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

Gulshan Sharma[1] , Abhinav Dhall[1,2], Ramanathan Subramanian[3]
[1]Indian Institute of Technology Ropar, [2]Monash University, [3]University of Canberra

NEURAL INFORMATION PROCESSING SYSTEMS

Background:

- Learning with a small-scale Electroencephalography (EEG) dataset is a non-trivial task.
- 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.

Contributions:

- 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.

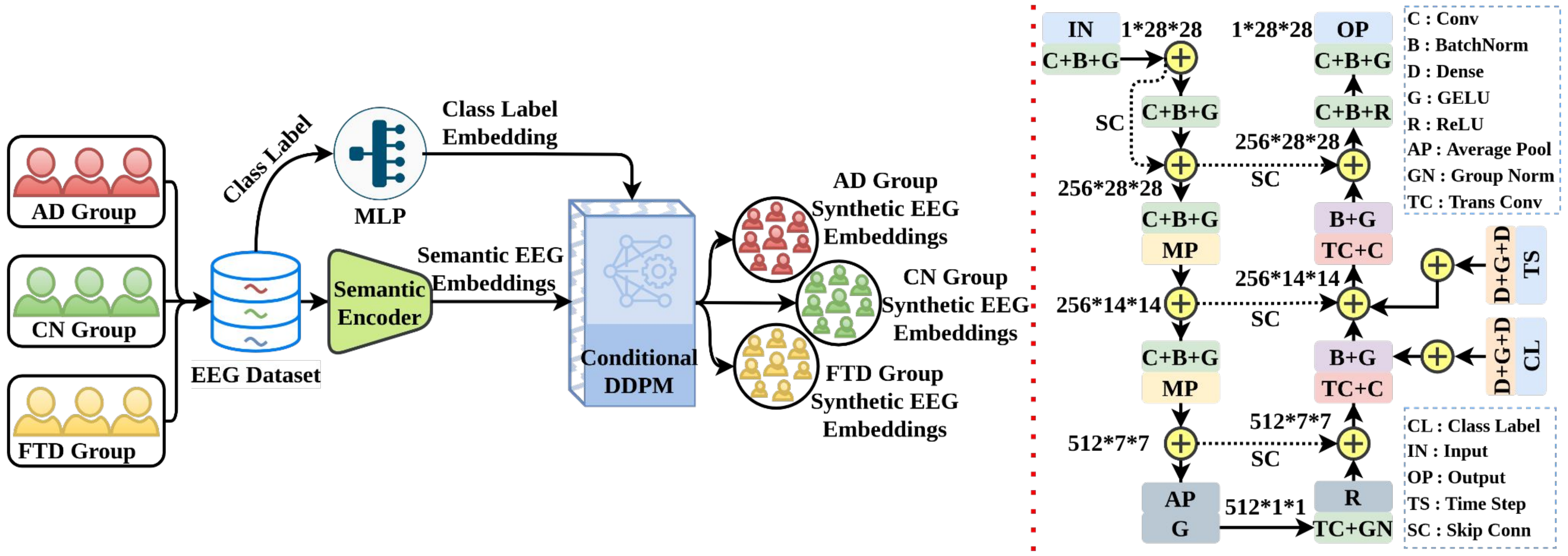

Fig. 1: MEDiC framework: EEG data is input to a semantic encoder to extract EEG embeddings. These embeddings, along with class label encodings, are then input into the class-conditioned DDPM to generate synthetic EEG embeddings (left). Conditional U-Net architecture used during backward diffusion process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

Results:

- 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 ± 0.01	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

Conclusion:

- Our method provides a scalable solution to generate high-quality synthetic EEG embeddings.
- Also reduces the privacy concerns associated with sharing actual patient data.