MEDiC: Mitigating EEG Data Scarcity Via Class-Conditioned Diffusion Model

Gulshan Sharma
Indian Institute of Technology Ropar, India

Abhinav Dhall
Indian Institute of Technology Ropar, India
Monash University, Australia

Ramanathan Subramanian
University of Canberra, Australia

Abstract

Learning with a small-scale Electroencephalography (EEG) dataset is a non-trivial task. On the other hand, collecting a large-scale EEG dataset is equally challenging due to subject availability and procedure sophistication constraints. Data augmentation offers a potential solution to address the shortage of data; however, traditional augmentation techniques are inefficient for EEG data. In this paper, we propose *MEDiC*, a class-conditioned Denoising Diffusion Probabilistic Model (DDPM) based approach to generate synthetic EEG embeddings. We perform experiments on a publicly accessible dataset. Empirical findings indicate that *MEDiC* efficiently generates synthetic EEG embeddings, which can serve as effective proxies to original EEG data.

1 Introduction

Over the past few years, there has been rapid growth in the applications of Artificial Intelligence (AI) in the healthcare sector. From diagnosing diseases through medical imaging to personalized treatment plans, generative AI-based solutions are transforming modern healthcare practices [1]. Within this transformation landscape, medical imaging is one area where generative AI has made significant progress [2]. EEG is a *brain-imagining* technique applied to diagnose various neurological disorders [3]. Unlike traditional imaging methods such as X-rays or MRI, which produce anatomical images; EEG measures the fluctuations in electrical potentials generated by neurons. It has a high temporal resolution, which allows real-time analysis of brain activity [4, 5]. It also serves as a well-established clinical tool for *automatic* diagnosis of neurological conditions such as epilepsy, sleep disorders, and neurodegenerative disorders like Alzheimer's disease [6].

The automated diagnosis via EEG signals primarily depends on classification techniques. These methods identify complex patterns within EEG signals and categorize them into normal vs. abnormal classes [7]. However, interpreting EEG data is challenging due to its highly dynamic nature and susceptibility to various artifacts [4, 8]. To address these issues, a substantial amount of research efforts have been directed toward developing Deep Neural Network (DNN) based classification models. DNN can effectively recognize EEG patterns but requires a large amount of data to achieve generalization [9]. EEG datasets are relatively small in comparison to other data modalities, such as text, image, and audio [10]. This is due to the complex and resource-intensive process involved in collecting EEG data. Collecting EEG data, particularly from individuals with neurological disorders, necessitates controlled experimental conditions, making the process more challenging. As a result, researchers frequently face generalization constraints when it comes to including DNN in their studies.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Methods such as data augmentation are useful in addressing the EEG data scarcity [11]. With data augmentation, one can generate synthetic EEG samples which exhibit properties similar to the original EEG data [12]. It is a cost-effective and promising solution which can improve DNN's performance and reduce the risk of overfitting by increasing the diversity in training data. In this paper, we present *MEDiC*, an augmentation framework capable of generating EEG latent embeddings comparable to real EEG data. Our research contributions are summarized below:

- Unlike traditional EEG augmentation methods, we consider latent EEG embeddings as a descriptor of EEG signals and perform augmentation based on these embeddings.
- We propose a class-conditioned Denoising Diffusion Probabilistic Model (DDPM) based approach for generating synthetic EEG embeddings for data corresponding to Alzheimer's disease, Frontotemporal Dementia, and Control Group classes.
- We validate the quality of synthetic EEG embeddings by measuring their ability to maintain class discrimination. Furthermore, we perform a similarity check via Jensen-Shannon Divergence scores. Additionally, we compare the fidelity of synthetic EEG embeddings generated via class-conditioned DDPM and VAE.

2 Background

2.1 EEG Data Augmentation

The majority of EEG augmentation methods can fall into (a) time-domain, (b) frequency-domain, (c) spatial-domain, and (d) latent feature space augmentation. The time-domain augmentation involves applying transformations such as time-shifting, amplitude scaling, filtering, waveform warping, and noise injection directly to the EEG time series [12]. Frequency-domain augmentation involves modifying the frequency content of the signals via applying frequency filtering, frequency warping, or spectral interpolation methods [13]. Spatial-domain augmentation involves changing the spatial distribution of EEG electrodes via channel interpolation and dropout [14]. Latent feature space augmentation enhances the EEG dataset by generating new samples within the learned latent space. Since EEG recordings are highly prone to artifacts which distort the neural information, leading to inaccurate diagnostic conclusions. Latent space provides a compact and enhanced signal-to-noise ratio proxies while retaining its essential information [15, 16].

2.2 Denoising Diffusion Probabilistic Model

DDPM falls under the class of score-based generative models, which learn the underlying data distribution by estimating the gradients of the log-likelihood of the data [17, 18]. DDPM comprises two sub-processes: forward diffusion and backward diffusion. The forward diffusion introduces noise to the input data via a predetermined noise scheduler until the data transforms into pure noise. On the other hand, the backward diffusion aims to reconstruct the original data from the noise. It begins with the noisy data obtained at the end of the forward diffusion process, and at each time step, a neural network predicts how much noise needs to be removed to return to the previous step. While generating data samples belonging to a specific class, model conditioning enables the incorporation of class-specific constraints into the generative process [19]. In context to DDPM, classifier guidance [20] and classifier-free guidance [21] are two main approaches for model conditioning.

3 Dataset

In this paper, we examine a publicly available EEG dataset which comprises EEG recordings of 88 participants, categorized into three groups: 36 individuals diagnosed with Alzheimer's disease (AD group), 23 diagnosed with Frontotemporal Dementia (FTD group), and 29 healthy subjects (CN group). The resting state EEG signals are collected via Nihon Kohden EEG 2100 clinical device, applying 19 scalp electrodes positioned according to the 10-20 international system. In our study, we use the pre-processed version of this dataset, which is publicly accessible. For more detailed information, interested readers may refer to the references [22, 23].

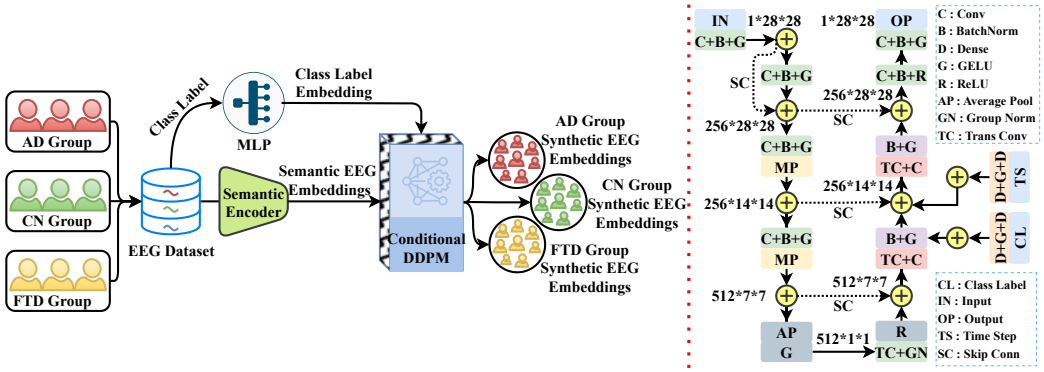

Figure 1: MEDiC framework: EEG data is input to a semantic encoder to extract EEG embeddings. These embeddings, along with class label encodings, are then input into the class-conditioned DDPM to generate synthetic EEG embeddings (left). Conditional U-Net architecture used during backward diffusion process (right).

4 Method

Our proposed approach can be divided into two stages: (a) Training a semantic encoder which removes irrelevant details and learns a latent EEG space semantically equivalent to the original EEG data, and (b) Developing a class-conditioned DDPM which operates within the latent EEG space and has the capacity to generate any number of synthetic EEG samples.

4.1 Latent EEG Embedding Generation

In the first stage, we begin by dividing EEG signals into 30-second segments with 25% overlap between consecutive segments. We assume that 30-second EEG data is sufficient to learn discriminative features across the AD, FTD, and CN groups. Afterward, we apply EEGNet [9] as a semantic encoder network on these segments. EEGNet is a neural network designed to learn class-disentangled representations from EEG signals. In our implementation, we train EEGNet on categorical cross-entropy loss with Adam optimizer, learning rate set to 0.0002. We continue the training process until EEGNet reaches its convergence point. We ensure that EEGNet learns well-discriminative features from the original EEG signals by performing a 5-fold cross-validation procedure. Finally, we extract the 784-dimensional latent embeddings from the feature representation layer of the EEGNet, resulting in 1265, 718, and 1048 embedding samples corresponding to AD, CN, and FTD groups. These latent embeddings serve as the basis for our augmentation method.

4.2 Class-Conditioned DDPM Development

In the second stage, we develop a class-conditioned DDPM tailored to train on the 784-dimensional latent EEG embeddings extracted via the EEGNet. In our implementation, the forward diffusion process comprises a predefined noise scheduler which returns pre-computed schedules for DDPM sampling. For the backward diffusion, we implement a simple U-Net architecture, where class conditioning is based on the classifier-free diffusion guidance process [21]. This U-Net incorporates context and time-step encoding, which are generated via separate MLP networks. These MLP comprise two dense layers, each with 256 neurons and GELU activation. Figure 3 (right side) illustrates the U-Net architecture. To train the DDPM, we apply the Adam optimizer, with the learning rate set to 0.0001 and the batch size to 32. We conduct training for 200 epochs and observe that the model reaches a stable value for the loss function. After the successful training, we generate 500 synthetic samples for each class.

5 Experiments & Validation Procedure

In our experiments, we divide data randomly into a 60-40 train-test split, with 60% data allocated for training the DDPM and the remaining 40% reserved to evaluate the efficacy of the synthetic

embeddings. After splitting, the Train set comprises 759, 431, and 628 embedding samples, while the Test set consists of 506, 287, and 420 embedding samples, corresponding to the AD, CN, and FTD groups, respectively. We conduct the following experiments to evaluate the quality of synthetic EEG embeddings.

5.1 Train Synthetic Test Original

The quality of the synthetic data can be measured as its ability to maintain class discrimination for the downstream tasks [24]. We perform AD–CN and FTD–CN classification, where an MLP is trained on the synthetically generated EEG embeddings and tested on the Test set of original EEG embeddings. If good classification performance is observed, then one can argue that synthetically generated embeddings effectively differentiate between classes in the data.

5.2 Mutual Information

We calculate the Jensen-Shannon Divergence (JSD) score [25] between synthetic and original embeddings, which measures the difference in information represented by the two distributions. This quantitative evaluation provides a statistical measure of how well the synthetic embeddings align with the original embeddings. The JSD is symmetrical, meaning the order of the input distributions does not affect the result. Its value lies in the range [0, 1], where 0 indicates that given two distributions are identical. A higher JSD suggests greater dissimilarity, indicating potential information loss during the data generation process.

5.3 Comparison with VAE

We implement a class-conditioned Variational Autoencoder (VAE) [26] to compare the fidelity of the synthetic EEG embeddings generated by the proposed DDPM. VAE uses a probabilistic approach to learn a latent distribution of the input data, enabling the generation of new data points by sampling from the learned distribution. We apply *Train Synthetic Test Original* procedure on the EEG embeddings generated via VAE and summarize the classification results in Table 1.

6 Results & Discussion

6.1 Train Synthetic Test Original

While training the MLP with synthetic embeddings and testing it on unseen original embeddings, we observed robust classification results for both AD–CN and FTD–CN. This observation confirms that the synthetic EEG embeddings generated via DDPM preserve crucial class discrimination information, which is necessary for downstream tasks. We also observe lower uncertainty across multiple training repetitions, which confirms the reliability of DDPM in generating high-quality synthetic embeddings. The classification results are summarized in Table 1.

6.2 Mutual Information

In our experiments, we calculate JSD between synthetic and both Train and Test of original embeddings. The results, depicted in Table 2, demonstrate relatively low JSD scores for all three classes (AD, CN, and FTD) in both Train and Test sets. This finding suggested a significant alignment between

Table 1: Classification results via MLP; trained on *Synthetic Embeddings* and evaluated on the Test set of *Original Embeddings*. Precision, recall, and F1 score are calculated via unweighted averaging. These results represent the mean score and its uncertainty across 10 training repetitions.

MLP–Classification	Precision	Recall	F1 Score
AD–CN (DDPM)	**0.98±0.01**	**0.98±0.01**	**0.98±0.01**
FTD–CN (DDPM)	**0.86±0.01**	**0.87±0.01**	**0.84±0.01**
AD–CN (VAE)	0.79±0.08	0.76±0.09	0.75±0.09
FTD–CN (VAE)	0.74±0.11	0.63±0.10	0.53±0.10

Table 2: The JSD score is computed between the *Synthetic Embeddings* and the Train/Test set of *Original Embeddings*. JSD = 0 means identical distributions; JSD = 1 means completely dissimilar.

Class	Embedding Set-1	Embedding Set-2	JSD Score
AD	Synthetic	Train Set	0.043
AD	Synthetic	Test Set	0.042
CN	Synthetic	Train Set	0.051
CN	Synthetic	Test Set	0.048
FTD	Synthetic	Train Set	0.092
FTD	Synthetic	Test Set	0.094

the synthetic and original embedding distributions. Notably, the JSD scores remained consistent across different classes, indicating the reliability of the synthetic data generation process. While a low JSD score between the synthetic and Train set indicates a high degree of similarity; a low JSD score between the synthetic and Test set confirms the proposed method's capacity to generalize effectively on unseen data.

6.3 Comparison with VAE

In Table 1, the comparison between DDPM and VAE reveals that synthetic embeddings generated by DDPM consistently yielded higher precision, recall, and F1 scores compared to those generated by VAE in both AD–CN and FTD–CN classification. Furthermore, after conducting 10 training repetitions, we observed a higher uncertainty in the classification results with the VAE embeddings. This higher uncertainty hinders the ability to generalize well.

7 Conclusion

To address the limited availability of EEG data, we propose *MEDiC*, a class-conditioned DDPM-based framework which generates synthetic embeddings in the latent EEG space. Our method provides a practical and scalable solution to generate high-quality synthetic EEG embeddings. EEG datasets in healthcare often exhibit long-tail distributions. Our method can generate a flexible number of synthetic EEG data points, which can be useful across healthcare scenarios where the availability of EEG data is limited or highly imbalanced. Moreover, the synthetic data does not represent real individuals; therefore, the privacy concerns associated with sharing actual patient information are minimized. Furthermore, synthetic data may escalate specific biases which may or may not be inherent in the original EEG data. Therefore, it is crucial to identify a suitable semantic encoder which can handle these biases. Additionally, incorporating a semantic information-informed loss into the diffusion model itself can be a promising direction for future research.

References

[1] Adam Bohr and Kaveh Memarzadeh. *Artificial intelligence in healthcare*. Academic Press, 2020.

[2] Manal Alamir and Manal Alghamdi. The role of generative adversarial network in medical image analysis: An in-depth survey. *ACM Comput. Surv.*, 55(5), dec 2022.

[3] Nancy C. Andreasen. Brain imaging: Applications in psychiatry. *Science*, 239(4846):1381–1388, 1988.

[4] Introduction to the Physiological Bases of EEG. In *Analyzing Neural Time Series Data: Theory and Practice*. The MIT Press, 01 2014.

[5] Gulshan Sharma, Pankaj Pandey, Ramanathan Subramanian, Krishna Prasad Miyapuram, and Abhinav Dhall. Neural encoding of songs is modulated by their enjoyment. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, ICMI '22, page 414–419, New York, NY, USA, 2022. Association for Computing Machinery.

[6] Junfeng Sun, Yingying Tang, Kelvin O. Lim, Jijun Wang, Shanbao Tong, Hui Li, and Bin He. Abnormal dynamics of eeg oscillations in schizophrenia patients on multiple time scales. *IEEE Transactions on Biomedical Engineering*, 61(6):1756–1764, 2014.

[7] Siddharth Panwar, Shiv Dutt Joshi, Anubha Gupta, and Puneet Agarwal. Automated epilepsy diagnosis using eeg with test set evaluation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(6):1106–1116, 2019.

[8] Vincent Fortuin, Matthias Hüser, Francesco Locatello, Heiko Strathmann, and Gunnar Rätsch. Som-vae: Interpretable discrete representation learning on time series. *arXiv preprint arXiv:1806.02199*, 2018.

[9] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5):056013, jul 2018.

[10] Data augmentation for deep-learning-based electroencephalography. *Journal of Neuroscience Methods*, 346:108885, 2020.

[11] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Data augmentation using synthetic data for time series classification with deep residual networks. In *International Workshop on Advanced Analytics and Learning on Temporal Data, ECML PKDD*, 2018.

[12] Qingsong Wen, Liang Sun, Fan Yang, Xiaomin Song, Jingkun Gao, Xue Wang, and Huan Xu. Time series data augmentation for deep learning: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4653–4660. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Survey Track.

[13] Fatemeh Fahimi, Strahinja Dosen, Kai Keng Ang, Natalie Mrachacz-Kersting, and Cuntai Guan. Generative adversarial networks-based data augmentation for brain–computer interface. *IEEE Transactions on Neural Networks and Learning Systems*, 32(9):4039–4051, 2021.

[14] Hristos S. Courellis, John R. Iversen, Howard Poizner, and Gert Cauwenberghs. Eeg channel interpolation using ellipsoid geodesic length. In *2016 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pages 540–543, 2016.

[15] Isabel Valera, Melanie F. Pradier, Maria Lomeli, and Zoubin Ghahramani. General latent feature models for heterogeneous datasets. *Journal of Machine Learning Research*, 21(100):1–49, 2020.

[16] Tsz-Him Cheung and Dit-Yan Yeung. Modals: Modality-agnostic automated data augmentation in the latent space. In *International Conference on Learning Representations*, 2020.

[17] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10850–10869, 2023.

[18] 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.

[19] Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. Towards diverse and natural image descriptions via a conditional gan. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[20] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794. Curran Associates, Inc., 2021.

[21] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

[22] Andreas Miltiadous, Katerina D. Tzimourta, Theodora Afrantou, Panagiotis Ioannidis, Nikolaos Grigoriadis, Dimitrios G. Tsalikakis, Pantelis Angelidis, Markos G. Tsipouras, Evripidis Glavas, Nikolaos Giannakeas, and Alexandros T. Tzallas. "a dataset of eeg recordings from: Alzheimer's disease, frontotemporal dementia and healthy subjects", 2023.

[23] Andreas Miltiadous, Emmanouil Gionanidis, Katerina D. Tzimourta, Nikolaos Giannakeas, and Alexandros T. Tzallas. Dice-net: A novel convolution-transformer architecture for alzheimer detection in eeg signals. *IEEE Access*, 11:71840–71858, 2023.

[24] Joshua Snoke, Gillian M. Raab, Beata Nowok, Chris Dibben, and Aleksandra Slavkovic. General and Specific Utility Measures for Synthetic Data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 181(3):663–688, 03 2018.

[25] J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

[26] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.