A Comparative Study of Machine Learning Models in Producing Synthetic iEEG Data for Epilepsy Analysis

Shoibolina Kaushik shoibolina.kaushik@emory.edu
Emory University

Aiden Maloney-Bertelli armalo2@emory.edu
Emory University

Editor: N/A

Abstract

The burgeoning need for comprehensive datasets in important domains like neurology, prompts a focus on synthetic data generation techniques to overcome data scarcity. Focusing on the 2014 American Epilepsy Society Seizure Prediction Challenge dataset, methodologies were devised to simulate data scarcity scenarios while validating against the real data. By leveraging Wasserstein Generative Adversarial Networks (WGANs) and Denoising Diffusion Probabilistic Models (DDPMs), synthetic EEG data resembling preictal and interictal states were generated for canine subjects. The quality and utility of this synthetic data were evaluated through comprehensive methodologies encompassing classifier performance, manual qualitative analysis, and comparison against real EEG signals. Results demonstrated convergence of the generative models without severe overfitting, yet highlighted nuanced discrepancies in frequency spectra and time domain characteristics between synthetic and real data. Classifier evaluations revealed varying performance trends as synthetic data proportion in training sets was modified, indicating shortcomings in capturing distinctive features essential for preictal and interictal signal classification. These findings underscore the complexity and limitations in accurately replicating nuanced EEG patterns through current synthetic data approaches. Thus, our results while yielding some successes in generating synthetic data also reveal the limitations in replicating specific features essential for classification tasks, which when subject to future improvements can overcome the potential and current limitations of synthetic data generation.

Keywords: Data scarcity, Synthetic Data Generation, WGANs, DDPMs, iEEG data

1. Introduction

In today's data-driven world, the protection of individual's privacy stands as a crucial priority. Consequently, valuable datasets across numerous fields, including neuroscience and beyond, remain confined, limiting the scope of research and innovation. The challenge of data scarcity presents a significant hurdle, impeding progress in critical areas where insights from comprehensive datasets could pave the way for groundbreaking discoveries.

This project undertakes a pivotal exploration into the realm of synthetic data generation, leveraging the power of machine learning models to address this pressing issue. Our primary goal is to devise methodologies that effectively replicate scenarios characterized by limited data availability, thereby expanding the horizons of research possibilities. This innovative approach aims to bridge the gap between the paucity of accessible data and the growing demands of researchers and practitioners in various domains.

To accomplish this, we turn our focus towards the 2014 American Epilepsy Society Seizure Prediction Challenge dataset available on Kaggle., recognized as a benchmark within the field. By harnessing the resources available on platforms like Kaggle, we intend to strategically curate a subset of this dataset, simulating conditions akin to a scarcity of information while retaining the option to validate our methodologies against the complete dataset. This approach provides a safety net for our exploratory methods, allowing us to validate and refine our techniques against a comprehensive dataset while still being adaptable to genuine data-scarce situations.

Our study aims to provide a comprehensive exploration of the methodology employed, the challenges encountered, and the potential implications of utilizing synthetic data in enhancing research capabilities in the field of epilepsy and beyond. Epilepsy, characterized by recurrent seizures, presents a significant health concern globally. To address this challenge, Electroencephalography (EEG) has emerged as an indispensable non-invasive tool for both diagnosis and management of the condition. By analyzing EEG data, distinct patterns corresponding to pre-ictal (before seizure onset) and ictal (during seizure) states have been identified. These patterns hold vital information that could potentially revolutionize our understanding of epilepsy, offering insights into its predictive markers and aiding in the development of more effective treatment strategies. Through these concerted efforts, we aim not only to fill the void caused by limited data availability but also to pave the way for a new paradigm in research methodologies. By expanding and enriching smaller datasets, we endeavor to empower researchers and practitioners across diverse domains to glean meaningful insights and drive unprecedented advancements, ultimately propelling the frontiers of knowledge and innovation forward.

2. Related Work

Carrle et al. (2023) systematically reviewed EEG data augmentation studies and also performed and evaluated their own GAN made up of CNNs for synthetic EEG creation, albeit for aiding classification of patients with major depressive disorder and healthy controls. Their review covered 27 papers, five of which pertained to epilepsy and two of which, like our group, aimed to predict upcoming seizures. By far, the most common method of synthetic EEG generation was GAN-based approaches, though two papers used generative pre-trained transformers (GPTs) and one used a variation autoencoder (VAE). Several published GAN models were compared against VAEs and found to be superior. The authors noted that GANs with RNN architectures seem to be an intuitive fit for EEG data but are underexplored compared with CNN architectures. In the authors' own implementation, they used the Wasserstein distance as a loss funciton in their GAN, which is frequent for EEG data generation. Augmentation with their synthetic data improved diagnostic accuracy by about 10% for one of their datasets but insignificantly for the other dataset, and they noted that qualititative examination of the frequency spectra of the synthetic data revealed shortcomings not apparent in the time domain.

The shortage of high-quality data in electroencephalography (EEG) analysis poses a significant challenge for accurate seizure prediction. To address this issue, Rasheed et al. (2021) proposed a deep convolutional generative adversarial network (DCGAN) which generates synthetic EEG data, enhancing prediction performance. The authors conducted patient-specific training of the DCGAN using real EEG data and validated the generated data's quality using one-class SVM and a unique method named convolutional epileptic seizure predictor (CESP). Furthermore, evaluation of VGG16, VGG19, ResNet50, and Inceptionv3 models, trained on augmented data using transfer learning (TL) with a 10-minute average temporal gap between true prediction and seizure onset samples, revealed noteworthy results. Their CESP model yielded a sensitivity of 78.11% with a false prediction rate of 0.27/h when tested on Epilepsyecosystem dataset and a sensitivity of 88.21% with a false prediction rate of 0.14/h when tested on the CHB-MIT dataset. Inceptionv3, leveraging TL and augmented data, attained the highest accuracy with a sensitivity of 90.03% and a 0.03 FPR/h. Their proposed data augmentation technique notably improved prediction outcomes for both CESP and Inceptionv3 models, highlighting the effectiveness of synthetic data in surpassing chance levels on augmented datasets.

3. Methodology

3.1 Data Description

The original dataset comprises EEG recordings from five dogs and two human patients, categorized as preictal (before a seizure) and interictal (away from seizure activity). These recordings are segmented into ten-minute intervals and annotated by epileptologists. For the dogs, the recordings span multiple months and were obtained from 16 leads at a 400 Hz sampling rate. The voltage measurements are referenced to a group average. In contrast, the human subjects' data, collected over varying spans (sometimes less than a week), involved either 15 leads (Patient 1) or 24 leads (Patient 2) at a 5000 Hz sampling rate, referenced to an electrode outside the brain.

Each ten-minute segment originates from a continuous hour-long sequence, with the data file indicating its position (indexed 1-6) within the hour of recording. Preictal recordings consist of hours leading up to a seizure event, ranging from one hour and five minutes before the seizure to five minutes before it. The five-minute gap (pre-seizure horizon) is excluded to minimize annotation errors and facilitate seizure prediction with adequate lead time for appropriate interventions. Interictal recordings were sampled deliberately distant from seizure events, occurring more than a week apart from seizures in dogs and more than four hours apart in humans.

Given the considerable size of the dataset and the inherent variability in electrode usage between dog and human subjects, our study concentrates solely on the canine EEG data. The fifth dog was excluded due to having one fewer channels of recordings than the other dogs' data. The training dataset now consists of 3224 interictal samples and 235 pre-ictal samples, while the test dataset contains 3399 unlabeled samples. Each of the considered subject's data distribution across the segments is as detailed in Table 1. This focused approach allows us to streamline our analysis.

Subject	Interictal	Preictal	Test
Dog 1	480	24	502
Dog 2	500	42	1000
Dog 3	1440	72	907
Dog 4	804	97	990
Total	**3224**	**235**	**3399**

Table 1: Dataset used for our study

3.2 Data Splitting

With each sample being considered a 10-minute recording segment, the samples were split in a stratified manner on the basis of class (interictal vs. preictal) into 70% for training the synthetic data generators and 30% for evaluating the synthetic data. From the 70% for training, 60 10-minute recordings were randomly sampled from each class to serve as a training set and 6 recordings from each class were held out as a validation set for the WGAN due to resource constraints. For DDPMs, the interictal class underwent downsampling to match the size of the preictal class. Additionally, 10% of the either data was allocated for the purpose of models' validation. Within the 30% used for evaluating the synthetic data, the data were split, again stratified by class, into 67% for training and 33% for testing. Within the 67% for training, the interictal class was undersampled to match the size of the preictal class, but the test set was left with the same class imbalance as the original dataset to preserve an understanding of generalizability. 5-fold cross-validation was used for hyperparameter optimization for the classification models used for evaluation.

After this initial splitting by 10-minute recording segment, the recordings were split into roughly 10-second segments (4000 samples at about 400 Hz) without overlap. Table 2 shows the numbers of 10-second samples used for each synthetic data generator as well as for the evaluation classifiers.

	Training Samples		**Validation Samples**		**Test Samples**	
Model	**Interictal**	**Preictal**	**Interictal**	**Preictal**	**Interictal**	**Preictal**
Interictal WGAN	3,540	-	354	-	-	-
Preictal WGAN	-	3,540	-	354	-	-
Interictal DDPM	-	8549	-	950	-	-
Preictal DDPM	8549	-	950	-	-	-
Evaluation Classifiers	2,714	2,714	5-fold	5-fold	18,939	1,357

Table 2: Numbers of samples used for training, validating, and testing each model. Each sample is a 10-second EEG segment.

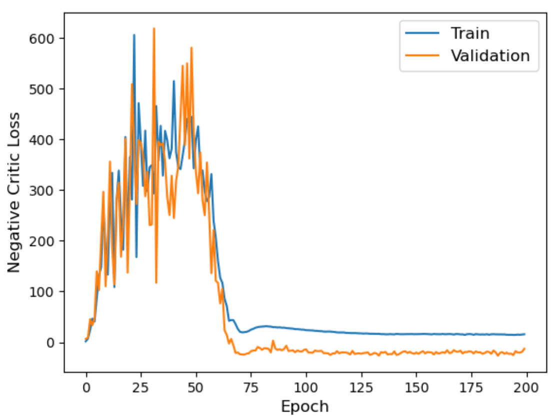
3.3 Wasserstein Generative Adversarial Network

The Wasserstein generative adversarial network (WGAN) was introduced by Arjovsky et al. (2017) to mitigate the traditional GAN's issues of stability, mode collapse, and lack of learning curve interpretability. The loss function of the WGAN is an estimate of the earth-mover (EM) distance, which is a measure of how much "mass" would need to be transported to transform the probability distribution of one random variable to match another. In practice, this means turning the "discriminator" of a GAN into a "critic" that does not classify between real and fake samples but rather is trained to have a higher score for real outputs than fake outputs. This is done by using a linear activation for the final dense layer of the critic, a loss function for the critic of $C(fake) - C(real)$ where C is the critic model, and a loss function for the generator of $-C(fake)$. In other words, the critic is trying to maximize the difference in scores between real and fake data while the generator is trying to maximize the critic's score for fake samples. The aim during training is to reach the critic's optimality.

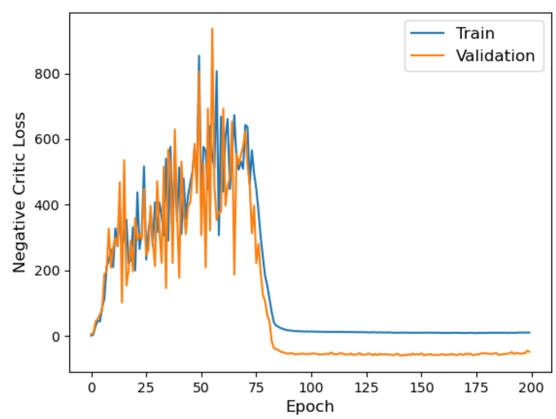
Here, we built upon a basic DCGAN architecture described in ten. The generator took an input vector of length 100 and fed it through a dense layer, then through three 1D convolutional transpose layers with some reshaping as well as batch normalization and leaky ReLu layers between each dense/convolutional layer. A hyperbolic-tangent activation function was used for the final convolutional layer. The critic was composed of two 2D convolutional layers (though each only convolved across the time dimension), each followed by leaky ReLu and dropout layers, with a final dense layer. As suggested for WGAN, weight clipping was used with a cutoff of 0.01 for the convolutional layers in the critic. Additionally, a hyperparameter was created for the number of times the critic was trained per time the generator was trained for each mini-batch, and the RMSProp optimizer was used.

Two separate WGANs were trained: one for generating interictal samples, and one for generating preictal samples. The input data were scaled per-channel by the maximum absolute values in the training set. Centering was not done since presumably the data is already "centered" by the use of a reference electrode. For hyperparameter tuning, loss curves were compared over 50 epochs for a grid search over learning rates of 2e-5 and 5e-5 as well as numbers of times the critic was trained per mini-batch of 1, 2, and 5. Mini-batch size was set at 16. The critic loss remained lowest and the generated signals appeared visually the best after 50 epochs for a learning rate of 2e-5 and critic-training-ratio of 1, so these were chosen as the hyperparameters going forward. Hyperparameters were tuned based on training for interictal samples, and the same hyperparameters were used for preictal samples since the classes are assumed to share some core characteristics. However, separate tuning should be explored in future work. For the best parameters, training was then done over 200 epochs, and critic loss appeared to reach a roughly state without divergence of the validation loss for both the interictal and preictal data as shown in Figure 1.

Once the final models were trained, synthetic data was generated by feeding batches of vectors of length 100 containing noise. These vectors are considered samples from the latent space from which the WGAN could generate 10s synthetic EEG signals.



(a) WGAN for preictal data generation



(b) WGAN for interictal data generation

Figure 1: Training and validation loss curves for WGAN

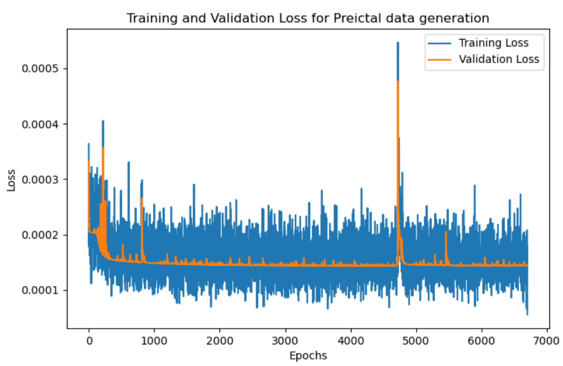
3.4 Denoising Diffusion Probabilistic Model

The Denoising Diffusion Probabilistic Model (DDPM) is a generative model used for modeling and generating high-dimensional data, particularly suited for tasks involving image, audio, or sequential data. It leverages a diffusion process to learn the data distribution and generate realistic samples from it, as described by Ho et al. (2020). The fundamental idea behind DDPM involves iteratively applying a diffusion process to a noise distribution to gradually approximate the true data distribution. The model learns this process during training, enabling it to generate high-quality synthetic data samples.

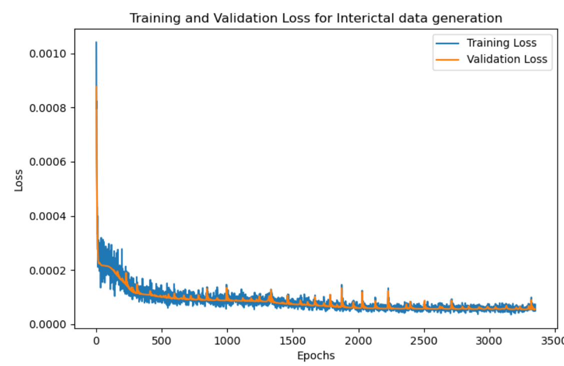
The DDPM architecture in this project is defined using a neural network with an encoder-decoder structure. Both the encoder and decoder each consist of two 1D convolutional layers with ReLU activations. The decoder is followed by a final layer with a Sigmoid activation to output values between 0 and 1. Two different DDPMs were trained for preictal and interictal data generation.

During training, for both preictal and interictal, the training data is converted into PyTorch tensors and loaded into DataLoader objects for batch processing. Adam optimizer is used to update the model parameters and Mean Squared Error (MSE) loss is chosen as the criterion for reconstruction. The training loop runs for the specified number of epochs during which the model is first set to train mode and training data is fed in batches to the model. Next, loss is calculated, backpropagation is performed, and model parameters are updated. Additionally, validation is performed using the validation dataset, and validation loss is computed. Among hyperparameters, 32 filters, 16 channels, and 25 epochs were used for both preictal and interictal data generation. However, the batch size and learning rate differed, with $batch\_size = 32$ and $learning\_rate = 0.01$ for preictal data generation, while $batch\_size = 64$ and $learning\_rate = 0.001$ for interictal data generation.

Training and validation losses are recorded for each epoch and stored in separate lists, which are plotted to track the training progress during each epoch. Finally, synthetic data samples are generated using the trained model by passing random noise through the decoder. Figure 2 show the training and the validation curve for the model while preictal and interictal data generation.



(a) DDPM for preictal data generation



(b) DDPM for interictal data generation

Figure 2: Training and validation loss curves for DDPM

3.5 Classification Models for Evaluation

To evaluate the synthetic data, we first compared the performances of a Naïve Bayes classifier, Ridge classifier, and AdaBoost classifier for classifying between interictal and preictal samples based on extracted features, trained and applied to real data. Then, using the best models with hyperparameters optimized based on real training data, we measured the classification performance when varying portions of the real training data were replaced with synthetic data. This evaluation technique was drawn from the methods for evaluating synthetic data outlined by Lu et al. (2023). Standard ML methods as opposed to convolutional networks were used for evaluation since convolutional methods were used for synthetic data generation. While performance of convolutional networks trained on synthetic data would also be an indicator of synthetic signal quality, classification based on interpretable features could provide better insight into whether the synthetic data captures known, measurable characteristics of preictal and interictal signals.

For feature extraction, we first split each 10-second sample into 1s windows with 0.5s overlap, totaling 19 windows per sample. Further preprocessing was omitted since we aimed for the synthetic data to mimic real data, including any noise patterns. For each window and for each of the 16 channels, we extracted 14 features: mean, standard deviation, absolute sum of changes, minimum, maximum, skewness, kurtosis, root mean square, root absolute energy, power in the delta band (1-4 Hz), power in the theta band (4-8 Hz), power in the alpha band (8-14 Hz), power in the beta band (14-30 Hz) and power in the gamma band (30-200 Hz). The *tsfresh* Python library was used for extracting all features except the band powers. Najafi et al. (2022) extracted these same features for classifying between focal and generalized epilepsy and found multiple to have statistically significant correlations with class labels. While we are performing a different classification task, we at least know these features distinguish different types of EEG signals. With these 14 features for each of 16 channels and 19 windows per sample, there were a total of 4,256 features per 10-second sample. Of the 4,256 features, the 20% with the lowest rank correlation with the labels within the real training set were dropped, leaving 3,404 features.

Hyperparameter optimization was done using 5-fold cross-validation with the real training data optimizing the macro-averaged F1-score. For the ridge classifier, the search was across $\alpha = [10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}]$. For the AdaBoost classifier, a grid search was done across learning rates of $[10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}]$ and maximum numbers of estimators of $[5, 20, 50, 100]$. We used the macro-averaged F1-score and AUC to determine the best classifier(s). These grid searches and models were implement using the *scikit-learn* Python library.

The best-performing model(s) were then retrained and evaluated with varying ratios of real and synthetic data from the two different synthetic data generation approaches (DDPM and WGAN). Ideally, the model performance would be steady across different ratios of real and synthetic data. Substantial increases or decreases in model performance would both suggest the synthetic data does not closely resemble the real data.

3.6 Manual Evaluation

Visual comparison of real and synthetic signals both in the time domain and frequency was also performed for quality analysis. Carrle et al. (2023) systematically reviewed EEG data augmentation studies and noted the importance of examining the frequency spectra of synthetic data as they can reveal shortcomings not apparent in the time domain. To compare the frequency spectra, we generated spectrograms covering 0-40 Hz, as these frequencies cover the delta, theta, alpha, and beta bands as well as part of the gamma band, for 10-second windows. Spectrograms were plotted with 256 datapoints (0.64 seconds) per block and 128 points (0.32 seconds) of overlap between blocks.

4. Results

Both the DDPM models and WGAN models appeared to converge without noticeable overfitting, which might show up as divergent losses for validation data, as shown in Figures 1 and 2. For the DDPM, losses converged to roughly the same level for training and validation data, whereas for the WGAN, the Wasserstein losses converged to separate values for training and validation data. Since the absolute value of the loss converged closer to zero for the training data than the validation data for the WGAN, the critic gave more different scores between validation data and synthetic data than it did between the training data and synthetic data. This suggested the synthetic data generated would still be meaningfully different from the real data.

Qualitative analysis in both the time and frequency domains also reveals noticeable shortcomings of the synthetic data. Graphs of the time domain signals are in Appendices A-F (6-6), and spectrograms are shown in Figure 3. Synthetic interictal data generated with the WGAN is noticeably lacking in in strength for frequencies under 10 Hz, but the range of general amounts of noise apparent in the time domain are consistant with real signals. The spectrogram also shows non-stationary properties for both the interictal and preictal WGAN synthetic signals with somewhat smooth transitions across time as seen in the spectrograms for real signals. More high-frequency spikes are apparent in the synthetic preictal WGAN data than in the synthetic interictal WGAN data as expected; however, the synthetic data can be lacking in broad high-frequency content. The DDPM synthetic data is starkly noisier with higher variance than the real data and the WGAN data but appears closer to the real data for low frequencies as shown in the spectrogram. For frequencies of 5 Hz and above, the power spectra appear somewhat randomly distributed than for the real and WGAN data. The DDPM preictal data does have more high-frequency content than the DDPM interictal data, which is consistent with real data.

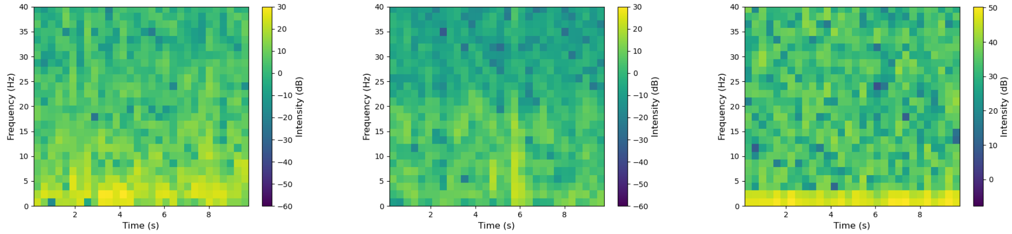
Finally, training on combinations of real and synthetic data revealed mostly slight declines in classifier performance as the fraction of training data that was synthetic was increased, as shown in Figure 4. Initial comparison of Naïve Bayes, ridge classification, and AdaBoost trained on only real data yielded the best macro-averaged F1-score and precision for Naïve Bayes but the best binary F1-score, recall, and AUC for AdaBoost, so we used both models for evaluating the synthetic data. AdaBoost maintained largely consistent performance until more than about 0.6 of the training data was synthetic, after which performance started to steeply decline. Naïve Bayes performance was better for the WGAN data than for the DDPM data, but both reached their highest AUCs when there was a roughly equal split of real and synthetic data in the training set. Overall, this suggests the synthetic data lacks some of the features that are useful for classifying between preictal and interictal signals.

Model	Macro F1	Binary F1	Precision	Recall	AUC
Naïve Bayes	**0.54**	0.18	**0.13**	0.26	0.63
Ridge	0.46	0.17	0.10	0.62	-
AdaBoost	0.50	**0.21**	0.13	**0.70**	**0.73**

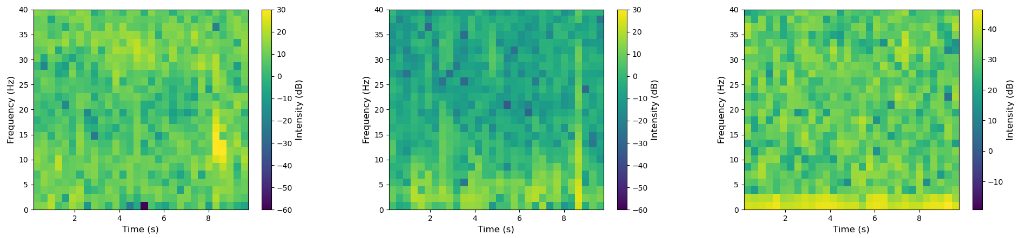
Table 3: Comparison of three models optimized, trained, and tested with real training data. For binary metrics, the preictal class was considered positive, and the interictal class was considered negative. The best result for each metric is in bold.

5. Discussion

For this project, GANs and DDPMs were selected as synthetic data generation models for EEG signals due to their adeptness in capturing complex patterns, spatial-temporal dependencies, and underlying data distributions present in EEG data. GANs, through their adversarial training mechanism, aim to generate samples that closely resemble real EEG data by training a generator to produce realistic signals and a discriminator to distinguish between real and generated samples. On the other hand, DDPMs model the diffusion process within the data, learning the intricate temporal dynamics and dependencies to generate synthetic EEG-like signals. While Large Language Models (LLMs) like GPT



(a) Real, Interictal (b) WGAN, Interictal (c) DDPM, Interictal



(d) Real, Preictal (e) WGAN, Preictal (f) DDPM, Preictal

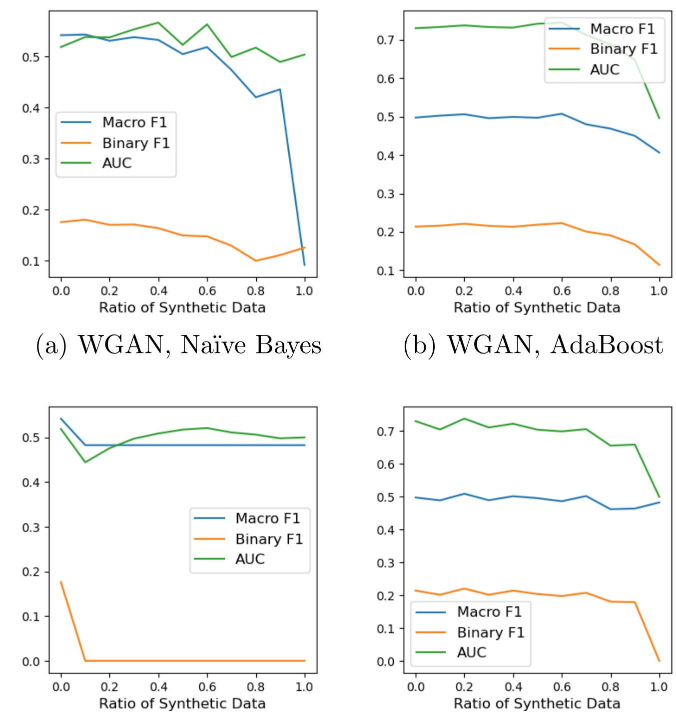
Figure 3: Spectrograms of real and synthetic, interictal and preictal signals from the second channel. Each spectrogram window covers 256 samples with 128 samples of overlap. The sampling rate was 399.6 Hz.

excel in text generation tasks, their adaptation for EEG data might be challenging due to the fundamental differences between textual data and the multi-dimensional, complex nature of EEG signals, making GANs and DDPMs more suitable choices for synthesizing realistic EEG data.

Having longer classification windows can potentially result in poorer classification outcomes due to the inclusion of diverse temporal patterns, loss of temporal specificity, increased noise and irrelevant information, heightened model complexity leading to overfitting, and the violation of stationarity assumptions within the data. Lengthy windows might aggregate varied patterns, making it challenging for models to discern relevant features while increasing the likelihood of including irrelevant or noisy data. Furthermore, the loss of temporal precision within extended windows might obscure crucial event timing essential for accurate classification. Thus, striking a balance between capturing pertinent information and avoiding noise by determining an optimal window length is pivotal in optimizing classification performance and ensuring effective feature extraction for complex data like EEG signals.

6. Conclusion

Here, we trained two DDPMs and two WGANs (one of each model type for preictal signals and one of each type for interictal signals) that generated synthetic 16-channel, 10-second iEEG samples sampled at about 400 Hz based on training data from four dogs and were able to demonstrate model convergence. Approximate signal ranges and time domain characteristics were captured, but there were still apparent differences between real and synthetic signals in both the time domain and spectral domain. Additionally, performance of classification models for classifying interictal and preictal signals generally declined as the



(a) WGAN, Naïve Bayes (b) WGAN, AdaBoost

(c) DDPM, Naïve Bayes (d) DDPM, AdaBoost

Figure 4: Classification performance for different ratios of real and synthetic data evaluated on real data. Models were trained to classify preictal and interictal signals with preictal considered the positive class.

fraction of training data that was synthetic increased, particularly above 0.6. However, the best classification performances we observed with entirely real training data was a macro-averaged F1-score of 0.54 with Naive Bayes and an AUC of 0.73 with AdaBoost, suggesting perhaps the classes are not very distinguishable for 10-second intervals. We encourage further work to explore WGAN behavior over more epochs across more hyperparameter combinations as well as to incorporate LSTM and/or other models specialized for sequence data to be within a WGAN and for evaluation. Extraction of additional features as well as exploration of EEG-specific preprocessing techniques could also be fruitful.

Contributions

The authors contributed equally. S. Kaushik explored preprocessing methods and developed the DDPM model. A. Maloney-Bertelli coded the classification models for evaluation as well as the WGAN.

Code

https://github.com/arberetum/synthetic_eeg

Dataset

bbrinkm, Will Cukierski. (2014). American Epilepsy Society Seizure Prediction Challenge. Kaggle. https://kaggle.com/competitions/seizure-prediction

Appendix A.

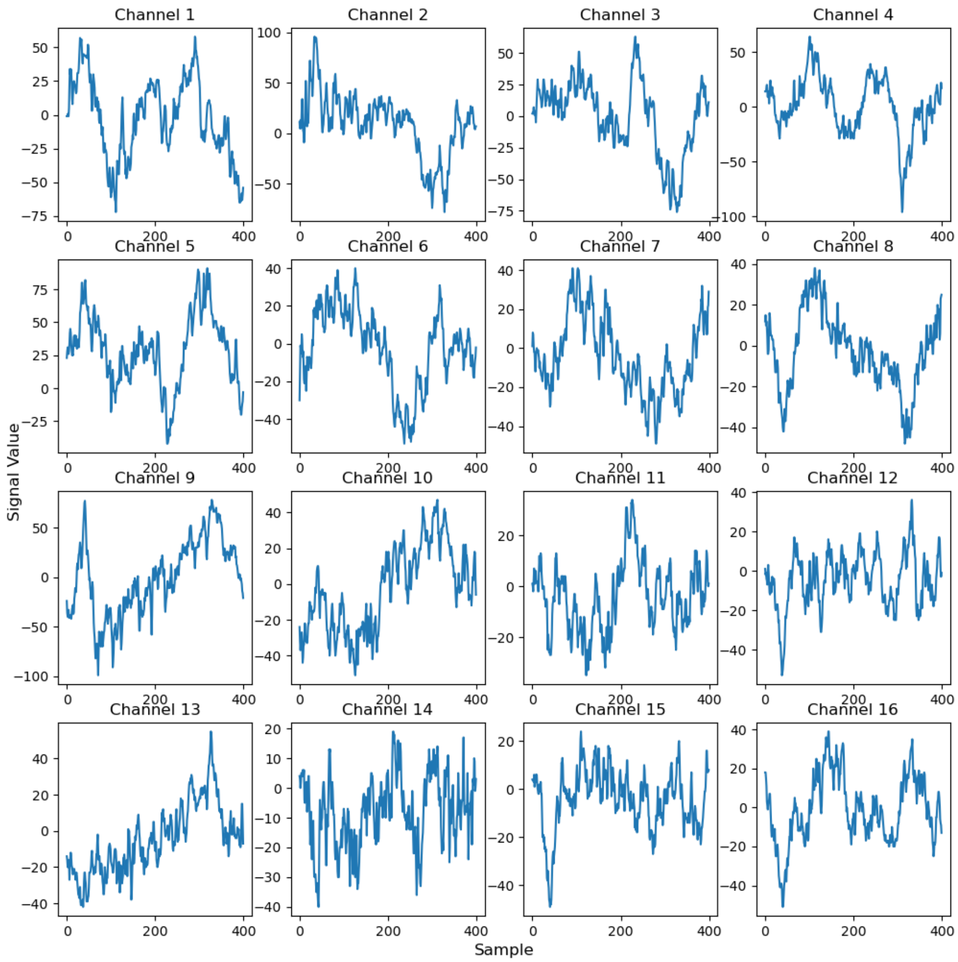


Figure 5: One second of real interictal data in the time domain.

Appendix B.

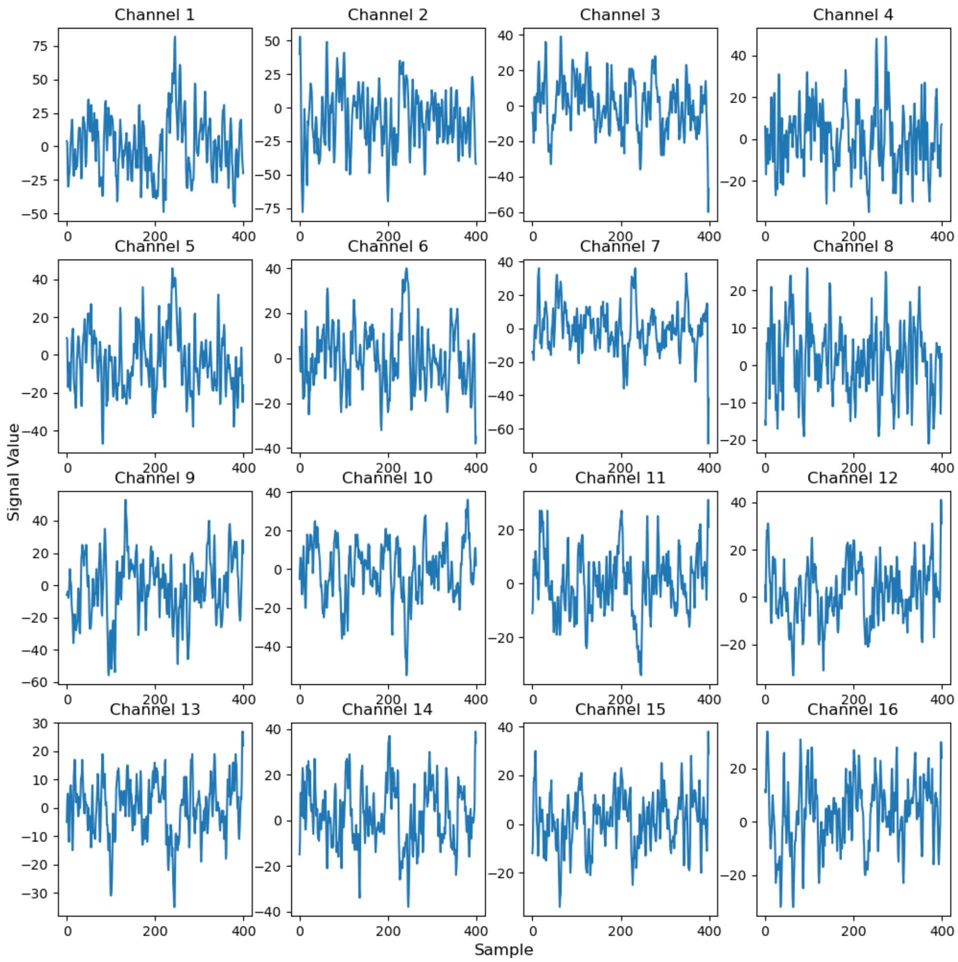


Figure 6: One second of real preictal data in the time domain.

Appendix C.

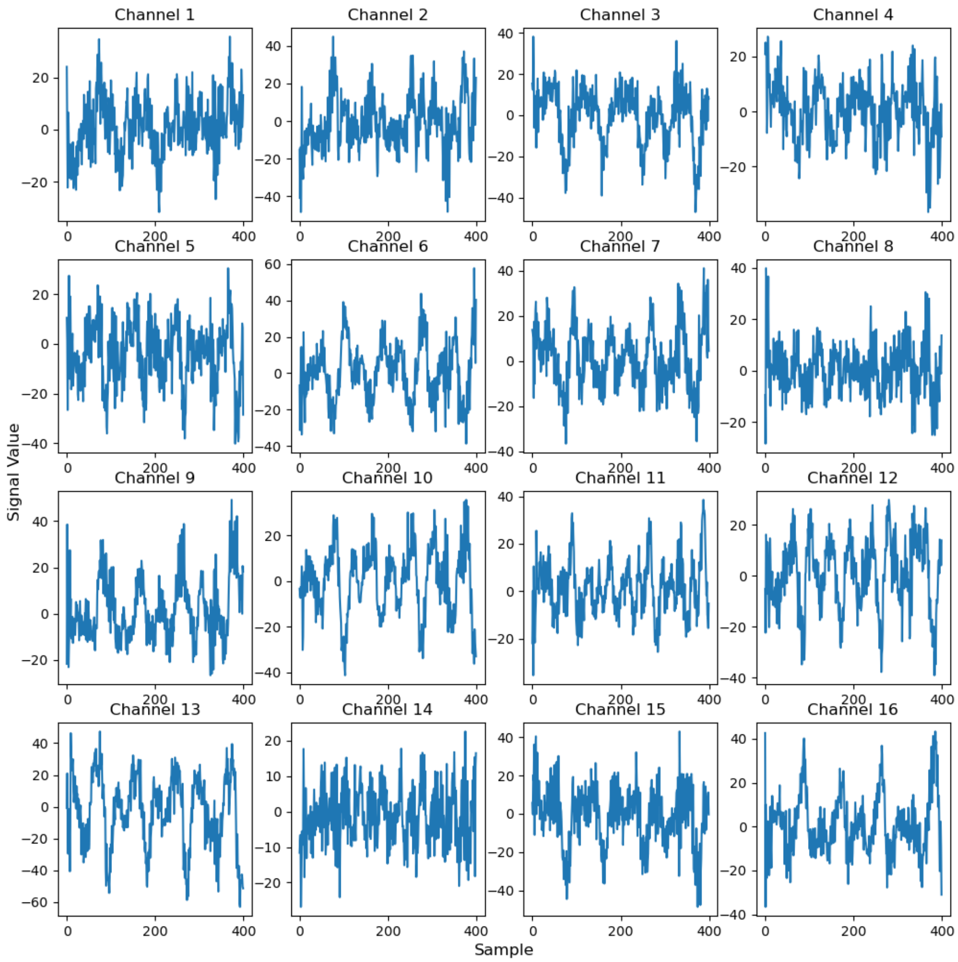


Figure 7: One second of synthetic interictal data generated with the WGAN.

Appendix D.

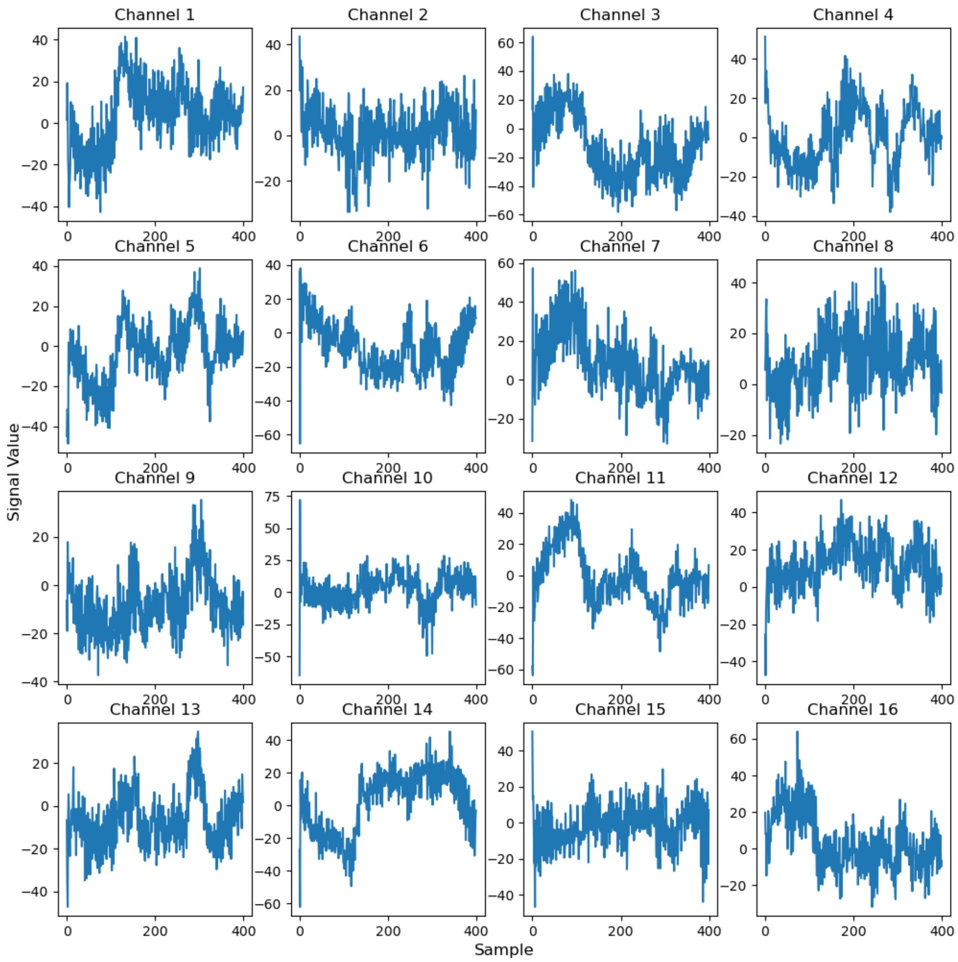


Figure 8: One second of synthetic preictal data generated with the WGAN.

Appendix E.

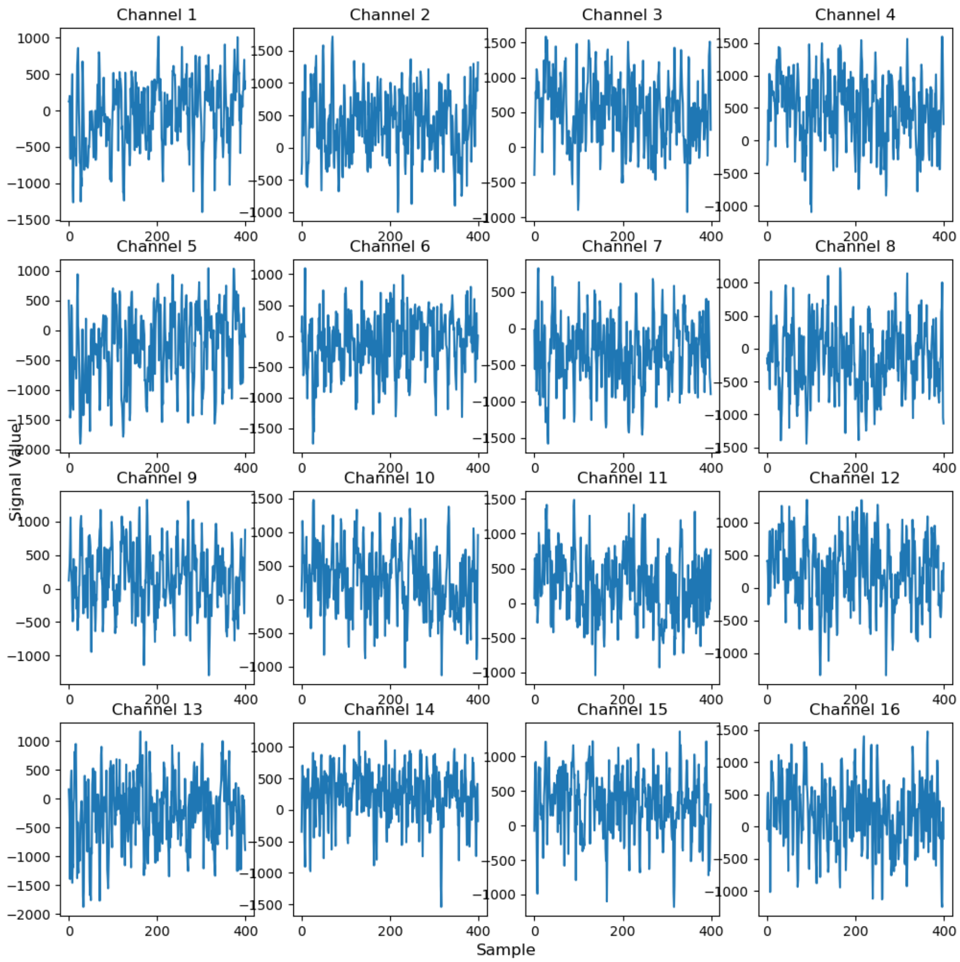


Figure 9: One second of synthetic interictal data generated with the DDPM.

Appendix F.

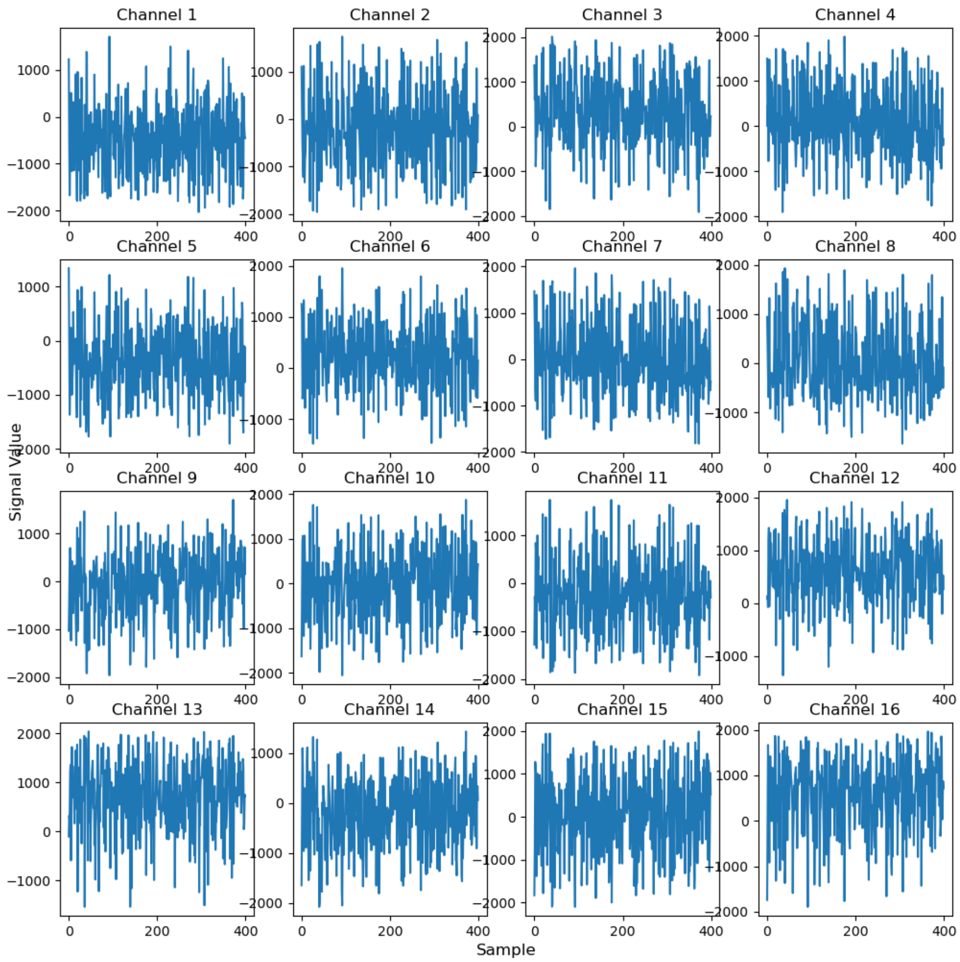


Figure 10: One second of synthetic preictal data generated with the DDPM.

References

Deep convolutional generative adversarial network. `https://www.tensorflow.org/tutorials/generative/dcgan`. Accessed: 2023-12-13.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR, 06–11 Aug 2017. URL `https://proceedings.mlr.press/v70/arjovsky17a.html`.

Friedrich Philipp Carrle, Yasmin Hollenbenders, and Alexandra Reichenbach. Generation of synthetic eeg data for training algorithms supporting the diagnosis of major depressive disorder. *Frontiers in Neuroscience*, 17:1219133, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf`.

Yingzhou Lu, Huazheng Wang, and Wenqi Wei. Machine learning for synthetic data generation: a review. *arXiv preprint arXiv:2302.04062*, 2023.

Tahereh Najafi, Rosmina Jaafar, Rabani Remli, and Wan Asyraf Wan Zaidi. A classification model of eeg signals based on rnn-lstm for diagnosing focal and generalized epilepsy. *Sensors*, 22(19): 7269, 2022.

Khansa Rasheed, Junaid Qadir, Terence J. O'Brien, Levin Kuhlmann, and Adeel Razi. A generative model to synthesize eeg data for epileptic seizure prediction. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:2322–2332, 2021. doi: 10.1109/TNSRE.2021.3125023.