Generative AI with WGAN-GP for boosting seizure detection accuracy

Lina Abou-Abbas[1,2,3]*, Khadidja Henni[1,2], Imene Jemal[2] and Neila Mezghani[1,2]

[1]Applied Artificial Intelligence Institute (I2A), TELUQ University, Montreal, QC, Canada, [2]Department of Science and Technology, TELUQ University, Montreal, QC, Canada, [3]Department of Electrical and Computer Engineering, Lebanese American University, Byblos, Lebanon

Background: Imbalanced datasets pose challenges for developing accurate seizure detection systems based on electroencephalogram (EEG) data. Generative AI techniques may help augment minority class data to facilitate automatic epileptic seizure detection.

New method: This study investigates the impact of various data augmentation (DA) approaches, including Wasserstein Generative Adversarial Network with Gradient Penalty (WGAN-GP), Vanilla GAN, Conditional GAN (CGAN), and Cramer GAN, on classification performance with Random Forest models. The best-performing GAN variant, WGAN-GP, was then integrated with a bidirectional Long Short-Term Memory (LSTM) architecture and compared against traditional and synthetic oversampling methods.

Results: The evaluation of different GAN variants for data augmentation with Random Forest classifiers identified WGAN-GP as the most effective approach. The integration of WGAN-GP with bidirectional LSTM yielded substantial performance improvements, outperforming traditional oversampling methods and achieving an accuracy of 91.73% on the augmented data, compared to 86% accuracy on real data without augmentation.

Comparison with existing methods: The proposed generative AI approach combining WGAN-GP and recurrent neural network models outperforms comparative synthetic oversampling methods on metrics relevant for reliable seizure detection from imbalanced EEG datasets.

Conclusions: Incorporating the WGAN-GP generative AI technique for data augmentation and integrating it with bidirectional LSTM elevates seizure detection accuracy for imbalanced EEG datasets, surpassing the performance of traditional oversampling and class weight adjustment methods. This approach shows promise for improving epilepsy monitoring and management through enhanced automated detection system effectiveness.

KEYWORDS

epileptic seizure, EEG, data augmentation, WGAN-GP, LSTM

Highlights

- Epilepsy detection systems often struggle with imbalanced datasets, impacting their accuracy.
- This study explores the potential of generative AI techniques for augmenting EEG data to facilitate improved seizure detection.

- The proposed WGAN-GP data augmentation approach, when integrated with a bidirectional LSTM architecture, outperforms conventional synthetic data generation methods for boosting seizure detection performance.

1 Introduction

Epilepsy is a neurological condition characterized by recurring seizures, abrupt disruptions in the brain's usual electrical activity (Fisher et al., 2005). These seizures can manifest in various symptoms, including muscle spasms, loss of consciousness, and alterations in behavior. Timely identification of an impending seizure empowers individuals to mitigate its impact by seeking medical assistance (World Health Organization, 2006).

Electroencephalography (EEG) serves as a widely employed diagnostic tool in epilepsy, playing a vital role in assessing treatment options for seizure management (Stafstrom and Carmant, 2015; Jemal et al., 2022). However, the practical implementation of EEG-based systems for seizure detection and prediction encounters significant hurdles, such as protracted calibration time and limited applicability. The primary challenge in seizure research lies in collecting data during actual seizure occurrences, primarily due to their unpredictable nature. Acquiring an extensive and diverse sample size required for obtaining reliable outcomes poses a substantial obstacle (Kuhlmann et al., 2018).

Imbalanced data in the context of data analysis refers to scenarios where the quantity of data points within one class significantly differs from another class. This imbalance can impede the accuracy of machine learning models in predicting the occurrence of epilepsy, as models can be more influenced by the more prevalent class, typically the negative examples (Kuhlmann et al., 2018). Consequently, this leads to subpar performance and increased false negatives, affecting the precision and reliability of machine learning models in predicting seizure outcomes or classifying different types of seizures.

Addressing the challenge of imbalanced data is crucial for advancing epilepsy research and treatment. Several approaches can be employed, including resampling techniques like oversampling or undersampling, and algorithms specifically designed for imbalanced data, such as decision trees or random forests.

Data augmentation is another valuable technique, which involves generating new synthetic examples based on existing data. Various methods for data augmentation include synthetic data generation, data transformation, data combination, and data noise injection (Dubey et al., 2014; Johnson and Khoshgoftaar, 2019; Kraiem et al., 2021). By using these techniques, it may be possible to overcome the challenge of imbalanced data and enhance the performance of machine learning models in epilepsy research. It is also important to carefully evaluate the performance of the model on the imbalanced data, using metrics such as precision, recall, and the F1 score, rather than solely relying on overall accuracy.

The primary objective of this investigation was to examine various data augmentation (DA) approaches in order to tackle the issue of imbalanced data. To amplify the representation of the minority classes, we explored diverse generative AI techniques. Initially, we employed four different variants of deep learning generative adversarial networks (GAN) to generate synthetic data targeting the underrepresented classes. We conducted a projection analysis and Qq plot analysis to compare the synthetic and real data and ascertain the GAN variant that accurately replicated the real data.

Furthermore, we utilized conventional DA methods such as SMOTE and ADASYN to expand the existing dataset. We evaluated the effectiveness of these techniques in comparison to the most effective GAN variant. To analyze the impact of data augmentation (DA) techniques on classification performance, we trained Long Short-Term Memory (LSTM) models and evaluated their performance on datasets with and without DA. Similarly, we used the Random Forest algorithm to compare performance metrics between classifiers trained solely on real data and those trained on a combination of real and synthetic data generated by DA techniques. This comparison helps us understand how different DA methods influence the LSTM and RF's classification accuracy and other performance metrics.

Our primary contribution lies in the systematic evaluation of various GAN architectures and conventional oversampling techniques for generative data augmentation in order to address class imbalance in seizure detection using LSTM classifiers. We demonstrate the effectiveness of WGAN-GP in synthesizing minority class data that closely aligns with the distribution of real data, resulting in significant performance improvement. This provides a crucial framework and benchmark for further research on data augmentation strategies for managing imbalanced data in EEG classifications and seizure detection.

2 Related state of art

In recent years, DA techniques have emerged as a promising approach to improve the performance of machine learning models in epilepsy detection tasks. By increasing the amount of available training data, DA can help to reduce overfitting and improve the generalization performance specifically when dealing with tasks based on electroencephalograms (EEGs).

For tasks such as emotion recognition (Wang et al., 2018; Salama et al., 2018; Luo and Lu, 2018; Chang and Jun, 2019), DA techniques have been shown to be instrumental in improving the performance of learning models. In the field of seizure detection, numerous investigations have employed DA techniques to advance the accuracy of deep learning models (Wei et al., 2019; Hussein et al., 2018).

A comprehensive review conducted by Lashgari et al. (2020) explored different DA approaches such as noise addition, GAN networks, sliding windows, sampling, Fourier transform, and recombination of segmentation and showed that DA is becoming a prevalent approach in EEG tasks, resulting in a significant improvement in accuracy, with an average increase of 29%.

The method proposed by Yuan et al. (2017), using the weighted extreme learning machine (ELM), showed to be an effective approach to mitigate the impact of the imbalanced class distribution on performance.

In a new study published in 2022, Zhang et al. (2022), a new algorithm called BNNSMOTE was proposed to improve seizure detection performance by synthesizing new samples to address the

challenge of imbalanced classification of seizure and non-seizure data.

In other studies (Ullah et al., 2018; Wang et al., 2021), augmentation techniques were used to address the problem of limited data and to improve the performance of the one-dimensional convolutional neural network (P-1D-CNN) model in the detection of seizures. Learned models showed high accuracy, sensitivity, and specificity.

Furthermore, researchers proposed using synthetic data generated by Temporal GAN to increase sample size and improve deep learning classification performance when faced with the limited availability of labeled thalamic EEG data (Ganti et al., 2022).

In addition, Zhao et al. (2022) presented an EEG augmentation method (EEGAug), involving the random selection of a few samples from the underrepresented class, the transformation of these samples into the frequency domain, the combination of different frequency bands, and conversion back to the time domain to generate new samples. Through the use of EEGAug, the authors balanced the unbalanced clinical iEEG data, which performed the best in most cases. Three studies have proposed using GANs for DA in different applications.

In one study, a WGAN-GP model is proposed for DA in emotion recognition using the DEAP dataset (Bhat and Hortal, 2021).

In another study (Hartmann et al., 2018), GANs are used to generate electroencephalographic (EEG) brain signals. A modification to the training of Wasserstein GANs is introduced to stabilize the training process.

In the third study (Haradal et al., 2018), a new method is presented to create synthetic time series data using GANs for the classification of biosignals. These approaches can have limitations and may not always be effective.

Based on the current state of the art in epilepsy detection, it is clear that DA and the issue of imbalanced datasets need to be further addressed. While recent studies have shown promising results using GANs for DA in EEG-based classification tasks, more research is needed to explore the potential of these techniques for improving the accuracy and reliability of epileptic seizure detection. As such, there is a pressing need for more studies to focus on developing effective strategies for DA and addressing data-imbalanced issues in epilepsy detection.

3 Method and materials

3.1 Database

The current research utilized data obtained from the Temple University Hospital EEG Seizure Corpus (TUSZ), version v1.5.1. This corpus consists of EEG recordings obtained in real-time from 341 patients undergoing clinical monitoring. The recordings were conducted utilizing the 10/20 international standard system, and among these patients, 188 were female. It includes 886 sessions, each ranging from one minute to one hour, thus providing a varied and diverse array of EEG signals. The TUSZ includes approximately 6% of files containing EEG seizure segments, which amount to a total duration of 40.41 hours. The TUSZ dataset offers several montages, but for consistency in this study, we focused exclusively

TABLE 1   Detailed overview of the Temple University Hospital EEG Seizure Corpus (TUSZ) used in our seizure detection experiments.

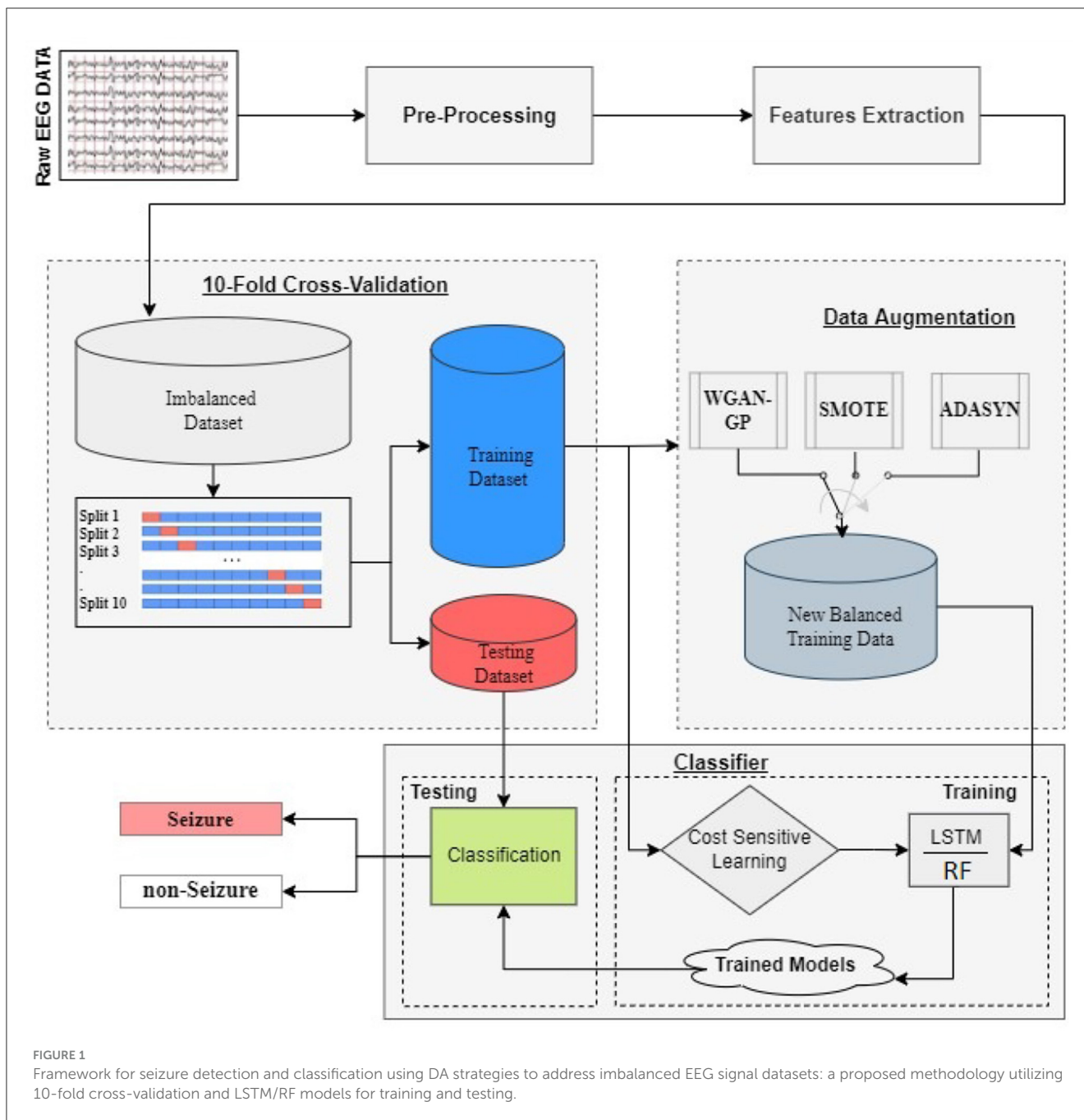| Number of patients (female)	341 (188 F)
Number of patients with seizure	133 (72 F)
Total number of sessions	886
Total number of files	7,634
Number of seizure files	1,780
Number of seizure-free files	5,854
Total duration in hrs.	655.36

on the most popular bipolar Temporal Central Parasagittal (TCP) Averaged Reference (AR) montage, ensuring uniformity across all analyzed data. All EEG signals were sampled at a minimum rate of 250 Hz to ensure recording accuracy. Prior to analysis, qualified researchers removed any artifacts and eye blinks present in the recordings by using an open-source annotation tool. Table 1 provides detailed information on the key features of the corpus used in the study. The corpus was carefully selected for its diverse range of EEG signals, providing an opportunity to conduct an in-depth analysis of seizure segments in a clinical setting. For more information about the database used, we invite readers to read Shah et al. (2018).

3.2 The proposed framework

Figure 1 displays the proposed framework for detecting and classifying seizures using DA techniques to address imbalanced datasets. The framework consists of two primary steps. In the first step, pre-processing of the EEG signal is performed, followed by feature extraction. For each of the 19 EEG channels, 44 time and/or frequency domain features are extracted, resulting in a total of 836 features. The list of features is detailed in Table 2 and for more details refer to our previous work (Abou-Abbas et al., 2023, 2022, 2021).

Next, a 10-fold cross-validation method is employed to partition the data. The cross-validation is performed using the GroupKFold technique, to ensure that all samples from a single patient are either included entirely in the training set or entirely in the test set for each fold. This prevents any overlap of patient information between the training and test phases. Specifically, the data is divided into 10 groups (or folds), and for each iteration, 9 of these folds are used to train the classifier, while the remaining fold is used for testing. This process is repeated ten times, with each fold serving as the test set once. The performance measures are averaged across all ten iterations to obtain the final results.

The training set is then utilized in two distinct ways. The first approach employs DA techniques, such as WGAN-GP, SMOTE, or ADASYN, to balance the dataset. The newly balanced dataset is then used to train the LSTM model. The second approach directly trains the LSTM model using the imbalanced dataset, but with a cost-sensitive loss function or class weighting approach. In both scenarios, the test data is reserved for testing the trained model.

FIGURE 1
Framework for seizure detection and classification using DA strategies to address imbalanced EEG signal datasets: a proposed methodology utilizing 10-fold cross-validation and LSTM/RF models for training and testing.

3.3 Data augmentation

DA is a technique utilized to create new training data by altering the existing data. Its primary goal is to enhance the performance of classification systems by forming new and different examples to train datasets. It is a valuable tool in situations where collecting new data is difficult or costly (Mumuni and Mumuni, 2022). It can be achieved by making small changes to the dataset or using deep learning models to generate new data points to expand the training set, enhance its diversity, and improve model generalization. Generative DA techniques, such as GANs and their variants like CramerGAN, CGAN, Vanilla GAN, and WGAN-GP, employ generative models to produce entirely new data points

that resemble the real data distribution (Goodfellow et al., 2020; Creswell et al., 2018; Gulrajani et al., 2017; Arjovsky et al., 2017). These generative approaches enable the creation of realistic synthetic samples that capture the underlying characteristics of the training data. On the other hand, synthetic DA techniques focus on creating new samples by applying transformations or modifying existing data points. This includes techniques like SMOTE and ADASYN, which generate synthetic samples for minority classes by interpolating between existing instances or adapting the generation process based on the data distribution (Chawla et al., 2002; He et al., 2008). Several studies have explored DA techniques in different fields. For instance, a study proposed a general methodology for small medical data classification that deploys an augmentation

TABLE 2 List of all the 44 features extracted.

Number	Features
1–5	Average value of power spectral density (delta, theta, alpha, beta, gamma)
6–10	Absolute value of power spectral density (delta, theta, alpha, beta, gamma)
11–15	Relative value of power spectral density (delta, theta, alpha, beta, gamma)
16–18	Skewness, variance and kurtosis
19–24	Features of empirical mode decomposition (energy, spectral entropy, mean, standard deviation, moment, skewness)
25–29	Sample entropy, permutation entropy (4 levels)
30–32	Hjorth (mobility, activity, complexity)
33	Spectral entropy of PSD
34	Features of discrete wavelet transform-Shannon entropy
35–36	Features of wavelet packet decomposition (log energy entropy and Shannon entropy)
37	Successive decomposition index
38–39	Mean Energy and its cumulative sum
40–44	Features of wavelet Decomposition (Percentage of energy-5 levels)

technique and a feature selection strategy (Alauthman et al., 2023). Another study presented a comprehensive survey of modern DA techniques (Mumuni and Mumuni, 2022). In this study, we will explore the generation of data using GAN (Generative Adversarial Network) and its variants, including CramerGAN, CGAN, Vanilla GAN, and WGAN-GP. Additionally, we will compare these generative approaches with synthetic data generation techniques such as SMOTE and ADASYN. Through comparative analysis with synthetic approaches like SMOTE and ADASYN, we will assess the effectiveness of generative DA in addressing class imbalance and improving the performance of machine learning models.

3.3.1 Generative Adversarial Network and its variants

Generative Adversarial Networks, or GANs for short, have become a very powerful deep learning technique for data generation and augmentation. GANs are made up of two key components - a generator and a discriminator. They work by having the generator creates synthetic data, while the discriminator evaluates this data to determine whether it is real or artificial. The original GAN architecture, introduced by Goodfellow et al. (2020) and Creswell et al. (2018) is commonly referred to as the Vanilla GAN. Over the years, researchers have proposed numerous GAN variants to address various challenges and limitations to the original formulation to improve performance and extend their applications (Gulrajani et al., 2017; Arjovsky et al., 2017; Bellemare et al., 2017; Mirza and Osindero, 2014; Gulrajani et al., 2017). One notable variant, the CramerGAN, replaces the discriminator with a Cramer distance estimator improving training stability and the quality of

generated samples (Bellemare et al., 2017). Another influential variant is the Conditional GAN (CGAN), which incorporates additional conditioning information such as class labels or auxiliary features, into both the generator and discriminator networks to enable better control over the generated data (Mirza and Osindero, 2014).The Wasserstein GAN with Gradient penalty (WGAN-GP) proposed by Gulrajani et al. (2017) and Arjovsky et al. (2017) addresses training instability and mode collapse issues by incorporating the Wasserstein distance as a more meaningful loss function. The gradient penalty further enhances training stability, encouraging the discriminator to behave as a true Wasserstein distance estimator. Figure 2 gives a detailed architecture of a basic GAN. A random noise (z) is used as input to the generator, a typical feature of GANs. The discriminator evaluates both real data from the dataset and the synthetic data generated. It assigns probabilities between 0 and 1 to these data samples, indicating their likelihood of being real or fake. The critic loss also called discriminator's loss is computed based on how well its predictions match the ground truth labels from the training dataset. This creates a feedback loop where the generator and discriminator continuously adjust: the generator improves the quality of its synthetic data to reduce its loss, while the discriminator enhances its ability to distinguish between real and synthetic data to minimize its own loss.

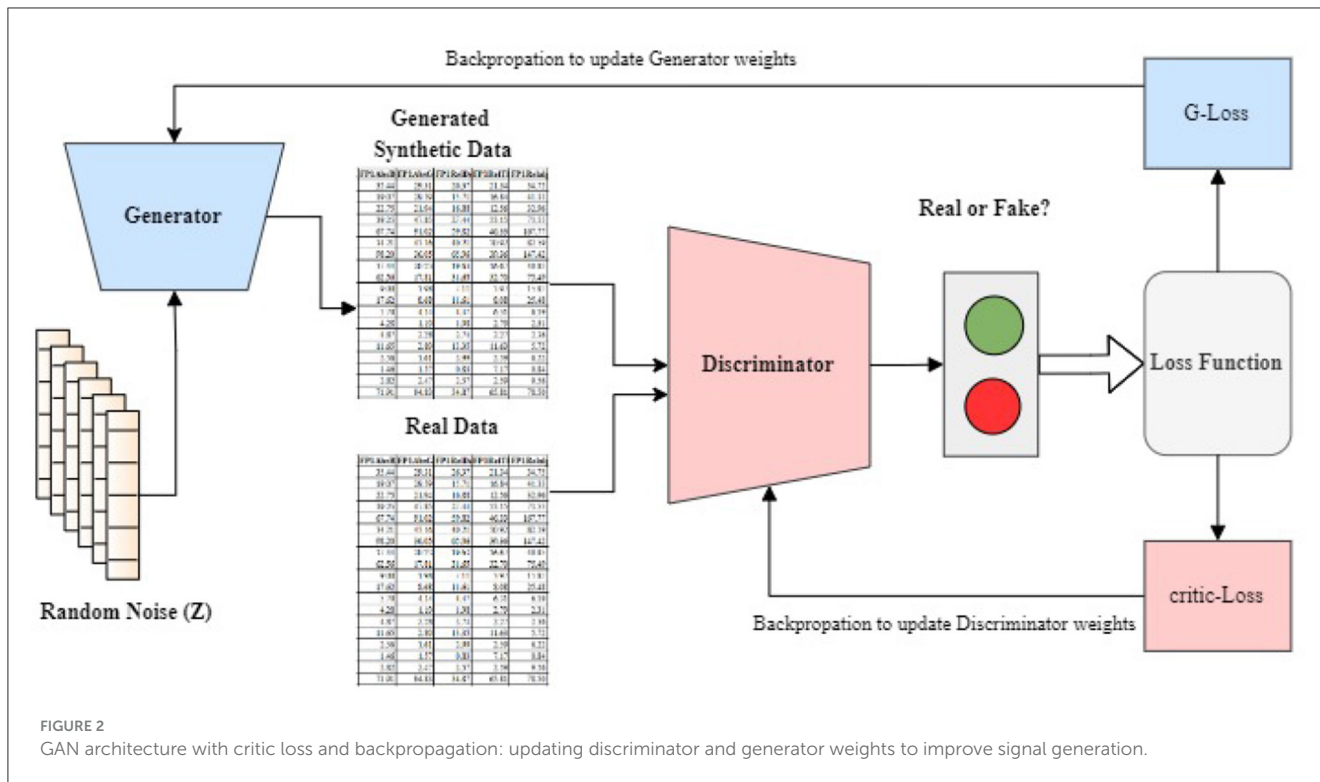### 3.3.2 Synthetic data generation techniques: SMOTE and ADASYN

The Synthetic Minority Over-sampling Technique (SMOTE) is a widely used method for addressing the issue of imbalanced datasets. SMOTE works by creating new samples for the minority class in a dataset. For each existing sample in the minority class, it finds a few similar samples, called its nearest neighbors. It then creates new samples by mixing the original sample with these neighbors. To do this mixing, SMOTE uses a simple math formula. For each minority class sample x_i, SMOTE selects k nearest neighbors. Specifically, if x_i and x_j are two samples from the minority class, a new synthetic sample x_{new} is generated using:

$x_{new} = x_i + \Lambda.(x_j - x_i)$ where Λ is a random number between 0 and 1. The algorithm repeats this process until the desired number of synthetic samples is generated. By creating these new samples, SMOTE expands the decision region of the minority class, helping to balance the dataset and potentially improving the performance of classification algorithms on the minority class (Chawla et al., 2002).

The Adaptive Synthetic Sampling (ADASYN), is another DA technique that generates synthetic samples of the underrepresented class. However, ADASYN diverges from other techniques by employing a distinct approach wherein the density distribution of the minority class is estimated and then the synthetic samples are produced in regions characterized by low density. Mathematically, ADASYN operates as follows: For each minority sample x_i, the density ratio r_i is calculated as:

$$r_i = \frac{\Delta_i}{K} \tag{1}$$

where Δ_i represents the number of majority class samples among its K nearest neighbors. These ratios are then normalized to form a

FIGURE 2
GAN architecture with critic loss and backpropagation: updating discriminator and generator weights to improve signal generation.

distribution:

$$\hat{r}_i = \frac{r_i}{\sum r_i} \tag{2}$$

The number of synthetic samples g_i to generate for each x_i is determined by:

$$g_i = G \cdot \hat{r}_i \tag{3}$$

where G is the total number of synthetic samples needed. New synthetic samples are created using linear interpolation:

$$x_{new} = x_i + \lambda \cdot (x_j - x_i) \tag{4}$$

where x_j is a randomly chosen minority neighbor of x_i, and λ is a random number in the interval $[0, 1]$. This process is repeated g_i times for each x_i, focusing on generating synthetic samples in regions where the minority class is underrepresented, thus adapting to the local data distribution (He et al., 2008).

3.4 LSTM architecture

Long Short-Term Memory (LSTM) networks are a special type of neural network designed to handle sequential data, like EEG signals. They're good at remembering important information over long periods, which is crucial for tasks like seizure detection. At its heart, an LSTM has a cell state, a sort of memory, and three principal components as gates: the input gate, which decides what new information to add to the cell state; the forget gate, which chooses what information to remove from the cell state; and the output gate, which determines what to output based on the cell

state. The combination of all these gates provides LSTMs with the ability to selectively remember or forget information over long sequences, making them really powerful in many tasks related to data that comes in a sequence or time series, such as EEG signal analysis data (Hochreiter and Schmidhuber, 1997).

Figure 3 illustrates the LSTM architecture implemented for classifying raw EEG data to detect seizures. This architecture is designed to leverage the strengths of LSTMs in processing sequential data. The LSTM network used in this study is unidirectional, meaning it processes data in a single direction, from past to future, which is appropriate for the nature of sequential EEG data.
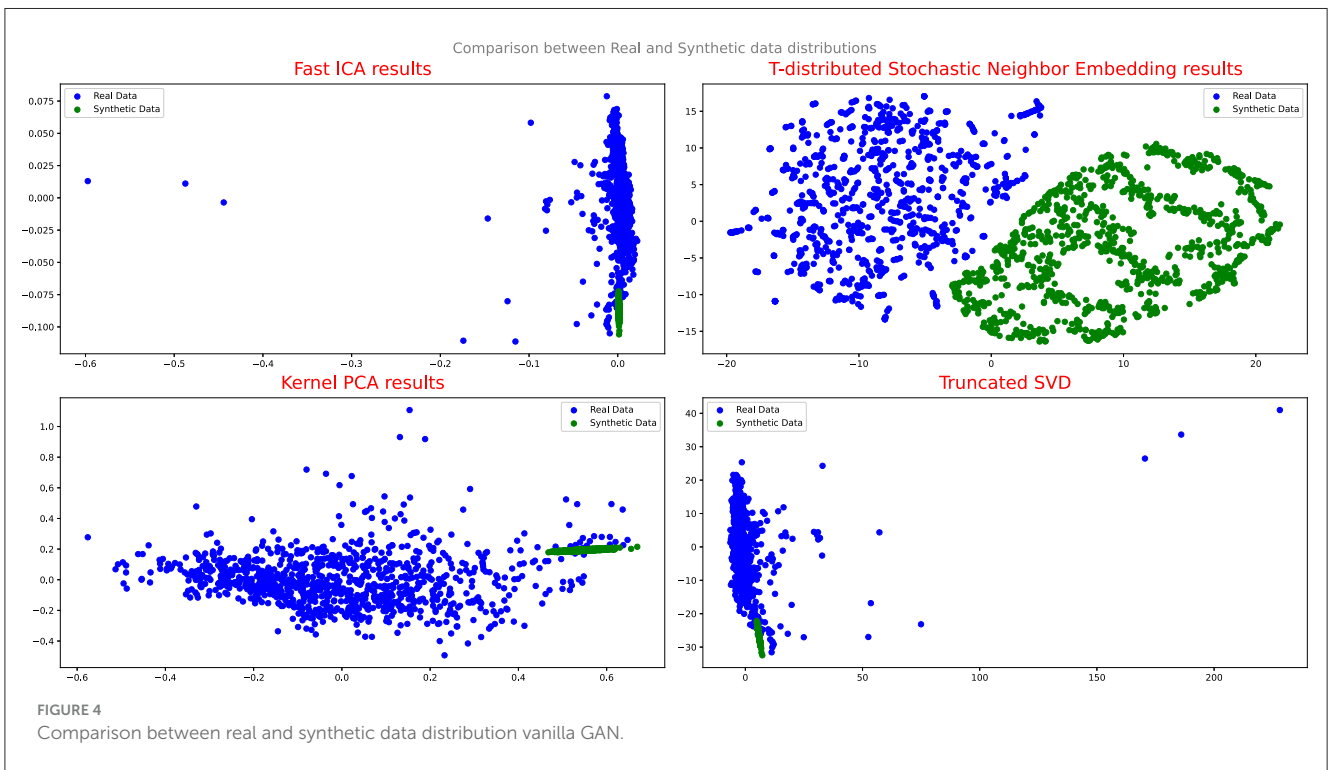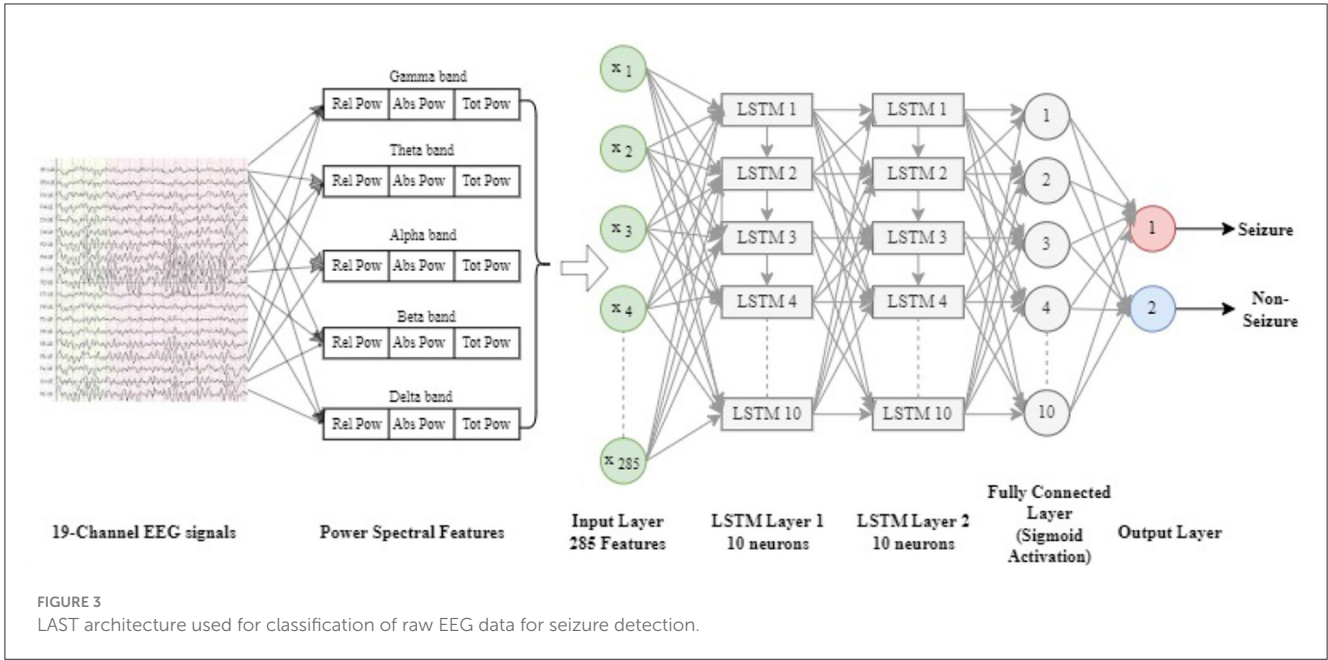
The input layer of the network receives 836 features derived from the raw EEG signals. These features are fed into two LSTM layers, each containing 10 neurons. Each LSTM unit within these layers applies the three gates–input, forget, and output gates–to manage the flow of information and maintain memory across time steps.

Following the LSTM layers, the network includes a fully connected (dense) layer. This layer transforms the output from the LSTM layers into a one-dimensional vector, which is then used for classification. The final output is passed through a Sigmoid activation function, which produces a probability score between 0 and 1, indicating the likelihood of a seizure event.
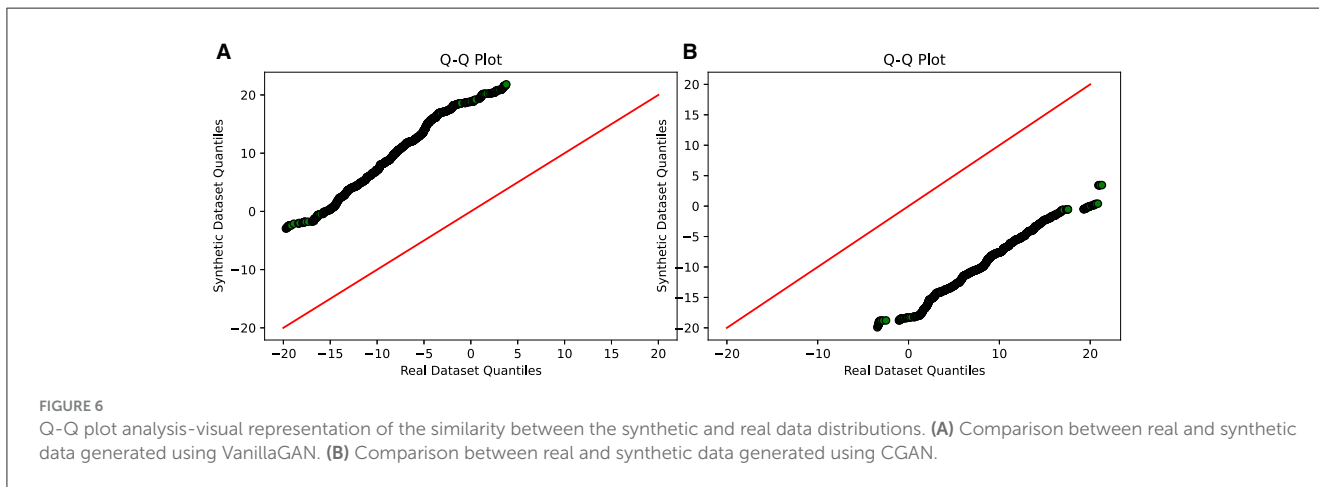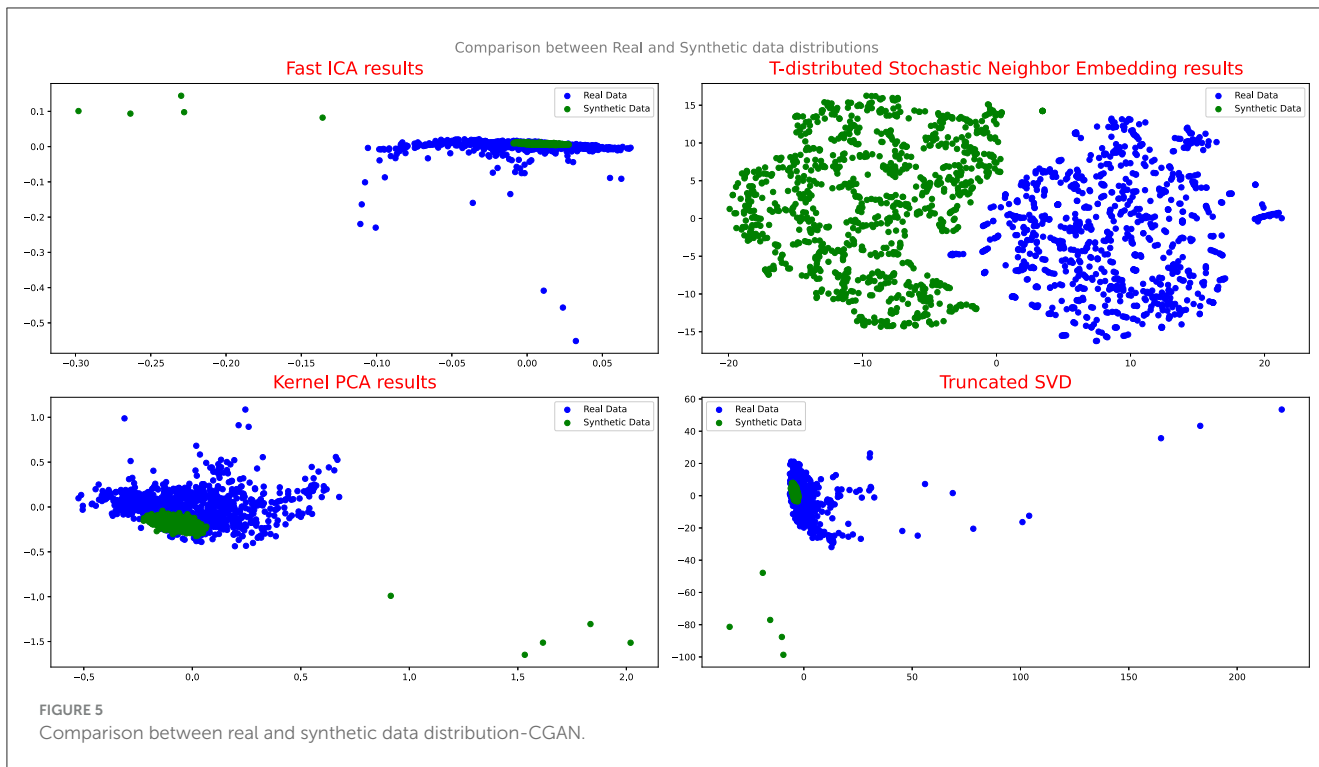
4 Results and discussion

4.1 Experimental setup

The framework illustrated in Figure 1 provides a comprehensive overview of the methodology. It starts with

FIGURE 3
LAST architecture used for classification of raw EEG data for seizure detection.

FIGURE 4
Comparison between real and synthetic data distribution vanilla GAN.

data augmentation (DA) using various Generative Adversarial Networks (GANs) and other DA techniques, including Vanilla GAN, CRAMER GAN, CGAN, WGAN-GP, SMOTE, and ADASYN. These methods were selected due to their ability to learn and generate realistic data distributions, enhancing the representation of minority classes in the augmented dataset. Once the augmented dataset is created, it is split into training and testing sets using a 10-fold cross-validation approach to ensure robust performance evaluation.

To evaluate the effectiveness of the DA approaches, the study employs several assessment methods. Initially, Quantile-Quantile (QQ) plot analysis is conducted to compare and analyze the distribution of real and synthetic data, providing insights into the fidelity of the synthetic data generation process and its alignment with the real data distribution. Additionally, projection techniques such as Fast Independent Component Analysis (Fast ICA), Kernel Principal Component Analysis (Kernel PCA), Truncated Singular Value Decomposition (Truncated SVD), and t-Distributed
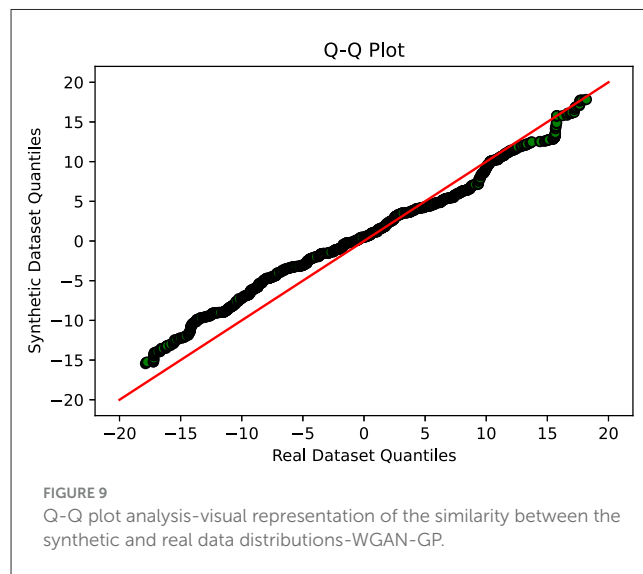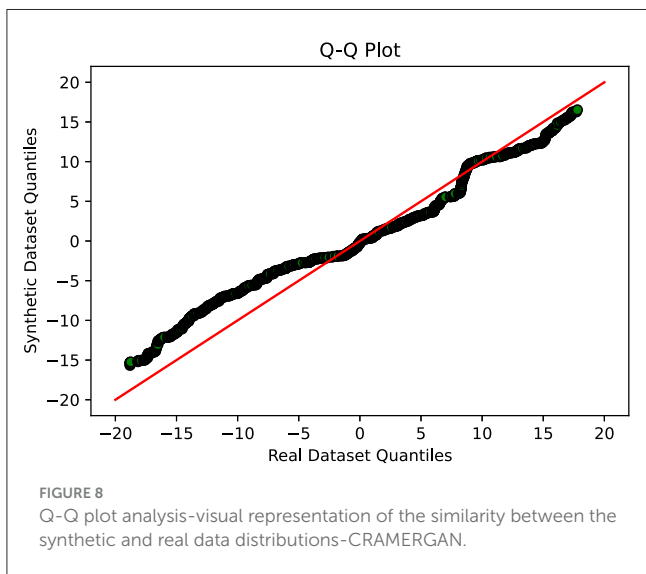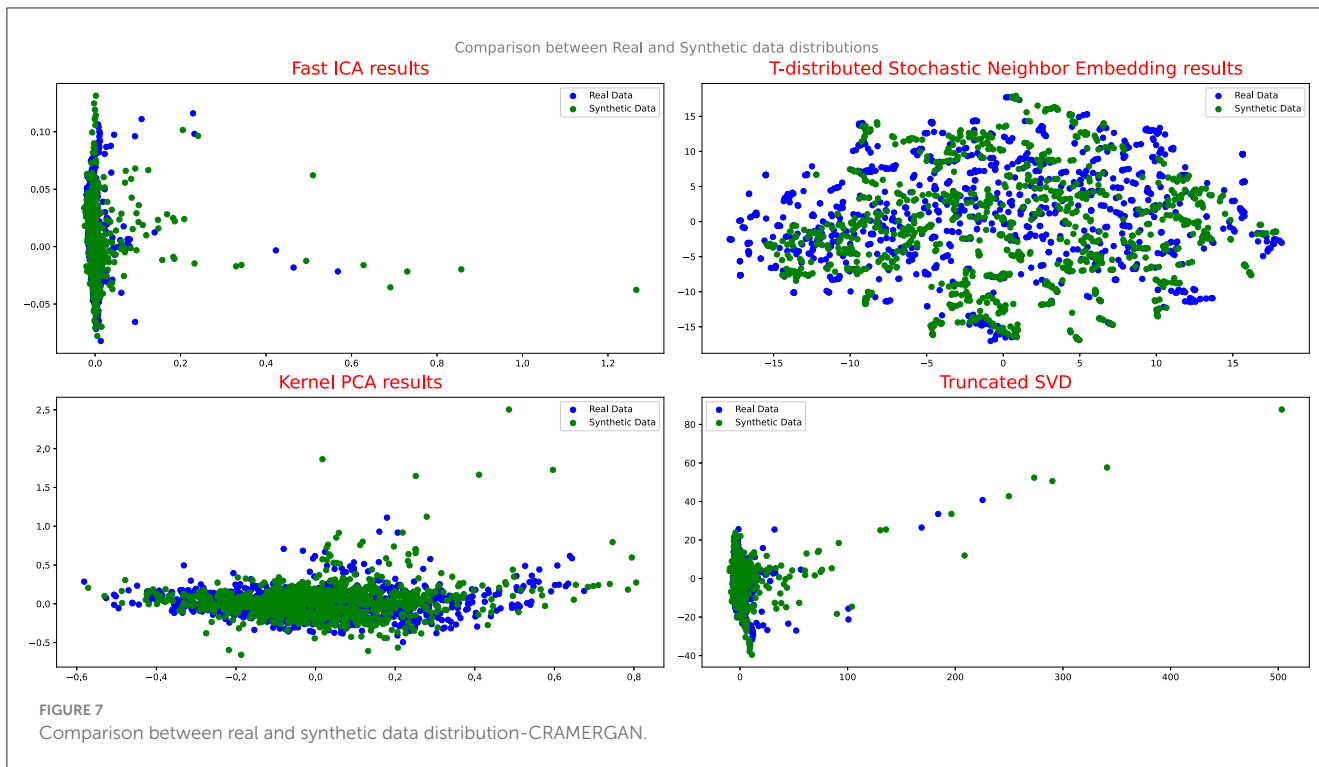
FIGURE 5
Comparison between real and synthetic data distribution-CGAN.

FIGURE 6
Q-Q plot analysis-visual representation of the similarity between the synthetic and real data distributions. **(A)** Comparison between real and synthetic data generated using VanillaGAN. **(B)** Comparison between real and synthetic data generated using CGAN.

Stochastic Neighbor Embedding (t-SNE) are utilized to analyze the results and evaluate the performance of the DA techniques.

The core of the study's experimental setup focuses on training and evaluating machine learning models using the augmented dataset. Both Random Forest (RF) and Long Short-Term Memory (LSTM) models are employed. The RF algorithm serves as a preliminary benchmark in our study, assessing the classifier's performance when trained on both real data and a mix of real and synthetic data. This evaluation is essential for understanding how data augmentation impacts a traditional machine learning model. Our previous work Abou-Abbas et al. (2022), has established that RF significantly outperforms other standard classifiers, including Support Vector Machines (SVM). Following this preliminary evaluation, the LSTM model, tailored for handling the sequential characteristics of EEG signals, is then trained and assessed. This approach allows us to explore how more advanced deep learning techniques compare with RF in seizure detection.

TensorFlow and Keras were used to implement the DA techniques, LSTM models, and cross-validation schemes (Fabio et al., 2013). The Adam optimizer as a gradient-based method with with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a learning rate of 0.001 was employed for model training. We used the early stopping criteria to prevent over-fitting where training runs up to 100 epochs, or until the validation loss does not decrease anymore for at least 20 epochs.

Several tests are performed to assess the effectiveness of each DA technique on the Temple University Seizure Detection Corpus (TUSZ) data. The experiments are repeated 10 times using different subsets of the dataset to ensure robust results. The performance of the LSTM model is compared across various DA techniques,

FIGURE 7
Comparison between real and synthetic data distribution-CRAMERGAN.

FIGURE 8
Q-Q plot analysis-visual representation of the similarity between the synthetic and real data distributions-CRAMERGAN.

FIGURE 9
Q-Q plot analysis-visual representation of the similarity between the synthetic and real data distributions-WGAN-GP.

providing a comprehensive analysis of how each method impacts the classification of imbalanced EEG signal datasets.

4.2 Comparative evaluation of GAN variants

A thorough comparative evaluation was conducted to assess the performance of different GAN variants in generating synthetic data that closely resembled real data distributions. The evaluation process involved analyzing the Q-Q plots which compare two sets of data by plotting their quantiles against each other. If the synthetic data matches the real data well, the points on the plot will form a nearly straight line, usually at a 45-degree angle. This line would mean the two sets of data have very similar distributions. When points stray from this line, it shows differences between the synthetic and real data. It also helps us judge if our methods for creating synthetic data are effectively balancing out our dataset, which originally had uneven amounts of data in different categories.

Figures 4, 5 illustrate the results of the projection analysis, specifically for the synthetic data generated respectively using Vanilla GAN and CGAN. As depicted in the figures, the projected synthetic data appears to be clustered together, significantly deviating from the distribution of the real data. These observation
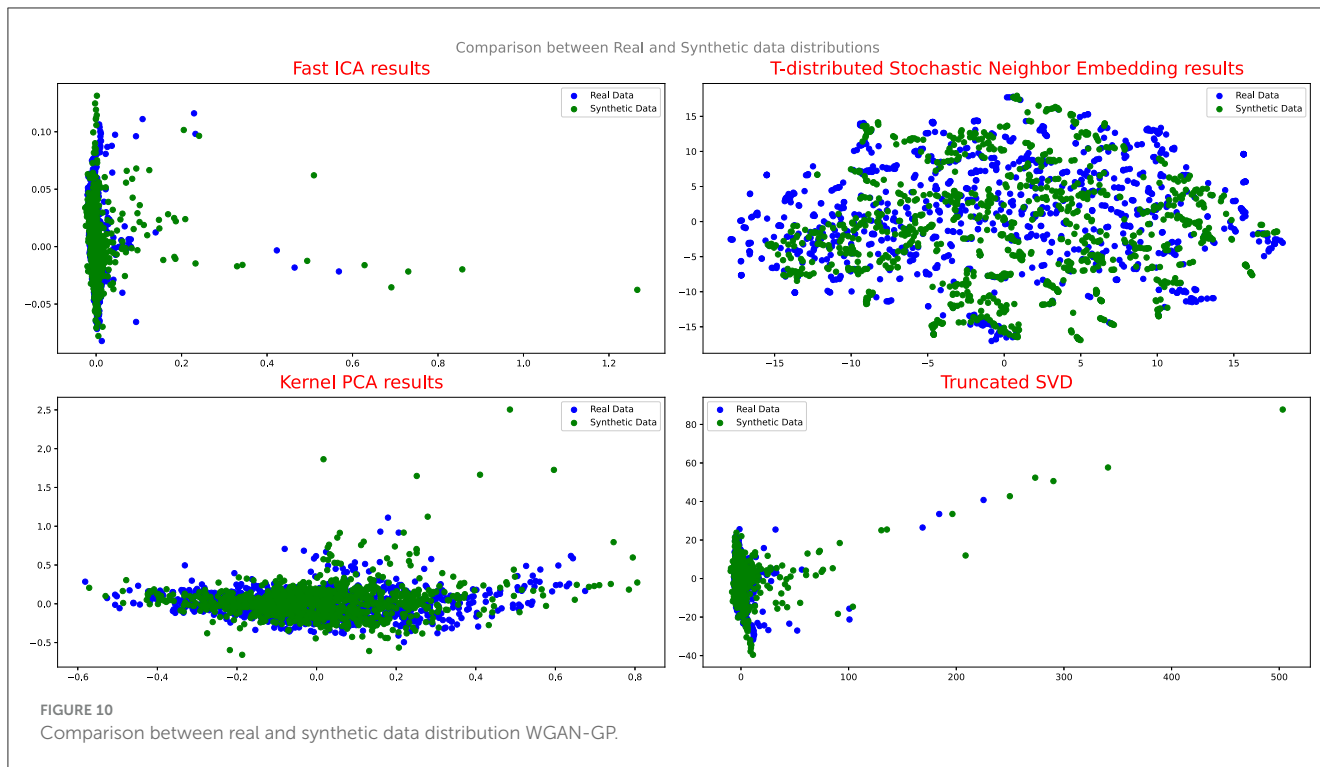
FIGURE 10
Comparison between real and synthetic data distribution WGAN-GP.

TABLE 3 Classification performance metrics for real data only and real + synthetic data.

Metrics	Real data only	Real + synthetic data
Accuracy	0.86	0.88
Recall	0.79	0.83
Specificity	0.67	0.72
F1 Score	0.83	0.86
Precision	0.81	0.84

TABLE 4 Classification performance metrics for seizure detection.

Approach	Loss	Precision	Recall	Specificity	Accuracy
SMOTE	0.1252	63.03%	70.57%	88.90%	85.25%
ADASYN	0.1281	61.50%	71.91%	88.47%	85.18%
Class Weight	0.1255	60.38%	72.93%	87.76%	84.82%
WGAN-GP	0.0681	80.81%	76.71%	95.45%	91.73%

suggests that the synthetic data generated by Vanilla GAN and CGAN fails to capture the inherent patterns and structure present in the real data distribution. Moreover, the form and shape of the synthetic data in Figures 4, 5 differ noticeably from that of the real data, indicating a significant disparity between the two distributions.

Furthermore, the Q-Q plot analysis reinforces these findings. The Q-Q plot in Figures 6A, B depicts a notable misalignment between the synthetic and real data distributions, with the points deviating considerably from the ideal diagonal line. This misalignment further supports the notion that Vanilla GAN and CGAN struggle to generate synthetic data that closely resembles the distribution of the real data.
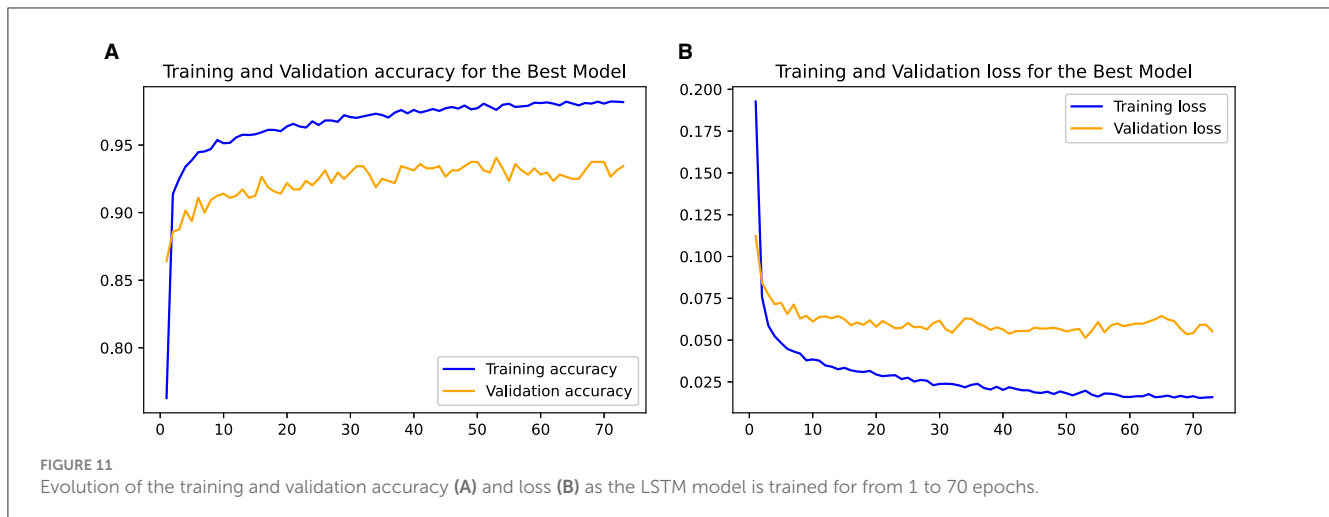
Upon evaluating the results of the projection analysis and Q-Q plots for Cramer GAN, it is observed that its performance is slightly improved compared to Vanilla GAN. Figure 7 displays the projection results of the synthetic data generated by Cramer GAN, where the dispersion of the synthetic data points appears to be more aligned with the real data distribution.

This indicates that Cramer GAN has a better ability to capture the underlying structure and patterns of the real data, resulting

in a more coherent and representative synthetic data distribution. Moreover, the Q-Q plot analysis in Figure 8 reveals a closer alignment between the synthetic and real data distributions, as the points in the plot exhibit a higher degree of adherence to the ideal diagonal line in the Q-Q plot. This suggests that CramerGAN generates synthetic data that closely mirrors the statistical characteristics of the real data. In both Q-Q plots and projection analyses, CramerGAN outperforms Vanilla GAN and CGAN. However, when comparing CramerGAN to WGAN-GP, the Q-Q plot results are quite similar as seen in Figure 9. The difference becomes visually apparent in the projection results shown in Figure 10, where WGAN-GP demonstrates superior performance. The synthetic data points generated by WGAN-GP show a remarkable alignment and integration with the real data, indicating a high degree of similarity and coherence. This enhanced performance suggests that WGAN-GP excels at capturing the underlying structure and patterns of the real data, resulting in a synthetic data distribution that closely resembles the real data distribution. The substantial improvements achieved by WGAN-GP in the projection analysis emphasize its efficacy in generating synthetic data that exhibits a remarkable resemblance to the real data distribution.

FIGURE 11
Evolution of the training and validation accuracy **(A)** and loss **(B)** as the LSTM model is trained for from 1 to 70 epochs.

The substantial improvements achieved by WGAN-GP in both projection analysis and Q-Q plots emphasize its efficacy in generating synthetic data that exhibits a remarkable resemblance to the real data distribution.

4.3 Impact of data augmentation on classification performance with random forest

The Impact of DA on Classification Performance with Random Forest was investigated, and the results are summarized in Table 1. The table compares the classification performance metrics for two scenarios: using real data only and using a combination of real and synthetic data. The metrics considered are Accuracy, Recall, Specificity, F1 Score, and Precision. The results in Table 3 demonstrate the impact of augmenting the data and training the Random Forest model with a combination of real and synthetic data to balance the dataset. By incorporating synthetic data, the overall performance of the classification model improved. The Accuracy of the model increased from 0.86 when using real data only to 0.88 when utilizing real and synthetic data. Similarly, the Recall increased from 0.79 to 0.83, indicating a higher ability to correctly identify positive instances.

Furthermore, the Specificity, which represents the model's ability to correctly identify negative instances, also improved from 0.67 to 0.72. The F1 Score, a harmonic mean of precision and recall, increased from 0.83 to 0.86, indicating a more balanced performance between these two metrics. Lastly, the Precision increased from 0.81 to 0.84, demonstrating the model's improved ability to correctly classify positive instances.

4.4 Classification for seizure detection using bi-directional LSTM

Table 4 presents the classification performance metrics for different DA approaches in the task of seizure detection. Four approaches were evaluated based on various metrics including Loss,

Precision, Recall, Specificity, and Accuracy. In terms of Loss, the SMOTE approach had a value of 0.1252, followed by ADASYN with 0.1281, Class Weight with 0.1255, and WGAN-GP with the lowest value of 0.0681. A lower loss value indicates better model performance in minimizing the discrepancy between predicted and actual values. Regarding Precision, the WGAN-GP approach achieved the highest precision of 80.81%. This indicates that it had a higher proportion of correctly classified positive instances compared to the other approaches. ADASYN had the second highest precision at 61.50%, followed by SMOTE with 60.38%, and Class Weight with the lowest precision of 63.03%. For Recall, the WGAN-GP approach achieved a value of 76.71%, indicating its ability to correctly identify positive instances. ADASYN had the second highest recall at 71.91%, followed by Class Weight with 72.93%, and SMOTE with the lowest recall of 70.57%. The WGAN-GP approach obtained the highest specificity of 95.45%, indicating its ability to distinguish negative instances accurately. ADASYN achieved the second highest specificity at 88.47%, followed by Class Weight with 87.76%, and SMOTE with the lowest specificity of 88.90%.Moreover, the WGAN-GP approach achieved the highest accuracy of 91.73%, followed by ADASYN at 85.18%, Class Weight at 84.82%, and SMOTE with the lowest accuracy of 85.25%.

In Figure 11, the plotted curves depict the training and validation accuracy and losses of the LSTM schemes employed for the seizure detection task. The training accuracy curve showcases the model's performance on the training data as the number of epochs increases, reflecting its learning progress. The ascending training accuracy indicates effective learning from the available data. Conversely, the validation accuracy curve illustrates the model's ability to generalize to unseen data. The decreasing training loss signifies convergence toward an optimal solution as the training progresses. Similar to the validation accuracy, the validation loss provides insight into the model's generalization to new data.

5 Conclusion

This study focused on seizure detection classification using a uni-directional LSTM model and explored the integration of

generative data augmentation techniques. The findings provide valuable insights into the effectiveness of data augmentation approaches, with a particular emphasis on the use of GANs for addressing imbalanced datasets. The investigated GAN models include Vanilla GAN, CRAMER GAN, CGAN, and WGAN-GP, selected for their ability to generate realistic data distributions and improve the representation of minority classes. Additionally, conventional techniques such as SMOTE and ADASYN were also examined. Notably, the most significant improvements were observed with WGAN-GP, which demonstrated a strong alignment and resemblance to the real data distribution in projection analysis and QQ plots. This highlights the effectiveness of WGAN-GP in generating synthetic data that closely matches the real data distribution, thereby enhancing seizure detection performance and improving the classification accuracy of the LSTM classifier. The study serves as a foundation for future investigations and inspires further advancements in seizure detection research. However, additional research and evaluation are required to assess the generalizability of these findings across different datasets and compare the performance against other state-of-the-art approaches that incorporate data augmentation techniques.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://isip.piconepress.com/projects/nedc/html/tuh_eeg/#c_tusz.

Author contributions

LA: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. KH: Formal analysis, Methodology, Software, Validation, Writing – original draft. IJ: Investigation, Methodology, Writing – original draft. NM: Funding acquisition, Project administration, Supervision, Writing – review & editing.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abou-Abbas, L., Henni, K., Jemal, I., Mitiche, A., and Mezghani, N. (2023). Patient-independent epileptic seizure detection by stable feature selection. *Expert Syst. Appl.* 232:120585. doi: 10.1016/j.eswa.2023.120585

Abou-Abbas, L., Jemal, I., Henni, K., Mitiche, A., and Mezghani, N. (2021). "Focal and generalized seizures distinction by rebalancing class data and random forest classification," in *Bioengineering and Biomedical Signal and Image Processing: First International Conference, BIOMESIP 2021* (Meloneras: Springer), 63–70.

Abou-Abbas, L., Jemal, I., Henni, K., Ouakrim, Y., Mitiche, A., and Mezghani, N. (2022). Eeg oscillatory power and complexity for epileptic seizure detection. *Appl. Sci.* 12:4181. doi: 10.3390/app12094181

Alauthman, M., Al-qerem, A., Sowan, B., Alsarhan, A., Eshtay, M., Aldweesh, A., et al. (2023). Enhancing small medical dataset classification performance using gan. *Informatics* 10:28. doi: 10.3390/informatics10010028

Arjovsky, M., Chintala, S., and Bottou, L. (2017). "Wasserstein generative adversarial networks," in *International Conference on Machine Learning* (New York: PMLR), 214–223.

Bellemare, M. G., Danihelka, I., Dabney, W., Mohamed, S., Lakshminarayanan, B., Hoyer, S., et al. (2017). The cramer distance as a solution to biased wasserstein gradients. *arXiv* [preprint] arXiv:1705.10743. doi: 10.48550/arXiv.1705.10743

Bhat, S., and Hortal, E. (2021). "Gan-based data augmentation for improving the classification of eeg signals," in *The 14th PErvasive Technologies Related to Assistive Environments Conference*, 453–458 .

Chang, S., and Jun, H. (2019). Hybrid deep-learning model to recognise emotional responses of users towards architectural design alternatives. *J. Asian Architect. Build. Eng.* 18, 381–391. doi: 10.1080/13467581.2019.1660663

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *J. Artif. Intellig. Res.* 16, 321–357. doi: 10.1613/jair.953

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., and Bharath, A. A. (2018). Generative adversarial networks: an overview. *IEEE Signal Process. Mag.* 35, 53–65. doi: 10.1109/MSP.2017.2765202

Dubey, R., Zhou, J., Wang, Y., Thompson, P. M., Ye, J., Initiative, A. D. N., et al. (2014). Analysis of sampling techniques for imbalanced data: An n= 648 adni study. *Neuroimage* 87, 220–241. doi: 10.1016/j.neuroimage.2013.10.005

Fabio, D. R., Fabio, D., and Carlo, P. (2013). Profiling core-periphery network structure by random walkers. *Sci. Rep.* 3:1467. doi: 10.1038/srep01467

Fisher, R. S., Boas, W. V. E., Blume, W., Elger, C., Genton, P., Lee, P., et al. (2005). Epileptic seizures and epilepsy: definitions proposed by the international league against epilepsy (ILAE) and the international bureau for epilepsy (IBE). *Epilepsia* 46, 470–472. doi: 10.1111/j.0013-9580.2005.66104.x

Ganti, B., Chaitanya, G., Balamurugan, R. S., Nagaraj, N., Balasubramanian, K., and Pati, S. (2022). Time-series generative adversarial network approach of deep learning improves seizure detection from the

human thalamic seeg. *Front. Neurol.* 13:93. doi: 10.3389/fneur.2022.755094

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2020). Generative adversarial networks. *Commun. ACM* 63, 139–144. doi: 10.1145/3422622

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. (2017). "Improved training of Wasserstein Gans," in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*.

Haradal, S., Hayashi, H., and Uchida, S. (2018). "Biosignal data augmentation based on generative adversarial networks," in *2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC)* (Honolulu, HI: IEEE), 368–371.

Hartmann, K. G., Schirrmeister, R. T., and Ball, T. (2018). Eeg-gan: Generative adversarial networks for electroencephalograhic (EEG) brain signals. *arXiv* [preprint] arXiv:1806.01875. doi: 10.48550/arXiv.1806.01875

He, H., Bai, Y., Garcia, E. A., and Li, S. (2008). "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)* (Hong Kong: IEEE), 1322–1328.

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Hussein, R., Palangi, H., Ward, R., and Wang, Z. J. (2018). Epileptic seizure detection: a deep learning approach. *arXiv* [preprint] arXiv:1803.09848. doi: 10.48550/arXiv.1803.09848

Jemal, I., Mezghani, N., Abou-Abbas, L., and Mitiche, A. (2022). An interpretable deep learning classifier for epileptic seizure prediction using eeg data. *IEEE Access* 10, 60141–60150. doi: 10.1109/ACCESS.2022.3176367

Johnson, J. M., and Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *J. Big Data* 6, 1–54. doi: 10.1186/s40537-019-0192-5

Kraiem, M. S., Sánchez-Hernández, F., and Moreno-García, M. N. (2021). Selecting the suitable resampling strategy for imbalanced data classification regarding dataset properties. An approach based on association models. *Appl. Sci.* 11:8546. doi: 10.3390/app11188546

Kuhlmann, L., Lehnertz, K., Richardson, M. P., Schelter, B., and Zaveri, H. P. (2018). Seizure prediction-ready for a new era. *Nat. Rev. Neurol.* 14, 618–630. doi: 10.1038/s41582-018-0055-2

Lashgari, E., Liang, D., and Maoz, U. (2020). Data augmentation for deep-learning-based electroencephalography. *J. Neurosci. Methods* 346:108885. doi: 10.1016/j.jneumeth.2020.108885

Luo, Y., and Lu, B.-L. (2018). "EEG data augmentation for emotion recognition using a conditional Wasserstein GAN," in *2018 40th Annual International Conference*

of the IEEE Engineering in Medicine and Biology Society (EMBC) (Honolulu, HI: IEEE), 2535–2538.

Mirza, M., and Osindero, S. (2014). Conditional generative adversarial nets. *arXiv* [preprint] arXiv:1411.1784. doi: 10.48550/arXiv.1411.1784

Mumuni, A., and Mumuni, F. (2022). Data augmentation: a comprehensive survey of modern approaches. *Array* 16:100258. doi: 10.1016/j.array.2022.100258

Salama, E. S., El-Khoribi, R. A., Shoman, M. E., and Shalaby, M. A. W. (2018). EEG-based emotion recognition using 3D convolutional neural networks. *Int. J. Adv. Comp. Sci. Appl.* 9:43. doi: 10.14569/IJACSA.2018.090843

Shah, V., Von Weltin, E., Lopez, S., McHugh, J. R., Veloso, L., Golmohammadi, M., et al. (2018). The temple university hospital seizure detection corpus. *Front. Neuroinform.* 12:83. doi: 10.3389/fninf.2018.00083

Stafstrom, C. E., and Carmant, L. (2015). Seizures and epilepsy: an overview for neuroscientists. *Cold Spring Harb. Perspect. Med.* 5:a022426. doi: 10.1101/cshperspect.a022426

Ullah, I., Hussain, M., Aboalsamh, H., et al. (2018). An automated system for epilepsy detection using eeg brain signals based on deep learning approach. *Expert Syst. Appl.* 107, 61–71. doi: 10.1016/j.eswa.2018.04.021

Wang, F., Zhong, S.-h., Peng, J., Jiang, J., and Liu, Y. (2018). "Data augmentation for eeg-based emotion recognition with deep convolutional neural networks," in *MultiMedia Modeling: 24th International Conference, MMM 2018* (Bangkok: Springer), 82–93.

Wang, X., Wang, X., Liu, W., Chang, Z., Kärkkäinen, T., and Cong, F. (2021). One dimensional convolutional neural networks for seizure onset detection using long-term scalp and intracranial eeg. *Neurocomputing* 459, 212–222. doi: 10.1016/j.neucom.2021.06.048

Wei, Z., Zou, J., Zhang, J., and Xu, J. (2019). Automatic epileptic eeg detection using convolutional neural network with improvements in time-domain. *Biomed. Signal Process. Control* 53:101551. doi: 10.1016/j.bspc.2019.04.028

World Health Organization (2006). *Neurological Disorders: Public Health Challenges.* Geneva: World Health Organization.

Yuan, Q., Zhou, W., Zhang, L., Zhang, F., Xu, F., Leng, Y., et al. (2017). Epileptic seizure detection based on imbalanced classification and wavelet packet transform. *Seizure* 50, 99–108. doi: 10.1016/j.seizure.2017.05.018

Zhang, P., Zhang, X., and Liu, A. (2022). Effects of data augmentation with the bnnsmote algorithm in seizure detection using 1D-mobilenet. *J. Healthc. Eng.* 2022. doi: 10.1155/2022/4114178

Zhao, X., Sole-Casals, J., Sugano, H., and Tanaka, T. (2022). Seizure onset zone classification based on imbalanced IEEG with data augmentation. *J. Neural Eng.* 19:065001. doi: 10.1088/1741-2552/aca04f