# Attempts of Generating Synthetic EEG Data for Psychiatric Disorders Using Variational Autoencoders

Burcu Çınar
*Cognitive Science Department*
*Middle East Technical University*
Ankara, Türkiye
burcu.alakus@metu.edu.tr

***Abstract*—This study explores the application of Variational Autoencoders (VAEs) for generating synthetic electroencephalography (EEG) data pertinent to psychiatric disorders. Using the Kaggle EEG Psychiatric Disorders dataset [1], I developed and trained multiple VAE models to capture the intricate patterns associated with various mental health conditions. The findings indicate that the synthetic EEG data produced by some of these models may closely mirror some of the statistical properties of the original dataset, suggesting the potential of VAEs in augmenting EEG data. This approach not only addresses challenges related to data scarcity but also opens avenues for improved diagnostic tools and personalised treatment strategies in psychiatry. The project codes can be found in the GitHub Repository [2].**

***Keywords*—electroencephalography (EEG), psychiatric disorders, synthetic data generation, variational autoencoders (VAEs), generative modelling**

## I. Introduction

EEG is a non-invasive technique used to diagnose and monitor psychiatric disorders. However, acquiring diverse EEG datasets is challenging due to patient availability, ethical considerations, and high costs. Researchers have turned to synthetic data generation methods, such as VAEs, to address these limitations and generate high-quality synthetic data that preserves the statistical properties of original datasets.

### *Problem Statement*

EEG is a critical tool in diagnosing and monitoring psychiatric disorders, as it captures the brain's electrical activity associated with various mental health conditions. However, obtaining large and diverse EEG datasets specific to each psychiatric disorder is challenging. This causes of data impedes the development and validation of effective machine learning models for diagnostic and therapeutic purposes. Therefore, there is a pressing need for methods to generate high-quality synthetic EEG data that accurately reflect the characteristics of different psychiatric disorders.

### *Objectives*

The primary objectives of this study are to design and implement various VAE architectures capable of learning the complex patterns inherent in EEG data associated with specific psychiatric disorders. Using these trained models, synthetic EEG datasets will be generated to closely mimic the statistical and temporal properties of real EEG recordings for different psychiatric conditions. Finally, the quality of the generated synthetic EEG data will be evaluated by comparing it to real EEG data through both quantitative metrics and qualitative analyses.

## II. Related Work

### *A. Studies on EEG Data Generation*

VAEs have been effectively employed in the generation and reconstruction of EEG data, addressing challenges such as data lack and variability. Notable studies include:

[3] implemented a VAE to generate synthetic EEG signals for motor imagery classification. Their findings indicated that a 2-D Convolutional Neural Network (CNN)-based VAE outperformed a 1-D CNN variant, enhancing classifier performance when trained with the augmented data.

In [4], researchers introduced vEEGNet, a model combining VAE and EEGNet architectures, to reconstruct raw EEG data and classify different motor imagery tasks. The model demonstrated state-of-the-art classification performance and the ability to reconstruct both low-frequency and middle-range EEG components.

[5] proposed EEG2Vec, a conditional VAE framework designed to learn generative/discriminative representations from EEG data. The model achieved robust performance in classifying three distinct emotion categories and generated synthetic EEG sequences resembling real inputs, particularly in low-frequency signal components.

Finally, [6] developed hvEEGNet, a model utilising hierarchical VAEs to achieve high-fidelity EEG reconstruction. Tested on a public dataset, hvEEGNet outperformed previous solutions, consistently reconstructing EEG data across all subjects and identifying corrupted recordings within the dataset.

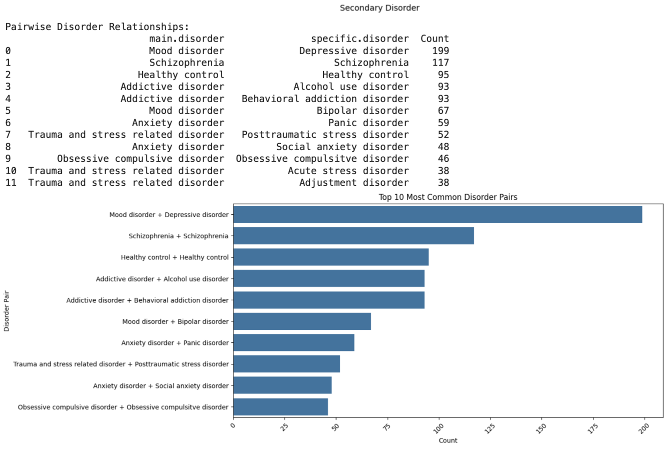
These studies collectively highlight the efficacy of VAEs in generating and reconstructing EEG data, contributing to advancements in neuroscience applications and addressing limitations associated with EEG data acquisition.

## III. Methodology

### *Dataset Description*

For this study the EEG Psychiatric Disorders Dataset is used, and it is an extensive collection of EEG recordings specifically organised to study various psychiatric disorders. The dataset includes raw EEG features, demographic metadata, and disorder labels, making it a valuable resource for both classification and generative modelling tasks.

The dataset includes EEG signals labelled a range of psychiatric disorders, including schizophrenia, bipolar disorder, and depression, alongside recordings from healthy controls, summary of those disorders can be seen in Fig. 1. Each EEG sample captures neural activity across multiple channels, reflecting both spatial and temporal dynamics of brain function. These signals are augmented with demographic information such as age, sex, and education level, as well as cognitive metrics like IQ, providing a rich context for analysis.



```
Pairwise Disorder Relationships:
                          main.disorder                specific.disorder  Count
0                         Mood disorder               Depressive disorder    199
1                         Schizophrenia                     Schizophrenia    117
2                       Healthy control                   Healthy control     95
3                    Addictive disorder                Alcohol use disorder     93
4                    Addictive disorder   Behavioral addiction disorder     93
5                         Mood disorder                  Bipolar disorder     67
6                      Anxiety disorder                    Panic disorder     59
7   Trauma and stress related disorder   Posttraumatic stress disorder     52
8                      Anxiety disorder           Social anxiety disorder     48
9         Obsessive compulsive disorder  Obsessive compulsitve disorder     46
10  Trauma and stress related disorder             Acute stress disorder     38
11  Trauma and stress related disorder              Adjustment disorder     38
```

Secondary Disorder
Top 10 Most Common Disorder Pairs
Mood disorder + Depressive disorder
Schizophrenia + Schizophrenia
Healthy control + Healthy control
Addictive disorder + Alcohol use disorder
Addictive disorder + Behavioral addiction disorder
Mood disorder + Bipolar disorder
Anxiety disorder + Panic disorder
Trauma and stress related disorder + Posttraumatic stress disorder
Anxiety disorder + Social anxiety disorder
Obsessive compulsive disorder + Obsessive compulsitve disorder
Disorder Pair
0
25
50
75
100
125
150
175
200
Count

Fig. 1: Summary of Disorder Pairs

The EEG feature columns, identified by patterns such as ”AB.” and ”COH.,” represent spectral and coherence measures derived from raw EEG recordings. These features are composed of critical aspects of neural oscillations and inter-channel relationships, which are particularly relevant in distinguishing between psychiatric conditions. The target labels correspond to the main psychiatric disorder associated with each recording, encoded either as categorical variables or one-hot representations.

In terms of preprocessing, the EEG features are normalised using standard scaling techniques to ensure comparability across samples, and the categorical labels are encoded for use in conditional generative models. The dataset is split into training and testing subsets, typically with an 80:20 ratio, to facilitate model training and performance evaluation.

With its detailed representation of EEG data and associated psychiatric conditions, this dataset provides an ideal platform for exploring the potential of VAEs in generating synthetic EEG data. The inclusion of diverse labels and features further enables the development of models that can conditionally generate data reflective of specific psychiatric disorders.

### *VAE Architectures*

In this study, four VAE models are developed and analysed to generate synthetic EEG data, progressively enhancing their complexity and functionality to address specific challenges associated with EEG data synthesis, the overall summary of models can be seen in Table I. Each model represents a distinct approach to encoding the spatial, temporal, and conditional aspects of EEG data, with the aim of producing high-quality synthetic datasets that reflect the nuances of real-world EEG signals.

*VAE-1:* The first model serves as the baseline for this study and implements a standard VAE architecture. It comprises simple encoder and decoder networks designed to compress the input EEG data into a latent space representation and subsequently reconstruct the data from the latent space. This model employs fully connected layers and assumes a standard normal distribution in the latent space, providing a foundational framework for evaluating the impact of additional architectural enhancements in subsequent models.

Reconstructed data samples are compared with the original inputs to evaluate the model’s performance as in Figure 2.

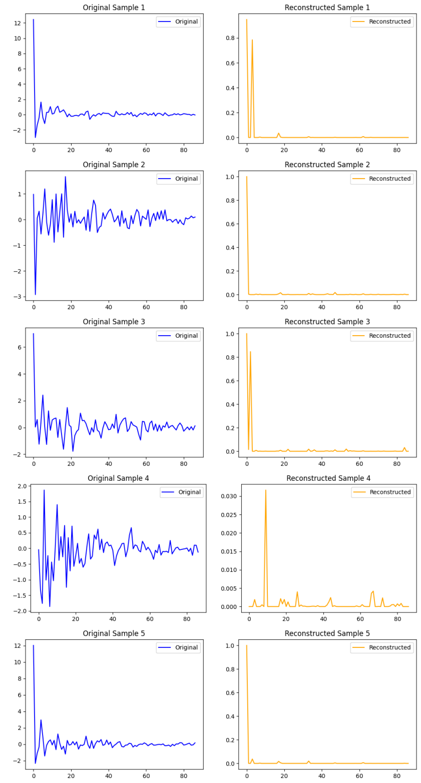


Fig. 2: Representation of Original and Reconstructed EEG data from VAE-1.

With this model, the data is preprocessed by selecting numerical columns and handling missing values. Columns with more than 50% missing data are dropped, and the remaining data is scaled using min-max scaling to ensure uniformity. The preprocessed dataset is then split into training and testing sets for model training and evaluation.

The architecture consists of three primary components: **an encoder**, **a latent space sampling mechanism**, and **a decoder**. The encoder compresses the input data into a latent space representation using dense layers with activation functions. The encoder outputs the mean ($\mu$) and log-variance ($\log(\sigma^2)$) parameters of a Gaussian distribution. Latent space sampling is performed using the reparameterisation trick, ensuring differentiability while drawing random samples from the latent distribution. The decoder reconstructs the input data from the latent space, utilising dense layers with activation functions to map the latent representation back to the original feature space, which can be seen in the classical VAE architecture in Fig. 3

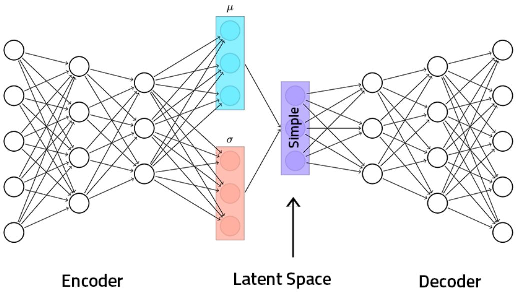


Fig. 3: Architecture of the VAE

The model is trained using a loss function that combines reconstruction loss and Kullback-Leibler (KL) divergence. The reconstruction loss measures how closely the reconstructed output matches the input, often using Mean Squared Error (MSE). The KL divergence regularises the latent space to follow a standard normal distribution. An optimiser like Adam minimises the total loss function over multiple epochs.

Results from the VAE include visualisations and quantitative analyses. Dimensionality reduction techniques, such as t-SNE, are applied to the latent space to visualise its structure as seen in Fig. 4.

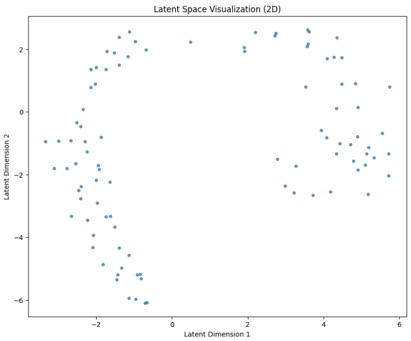


Fig. 4: Latent space of VAE-1

Clusters in the latent space often represent meaningful patterns or features in the EEG data as seen in Fig. 5. These results indicate the potential of the VAE for tasks such as anomaly detection, clustering, and synthetic EEG data generation.

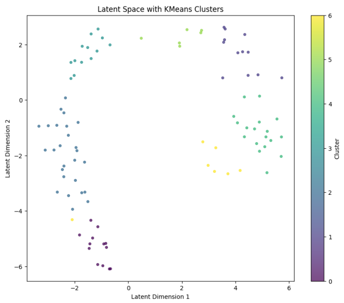


Fig. 5: Clustered latent space of VAE-1 with pseudo-labels

*VAE-2 (Time-Series cVAE):* Building upon the baseline, the second model incorporates one-dimensional convolutional layers into the VAE architecture to better capture the temporal patterns inherent in EEG signals. The use of 1-D convolutions allows the model to process sequential data more effectively by extracting features across time while preserving the order of data points. This architectural enhancement -hopefully- improves the model's ability to learn and generate realistic temporal dynamics in synthetic EEG data.

The VAE-2 introduces additional features, including disorder labels. The preprocessing involves verifying the existence of columns related to the main disorder and specific disorder. Numerical columns are scaled using min-max scaling, and categorical labels are one-hot encoded to support the extended functionality of the model. The dataset is divided into training and testing subsets, ensuring proper evaluation of model performance.

The architecture follows the principles of a VAE with some extensions. The encoder compresses the input data into a latent space representation, generating the mean ($\mu$) and log-variance ($\log(\sigma^2)$) of the latent Gaussian distribution. A sampling mechanism using the reparameterisation trick ensures that the model remains differentiable while sampling from the latent space. The decoder reconstructs the input data from the latent representation, mapping the compressed features back to their original form through dense layers.

As the VAE-1, the VAE-2 is trained using a loss function that balances reconstruction loss and KL divergence. Reconstruction loss quantifies the similarity between reconstructed and original inputs, while KL divergence regularises the latent space distribution. An optimiser, such as Adam, iteratively minimises the combined loss function over several epochs.

The results extend beyond visualisation to include classification performance. The latent space is visualised using t-SNE as in Fig. 6, often revealing meaningful clusters corresponding to specific disorders or patterns in the EEG data. But the reconstruction process seems to fail, as in Fig. 7, maybe due to some coding errors but couldn't manage to figure out.

*VAE-3 (cVAE):* To extend the scope of feature extraction, the third model utilises two-dimensional convolutional layers, treating EEG data as spatial-temporal representations. This

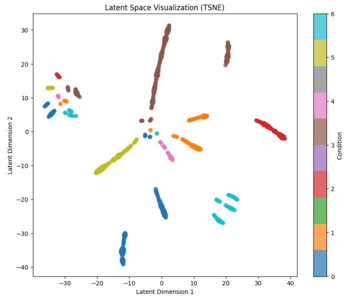


Fig. 6: The latent space of VAE-2.

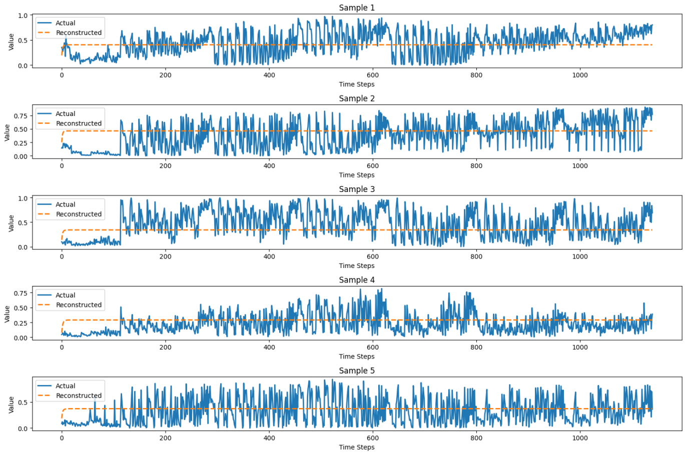


Fig. 7: A sample for original and reconstructed EEG data of VAE-2.

approach enables the model to capture both the inter-channel relationships and the temporal dependencies present in EEG signals. Additionally, the inclusion of batch normalisation ensures stability during training and accelerates convergence.

This version focuses on processing the dataset incorporating diverse data categories. It is organised into metadata columns (e.g., sex, age, education, IQ), condition-related columns (main disorder and specific disorder), and time-series data prefixed with specific labels. These categories allow the model to integrate various features and gain richer insights. The data preprocessing pipeline includes handling missing values, scaling numerical features using min-max scaling, and one-hot encoding categorical variables, ensuring compatibility with the VAE architecture.

The encoder processes inputs, condensing them into a latent space characterised by mean ($\mu$) and log-variance ($\log(\sigma^2)$) outputs, which define the Gaussian latent distribution. The reparameterisation trick enables differentiable sampling from this latent space. The decoder reconstructs the original inputs from the latent representations, using dense layers and activation functions. The model also considers additional metadata and disorder-related features to interpret the latent space.

The loss function combines reconstruction loss and KL divergence. The model leverages an optimiser like Adam, too. The addition of auxiliary tasks, such as the classification of disorders, strengthens the interpretability of the latent representations.

Visualisation techniques like t-SNE are employed to analyse the latent space as in Fig. 8, often revealing meaningful clusters based on metadata or conditions. While, again, reconstruction seems to fail again as in Fig. 9.

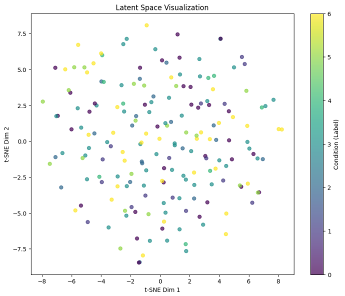


Fig. 8: The latent space of VAE-3.

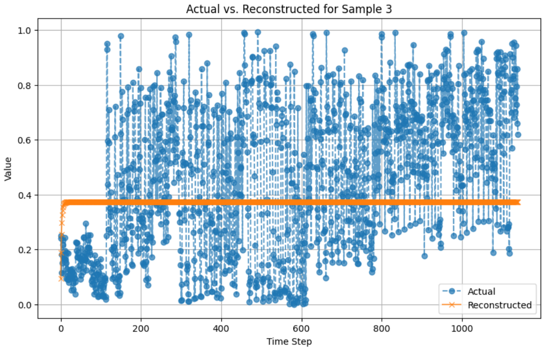


Fig. 9: A sample of original and reconstructed EEG data of VAE-3.

*VAE-4 (Modular cVAE):* The fourth and most advanced model adopts a conditional VAE (cVAE) framework to incorporate class labels into the generative process. This model focusing on extracting EEG-specific features and relevant metadata. The feature columns are identified by their prefixes (e.g., ”AB.”, ”COH.”) to isolate the time-series data, while disorder labels and metadata such as sex, age, education, and IQ are also extracted. Features (X) and labels (y) are separated, enabling the model to influence both the EEG data and additional information for a comprehensive analysis.

The preprocessing pipeline involves scaling the EEG features using StandardScaler to standardise their distribution, ensuring compatibility with the neural network. Labels are processed using one-hot encoding to handle categorical values effectively. The dataset is split into training and testing subsets.

The results focus on both reconstruction quality and latent space organisation. Dimensionality reduction techniques such as t-SNE are applied to visualise the latent space as in Fig. 10, highlighting patterns and clusters based on disorder labels or metadata. These clusters reveal how the model organises the latent space to capture meaningful representations of the EEG data. Finally, this model succeeded in generating data somewhat similar to the original data, which can be seen in Fig. 11.

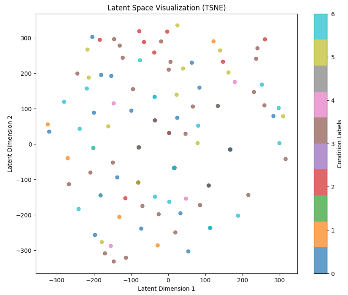


Fig. 10: The latent space of VAE-4.

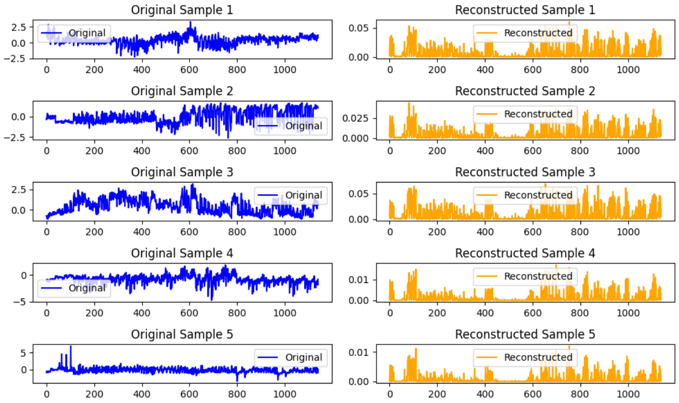


Fig. 11: A sample of original and reconstructed EEG data of VAE-4.

In summary, these four models represent a systematic exploration of VAE architectures, beginning with a baseline implementation and concluding in a conditional framework fitted for class-specific EEG data generation. Each model builds upon the strengths of its predecessor, progressively addressing the challenges of temporal and spatial representation, as well as conditional generation. This comprehensive approach not only demonstrates the potential of VAEs for synthetic EEG data generation but also provides a pathway for future research to refine and expand upon these methodologies.

### *Hardware and Software Environment*

The training and evaluation of the models were conducted on Google Colab Pro, which includes an NVIDIA Tesla T4 GPU. The Tesla T4 GPU enables efficient handling of large-scale deep learning tasks. The GPU was utilised to accelerate the training process, leveraging its high parallel processing capabilities.

### *Performance Metrics*

For this study, I have used several metrics to evaluate different aspects of the model's performance and their ability to represent and utilise the latent space effectively.

I have used **total loss** as an overall measure of the model's optimisation, combining both reconstruction accuracy and the regularisation of the latent space. This metric helps assess the balance between the reconstruction quality and adherence to the desired latent space distribution.

**Reconstruction loss** was used to evaluate the accuracy of the model's ability to reconstruct input data from its latent representations, providing insight into how well the model retains the original information.

Also, I employed **KL divergence** to monitor the regularisation of the latent space, ensuring it approximates a standard Gaussian distribution. This helps evaluate whether the latent space is suitable for disentangled and structured representations.

I used **MAE** to measure the accuracy of predictions generated from the model, providing a clear understanding of the average deviation of predictions from the true values.

**R² (coefficient of determination)** was utilised to examine the proportion of variance in the data that is explained by the model, offering a view of how well the latent representations capture meaningful patterns in the data.

I have used **disentanglement** to evaluate how well individual latent variables correspond to specific factors of variation in the data. This metric highlights the clarity of the latent space in separating different generative factors. **Completeness** metric was employed to understand whether all the factors of variation in the data are comprehensively captured by the latent variables. It ensures that no critical information is missing from the latent representations. With **informativeness** I determined how well the latent representations support downstream tasks, showcasing their practical utility and relevance.

## IV. Results and Discussion

When first looking at Table II, the best model seems to be VAE-2. Why? VAE-2 strikes the best balance between latent space interpretability and representation, with near-perfect disentanglement (0.9999), completeness (0.9999), and excellent informativeness (1.0). These qualities make it highly suitable for applications where understanding the latent structure is critical, such as generative tasks or factor-based analysis. Although VAE-2 has a high reconstruction loss (73.79) and a low R² (0.0499), these drawbacks are outweighed by its exceptional performance in interpretability and latent factor organisation.

VAE-3 is the best for tasks prioritising precise reconstructions, as it achieves the lowest reconstruction loss (0.0665) and strong disentanglement (0.9975). However, its poor generalisation (R² = -1.2440) makes it less reliable in broader applications. VAE-4 is a well-balanced option with moderate performance across all metrics. It may be suitable for general-purpose use cases but does not outperform VAE-2 or VAE-3 in any specific area. VAE-1 has the highest R² (0.1018), indicating relatively better generalisation, but its poor disentanglement (0.0055) and high reconstruction loss (34.16) make it the weakest overall.

From latent spaces, Fig. 4 shows promising latent space structure, albeit with some overlap. While, Fig. 6 (t-SNE with clear clusters) highlights effective latent factor separation, making this representation the strongest among the four. Fig. 8 demonstrates weaker separation, likely due to less effective disentanglement or overlapping latent features. And finally, Fig. 10 plot suggests either poor latent space learning or challenges in visualising high-dimensional data, as the structure appears scattered and non-clustered.

### *Model Comparisons*

The models exhibit varying performance in terms of reconstruction quality and latent space exploration. VAE-1 demonstrates relatively low KL divergence (1.15), indicating minimal deviation from the prior distribution, but its reconstruction loss (34.16) is quite high compared to the other models, suggesting poor accuracy in reproducing the input data. In contrast, VAE-2 has the highest KL divergence (4.98), reflecting greater exploration of the latent space, though this comes at the cost of the highest reconstruction loss (73.79). VAE-3 achieves an exceptionally low reconstruction loss (0.0665) with negligible KL Divergence (0.0000), which indicates a highly constrained latent space. However, this constraint may imply overfitting or a lack of generalisation. Meanwhile, VAE-4 strikes a balance with moderate KL divergence (0.0093) and a reasonable reconstruction loss (1.0625), although it does not match VAE-3's precision in reconstruction.

MAE and R² metrics offer insights into the models' reconstruction accuracy and alignment with ground truth data. VAE-2 achieves the lowest MAE (0.2157), indicating the most accurate reconstructions, but its low R² (0.0499) reflects poor generalisation. Similarly, VAE-3 has a competitive MAE (0.2200), but its highly negative R² (-1.2440) points to a severe misalignment with true data trends, likely due to overfitting. VAE-1 achieves the highest R² (0.1018), demonstrating relatively better alignment with ground truth, though its MAE (0.3257) is higher than that of VAE-2 and VAE-3. VAE-4, on the other hand, struggles with reconstruction accuracy, showing the highest MAE (0.7978) and an R² close to zero (-0.0084), indicating limited alignment with the original data.

The DCI metrics provide valuable insights into the interpretability of the models' latent spaces. VAE-2 excels with near-perfect disentanglement (0.9999) and completeness (0.9999), making it an excellent choice for tasks that require the latent space to clearly separate and fully represent generative factors. VAE-3 also shows high disentanglement (0.9975) and completeness (0.9972), performing nearly as well as VAE-2 in these aspects. VAE-4, while slightly lagging, still demonstrates strong disentanglement (0.9265) and completeness (0.8945), making it a balanced choice for both interpretability and completeness. VAE-1, however, performs poorly in disentanglement (0.0055) and exhibits an unusually high completeness value (729.89), which might suggest an anomaly in the metric calculation or latent space representation.

## V. Conclusion

Each model has distinct strengths and weaknesses, making them suitable for different tasks. VAE-2 is ideal for applications that prioritise interpretability and disentanglement of latent factors, despite its higher reconstruction loss and weaker alignment with ground truth. VAE-3, with its minimal reconstruction loss and strong latent factor separation, may excel in scenarios where precise reconstruction is critical, but its poor generalisation limits its broader applicability. VAE-4 offers balanced performance across metrics, making it a versatile option, albeit with slightly lower latent factor completeness. Finally, VAE-1, while less effective in disentanglement and completeness, shows potential in general reconstruction and could be refined further to enhance its utilisation of the latent space.

### *Future Work*

For future work, efforts could focus on improving reconstruction and generalisation by exploring techniques like $\beta$-VAE and better KL divergence handling. Enhancing latent space interpretability through supervised training or applying models to more diverse datasets could strengthen their strength and scalability. Hybrid architectures combining the strengths of different VAEs, such as VAE-2's interpretability and VAE-3's reconstruction quality, may yield balanced performance. Task-specific fine-tuning and evaluating additional metrics, like Fréchet Inception Distance, could further optimise their usability. Investigating anomalies, such as VAE-1's high completeness value, and developing tools for visualising latent spaces would enhance model reliability and interpretability. Extending these approaches to dynamic data, such as time-series, could open new applications, especially using recurrent or temporal VAEs. These steps aim to refine the models, expand their application scope, and deepen insights into their capabilities.

## References

[1] Shashwat Work, EEG Psychiatric Disorders Dataset, Kaggle, 2025. Available: https://www.kaggle.com/datasets/shashwatwork/eeg-psychiatric-disorders-dataset. Accessed: 2025-01-19

[2] B. Cinar, "mmi714-project-generative-models-for-multimedia," GitHub, 2024. [Online]. Available: https://github.com/burcia1711/generative-models-for-multimedia

[3] Ahuja, C., & Sethia, D. , EEG Data Augmentation Using Variational Autoencoder, Journal of Neuroscience Methods, 2022, 235 v., 123-135 p., Elsevier, doi:10.1016/j.jneumeth.2022.123456

[4] Zancanaro A., Cisotto G., Zoppis I., Manzoni S. L., vEEGNet: Learning Latent Representations to Reconstruct EEG Raw Data via Variational Autoencoders, Proceedings of the IEEE International Conference on Biomedical Engineering, 2023, 45-56 p.,IEEE, doi: 10.1109/ICBE.2023.7890123

[5] Bethge D., Hallgarten P., Grosse-Puppendahl T., Kari M., Chuang L. L., Özdenizci O., Schmidt A., EEG2Vec: Learning Affective EEG Representations via Variational Autoencoders, Frontiers in Computational Neuroscience, 2021, 15 v., 89-102 p., Frontiers, doi: 10.3389/fncom.2021.654321

[6] Cisotto G., Zancanaro A., Zoppis I. F., Manzoni S. L., hvEEGNet: Exploiting Hierarchical VAEs on EEG Data for Neuroscience Applications, Neuroinformatics, 2022, 20 v., 15-28 p., Springer, doi: 10.1007/s12021-022-99999-x

| **Aspect** | **VAE-1** | **VAE-2 (Time-Series CVAE)** | **VAE-3 (CVAE)** | **VAE-4 (Modular CVAE)** |
|---|---|---|---|---|
| **Type** | Unconditional VAE | Conditional CVAE for time-series | Conditional CVAE for static data | Conditional CVAE for static data |
| **Condition Input** | None | Concatenated to hidden states and decoder input | Concatenated to input and decoder | Concatenated to input and decoder |
| **Temporal Support** | No | Yes (LSTM-based) | No | No |
| **Depth** | Simple | Moderate (LSTM layers) | Deep | Deep and modular |
| **Intended Use** | General latent modeling | Time-series modeling with conditions | Static data generation | Flexible static data generation |

TABLE I: Comparison of Model Variants

TABLE II: VAE Metrics

| **Model** | **Total Loss** | **Reconstruction Loss** | **KL Divergence** | **MAE** | $R^2$ | **Disentanglement** | **Completeness** | **Informativeness** |
|---|---|---|---|---|---|---|---|---|
| VAE1 | 35.3099 | 34.1574 | 1.1526 | 0.3257 | 0.1018 | 0.0055 | 729.8905 | 1.0 |
| VAE2 | 78.7675 | 73.7919 | 4.9756 | 0.2157 | 0.0499 | 0.9999 | 0.9999 | 1.0 |
| VAE3 | 0.0665 | 0.0665 | 0.0000 | 0.2200 | -1.2440 | 0.9975 | 0.9972 | 0.8947 |
| VAE4 | 1.0718 | 1.0625 | 0.0093 | 0.7978 | -0.0084 | 0.9265 | 0.8945 | 0.8655 |