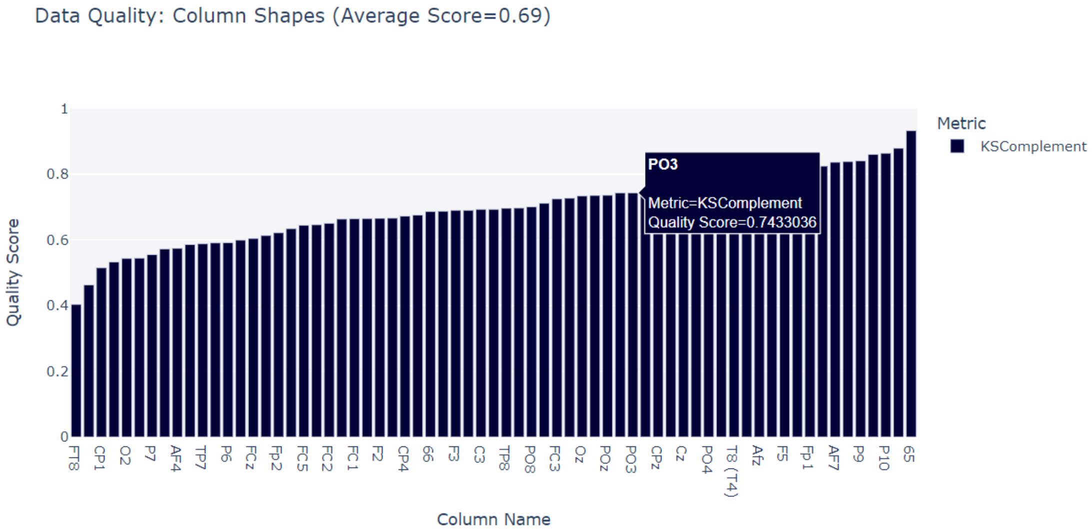
(d)

Fig. 7 Comparison of cumulative sum per feature of channel data (Real (blue) vs. Synthetic (red)) between (a, c) ND-CTGAN and (b, d) EEG-GAN.

Table 1 shows the comparative analysis of classification accuracy for the SMOTE variants and GAN methods. In SMOTE, augmentation was applied to the features matrix, whereas EEG-GAN and ND-CTGAN were applied to pre-processed data. We tested the performance of seven classifier models, and the result shows a 16% to 18% improvement in accuracy after augmentation. SMOTE borderline and ADASYN perform well among other SMOTE variants. Likewise, MLP reaches its peak accuracy of 0.69 with ND-CTGAN pre-processed data. Random Forest, SVM, KNN, and DT models also exhibit their best performance after using ND-CTGAN on pre-processed data compared to EEG-GAN.

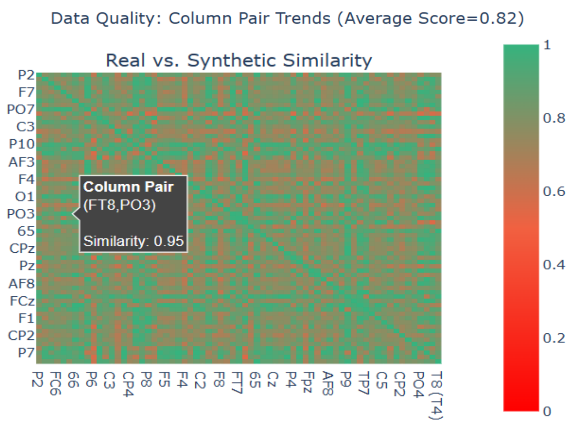


(a)

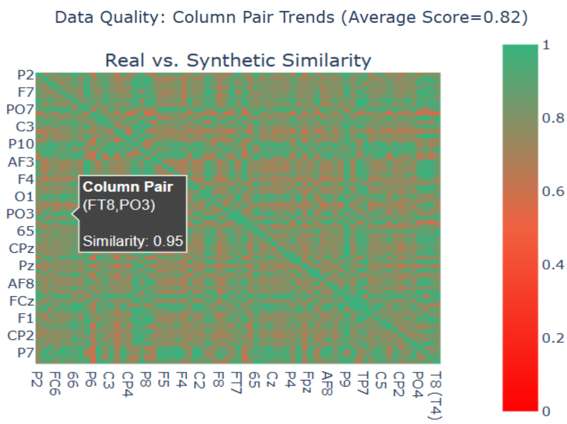


(b)

Fig. 8 Quality Score of all channels using the KSComplement metric of synthetic data: (a) ND-CTGAN, (b) EEG-GAN.



(a)



(b)

Fig. 9 Correlation similarity (column pair trends) of channel data (real vs synthetic): (a) ND-CTGAN, (b) EEG-GAN.

Table 1 Performance evaluation of SMOTE variants, EEG-GAN, and ND-CTGAN.

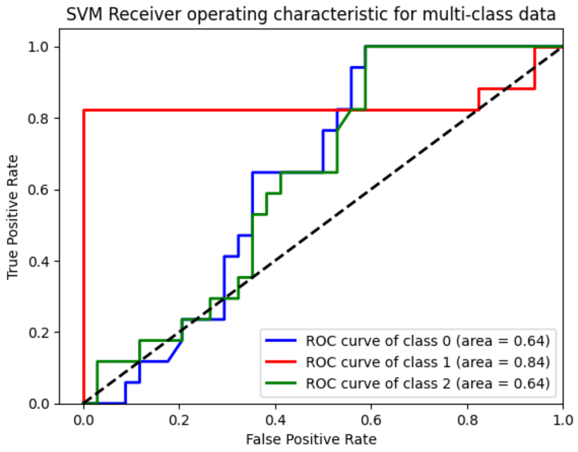
				Random Forest	MLP	SVM	KNN	Logistic	DT	Gaussian NB
before augmentation			Accuracy	0.32	0.37	0.38	0.4	0.46	0.38	0.35
			Recall	0.23	0.27	0.37	0.3	0.33	0.38	0.35
			Precision	0.16	0.23	0.24	0.3	0.31	0.27	0.35
after augmentation (augmentation applied on)	**feature matrix**	**SMOTE**	Accuracy	0.59	0.55	0.51	0.49	0.64	0.61	0.57
			Recall	0..583	0.57	0.5	0.48	0.66	0.54	0.54
			Precision	0.533	0.49	0.51	0.35	0.59	0.53	0.53
		Borderline SMOTE	Accuracy	0.58	0.63	0.59	0.51	0.7	0.64	0.57
			Recall	0.63	0.65	0.5	0.48	0.7	0.66	0.53
			Precision	0.62	0.57	0.43	0.39	0.66	0.59	0.52
		ADASYN	Accuracy	0.59	0.62	0.4	0.53	0.67	0.61	0.56
			Recall	0.55	0.64	0.4	0.53	0.66	0.63	0.53
			Precision	0.49	0.57	0.4	0.53	0.67	0.54	0.55
		SMOTE SVM	Accuracy	0.47	0.54	0.32	0.5	0.63	0.51	0.45
			Recall	0.48	0.63	0.3	0.5	0.65	0.53	0.48
			Precision	0.5	0.59	0.3	0.5	0.62	0.52	0.45
		SMOTE Tomek	Accuracy	0.57	0.61	0.51	0.47	0.57	0.59	0.61
			Recall	0.56	0.6	0.44	0.48	0.58	0.65	0.6
			Precision	0.48	0.5	0.32	0.32	0.48	0.58	0.55
	pre-processed data	**ND-CTGAN**	Accuracy	0.6	0.69	0.67	0.59	0.65	0.61	0.61
			Recall	0.58	0.6	0.66	0.58	0.65	0.55	0.61
			Precision	0.43	0.52	0.58	0.53	0.64	0.55	0.45
		TTS-WGAN-GP (EEG-GAN)	Accuracy	0.56	0.54	0.6	0.56	0.6	0.54	0.58
			Recall	0.51	0.54	0.61	0.58	0.61	0.51	0.61
			Precision	0.51	0.47	0.62	0.55	0.58	0.51	0.45

The ROC curve was generated for three classes: Combine (Class 0), Hyperactive (Class 1), and Inattentive (Class 2). Fig. 10 illustrates that the 0.83 to 0.85 AUC for the hyperactive class (class 1) shows the good performance of classifiers compared to the other two classes.

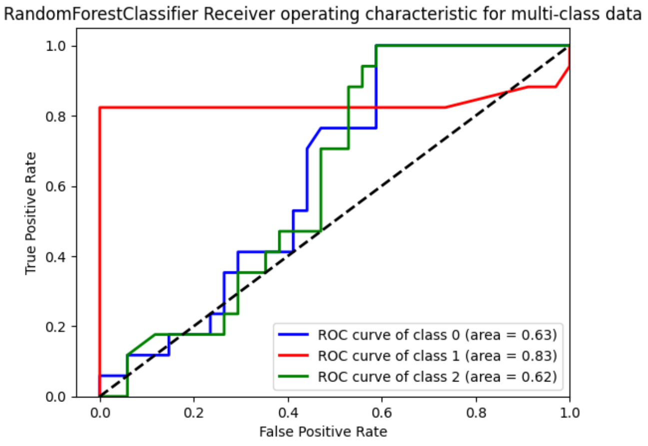
4 Discussion and Conclusion

This study focuses on the issues associated with less EEG data availability and imbalances within classes that frequently hamper ADHD diagnosis and classification accuracy. Conventional methods that depend on EEG image data have difficulties with the number of dimensions and maintaining the inter-channel correlations. To address these problems, the study focuses on preserving the spectral properties and improving the inter-channel relationships of synthetic EEG data to enhance classification accuracy, explicitly emphasizing the Hyperactive ADHD subtype.

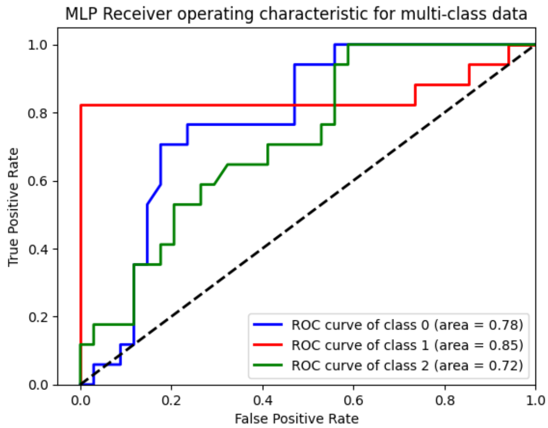
The study employed three primary techniques for data augmentation: several versions of the Synthetic Minority Over-sampling technique (SMOTE), EEG-GAN variants (TTS-WGAN-GP), and a novel method called Non-Deterministic Conditional Tabular Generative Adversarial Network (ND-CTGAN). EEG data from subjects with various subtypes of ADHD was used, and multiple SMOTE modifications were employed to address class imbalances. The expanded datasets were classified using seven various classifiers, resulting in notable improvements in accuracy. ND-CTGAN generates high-quality synthetic data that maintains complex, non-linear relationships and essential inter-channel correlations compared to EEG-GAN. This strategy is effective, as the data generated resembles real data features similar to the individual's.



(a)



(b)



(c)

Fig. 10 ROC for Class 0 (Combine), Class 1 (Hyperactive), Class 2 (Inattentive) data: (a) for SVM, (b) for Random Forest, (c) for MLP using ND-CTGAN classifiers.

The synthetic and real data analysis using statistical and quality metrics comprise mean and STD, CDF, quality score: KS statistic, correlation similarity indicates that implementing sophisticated data augmentation techniques, ND-CTGAN can significantly enhance the classification accuracy of EEG data for individuals with Hyperactive ADHD. It primarily maintains the unique characteristics of each participant, facilitating meaningful comparisons between actual and synthetic datasets. This research emphasizes the potential of sophisticated generative models in dealing with class imbalances and improving the ability of machine learning models to generalize when applied to EEG data. Further research can enhance and explore the applicability of these methods to different forms of EEG and relevant neurological disorders.

Data Availability Statement

The data which was used in this research is publicly available on the following website: https://osf.io/4285y/.

Conflict of Interest

The authors declare no conflicts of interest.

Ethical Standards

Not applicable (Datasets are mentioned in the dataset availability statement).

Clinical Trials

This research did not undertake any clinical trials.

Funding Statement

The authors did not receive support from any organization for the submitted work.

References

1. R. J. Barry, A. R. Clarke, and S. J. Johnstone, "A review of electrophysiology in attention-deficit/hyperactivity disorder: I. Qualitative and quantitative electroencephalography," Clinical Neurophysiology 114(2), 171–183 (2003).
2. M. N. I. Qureshi, B. Min, H. J. Jo, and B. Lee, "Multiclass classification for the differential diagnosis on the ADHD subtypes using recursive feature elimination and hierarchical extreme learning machine: structural mri study," PLoS ONE 11(8), e0160697 (2016).
3. S. L. Karalunas, J. T. Nigg, "Heterogeneity and subtyping in attention-deficit/hyperactivity disorder—considerations for emerging research using person-centered computational approaches," Biological Psychiatry 88(1), 103–110 (2020).
4. D. Katsuki, H. Yamashita, K. Yamane, S. Kanba, and K. Yoshida, "Clinical subtypes in children with attention-deficit hyperactivity disorder according to their child behavior checklist profile," Child Psychiatry & Human Development 51(6), 969–977 (2020).
5. M. Rostami, S. Farashi, R. Khosrowabadi, and H. Pouretemad, "Discrimination of ADHD subtypes using decision tree on behavioral, neuropsychological and neural markers," Basic and Clinical Neuroscience Journal 11(3), 359–368 (2019).
6. A. Kamida, K. Shimabayashi, M. Oguri, T. Takamori, N. Ueda, Y. Koyanagi, N. Sannomiya, H. Nagira, S. Ikunishi, Y. Hattori, K. Sato, C. Fukuda, Y. Hirooka, and Y. Maegaki, "EEG power spectrum analysis in children with ADHD," Yonago Acta Medica 59(2), 169–173 (2016).
7. M. Rostami, R. Khosrowabadi, B. Albrecht, A. Rothenberger, and H. Pouretemad, "Classifying ADHD subtypes/presentations considering the joint effect of three levels of investigation," Nordic Journal of Psychiatry 75(1), 31–37 (2021).
8. R. Aldemir, E. Demirci, H. Per, M. Canpolat, S. Özmen, and M. Tokmakçı, "Investigation of attention deficit hyperactivity disorder (ADHD) sub-types in children via EEG frequency domain analysis," International Journal of Neuroscience 128(4), 349–360 (2018).
9. A. R. Clarke, R. J. Barry, R. McCarthy, and M. Selikowitz, "EEG-defined subtypes of children with attention-deficit/hyperactivity disorder," Clinical Neurophysiology 112(11), 2098–2105 (2001).
10. C. Rommel, J. Paillard, T. Moreau, and A. Gramfort, "Data augmentation for learning predictive models on EEG: a systematic comparison," Journal of Neural Engineering 19(6), 066020 (2022).
11. A. G. Habashi, A. M. Azab, S. Eldawlatly, and G. M. Aly, "Generative adversarial networks in EEG analysis: an overview," Journal of NeuroEngineering and Rehabilitation 20(1), 40 (2023).
12. Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert, "Deep learning-based electroencephalography analysis: a systematic review," Journal of Neural Engineering 16(5), 051001 (2019).
13. V. Joshi, N. Nanavati, "Inter-class correlation-based EEG channel selection for ADHD classification," IAENG International Journal of Computer Science 51(8), 961–975 (2024).
14. J. J. Newson, T. C. Thiagarajan, "EEG frequency bands in psychiatric disorders: a review of resting state studies," Frontiers in Human Neuroscience 12, 521 (2019).
15. J. Kwasa, A. Noyce, and B. Shinn-Cunningham, "Top-down attention modulates neural responses in neurotypical, but not ADHD, young adults," The Journal of the Acoustical Society of America 150(4_Supplement), A64–A64 (2021).
16. C. C. Williams, D. Weinhardt, M. Wirzberger, and S. Musslick, "Augmenting EEG with generative adversarial networks enhances brain decoding across classifiers and sample sizes," Proceedings of the Annual Meeting of the Cognitive Science Society 45(45) (2023).
17. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," Journal of Artificial Intelligence Research 16, 321–357 (2002).
18. N. Alsharif, M. H. Al-Adhaileh, and M. Al-Yaari, "Accurate identification of attention-deficit/hyperactivity disorder using machine learning approaches," Journal of Disability Research 3(1), (2024).
19. L. Shahmoradi, Z. Liraki, M. Karami, B. Savareh, and M. Nosratabadi, "Development of decision support system to predict neurofeedback response in adhd: an artificial neural network approach," Acta Informatica Medica 27(3), 186 (2019).
20. T. V. Padmavathy, M. Vinothkumar, and M. N. Vimal Kumar, "Modified cuckoo search- cascade forest (Mcs-cf) for attention deficit hyperactivity disorder (adhd) diagnosis," NeuroQuantology 18(7), 83–94(2020).
21. H. Alkahtani, T. Aldhyani, Z. Ahmed, and A. Alqarni, "Developing system-based artificial intelligence models for detecting the attention deficit hyperactivity disorder," Mathematics 11(22), 4698 (2023).
22. A. Alhudhaif, "A novel multi-class imbalanced EEG signals classification based on the adaptive synthetic sampling (ADASYN) approach," PeerJ Computer Science 7, e523 (2021).
23. Z. Jiang, W. Zhao, "Fusion algorithm for imbalanced EEG," Seizure 91, 207–211 (2021).
24. P. Zhang, X. Zhang, and A. Liu, "Effects of data augmentation with the bnnsmote algorithm in seizure detection using 1D-MobileNet," Journal of Healthcare Engineering 2022, 1–13 (2022).

25. R. Wardoyo, I. M. A. Wirawan, and I. G. A. Pradipta, "Data processing in seizure detection oversampling approach using radius- SMOTE for imbalance electroencephalography datasets," Emerging Science Journal 6(2), 382–398 (2022).
26. Y. Chen, R. Chang, and J. Guo, "Effects of data augmentation method borderline-smote on emotion recognition of eeg signals based on convolutional neural network," IEEE Access 9, 47491–47502 (2021).
27. T. Lee, M. Kim, and S. P. Kim, "Data augmentation effects using borderline-SMOTE on classification of a P300-based BCI," 8th International Winter Conference on Brain-Computer Interface, 9–12 (2020).
28. T. Lee, M. Kim, and S.-P. Kim, "Improvement of P300-based brain–computer interfaces for home appliances control by data balancing techniques," Sensors 20(19), 5576 (2020).
29. K. Rasheed, J. Qadir, T. J. O'Brien, L. Kuhlmann, and A. Razi, "A generative model to synthesize EEG data for epileptic seizure prediction," IEEE Transactions on Neural Systems and Rehabilitation Engineering 29, 2322–2332 (2021).
30. D. Geng, Z. S. Chen, "Auxiliary classifier generative adversarial network for interictal epileptiform discharge modeling and EEG data augmentation," 10th International IEEE/EMBS Conference on Neural Engineering, 1130–1133 (2021).
31. M. P. Kalashami, M. M. Pedram, and H. Sadr, "EEG feature extraction and data augmentation in emotion recognition," Computational Intelligence and Neuroscience 2022, 1–16 (2022).
32. A. Zhang, L. Su, Y. Zhang, Y. Fu, L. Wu, and S. Liang, "EEG data augmentation for emotion recognition with a multiple generator conditional Wasserstein GAN," Complex & Intelligent Systems 8(4), 3059–3071 (2022).
33. M. Xu, Y. Chen, Y. Wang, D. Wang, Z. Liu, and L. Zhang, "BWGAN-GP: an EEG data generation method for class imbalance problem in RSVP tasks," IEEE Transactions on Neural Systems and Rehabilitation Engineering 30, 251–263 (2022).
34. Y. Song, L. Yang, X. Jia, and L. Xie, "Common spatial generative adversarial networks based EEG data augmentation for cross-subject brain-computer interface," arXiv preprint arXiv:2102.04456 (2021).
35. R. Zhang, Y. Zeng, L. Tong, J. Shu, R. Lu, K. Yang, Z. Li, and B. Yan, "ERP-WGAN: A data augmentation method for EEG single-trial detection," Journal of Neuroscience Methods 376, 109621 (2022).
36. J. Yoon, D. Jarrett, and M. van der Schaar, "Time-series generative adversarial networks," Advances in Neural Information Processing Systems 32, 1–11 (2019).
37. B. Ganti, G. Chaitanya, R. S. Balamurugan, N. Nagaraj, K. Balasubramanian, and S. Pati, "Time-series generative adversarial network approach of deep learning improves seizure detection from the human thalamic SEEG," Frontiers in Neurology 13, 755094 (2022).
38. O. George, R. Smith, P. Madiraju, N. Yahyasoltani, and S. I. Ahamed, "Data augmentation strategies for EEG-based motor imagery decoding," Heliyon 8(8), e10240 (2022).
39. L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional GAN," Advances in Neural Information Processing Systems 32, (2019).
40. M. J. Cheon, D. H. Lee, J. W. Park, H. J. Choi, J. S. Lee, and O. Lee, "CTGAN VS TGAN? Which one is more suitable for generating synthetic EEG data," Journal of Theoretical and Applied Information Technology 99(10), 2359–2372 (2021).
41. V. Joshi, N. Nanavati, "A review of EEG signal analysis for diagnosis of neurological disorders using machine learning," Journal of Biomedical Photonics and Engineering 7(4), (2021).