



eBRAIN-Health

Public report D8.1

Identified biomarkers using explainable AI

Project number	101058516
Project title	eBRAIN-Health - Actionable Multilevel Health Data
Submission date	July 2025
Authors	Ramesh Upreti, Lukas Gemein, Anis Yazidi, Ira Haraldsen (OUS)
Dissemination level	Public (PU)
Public project website	https://ebrain-health.eu



**Funded by
the European Union**

Table of content

Abbreviations	3
1. eBRAIN-Health.....	4
2. eBRAIN-Health partners	4
3. Introduction	5
4. Partners involved	6
5. Description of work performed	7
5.1 GAN	7
5.2 Model Architecture	9
5.3 WP8`s Personalized Variations	11
5.4 Machine-learning pipeline.....	13
5.5 Explainability	14
6. Results	14
6.1 Extracted EEG code features association with alpha peaks	15
6.2 Personalized EEG variations	15
6.2.1 Synthetic EEG	15
6.2.2 PSD of generated variations	16
6.2.3 PSD of generated variations extracted from MEEGLET	17
6.3 Machine learning prediction	18
6.4 Model Explainability	20
7. Future work and limitations	21
Acknowledgements	22
References	23

Abbreviations

AD	Alzheimer's disease
MCI	Mild cognitive impairment
APOE	Apolipoprotein E
p-tau	Phosphorylated Tau
GAN	Generative Adversarial Networks
VQ	Vector Quantization
WP	Working Package
MoCA	Montreal Cognitive Assessment
MADRS	Montgomery Asberg Rating Scale
CANTAB	Cambridge Neuropsychological Test Automated Battery
MEG	Magnetoencephalography
EC	Eye Close
EO	Eye Open
PSD	Power Spectral Density
GREEN	Gabor Riemann EEGNET
BCI	Brain Computer Interface
CNN	Convolutional Neural Network
ECG	Electrocardiogram

1. eBRAIN-Health

The Project eBRAIN-Health will deliver a distributed research platform for modeling and simulating complex neurobiological phenomena of human brain function and dysfunction in a data protection compliant environment. It will provide thousands of multilevel virtual brains from patients and healthy human controls for research and innovation. Brain data from multiple sources will be pre-processed. Solving the societal grand challenge of dementia is a big task. Yet it appears feasible in a collective approach. Therefore, we intend to build an interdisciplinary digital twin for dementia for modeling and simulating complex phenomena at the service of research infrastructure communities. eBRAIN-Health-Cloud will offer end-to-end services for personalized complex brain modeling and simulations in distributed e-infrastructures with data protection by design and by default and simulation-ready human multiscale brain data that range from molecular (genomics, proteomics, metabolomics) and cellular to electrophysiology and imaging to behavioural, clinical, life-style and environmental data as well as data from wearables. Brain data are pre-processed and annotated such that they all relate to a common reference 3D brain space.

eBRAIN-Health-Cloud constitutes a blend of three large-scale research programs: the FET Flagship Human Brain Project with its EBRAINS Research Infrastructure, the EOSC project Virtual Brain Cloud with its Virtual Research Environment for sensitive data and the H2020 project AI-MIND with intelligent tools for dementia risk estimation. The project will have synergies to topics of the Digital Europe Program, such as artificial intelligence, cybersecurity and supercomputing and the Health Data Space. eBRAIN-Health-Cloud offers a next generation clinical research infrastructure and creates an open yet protected space for groundbreaking digital health innovation by the research infrastructure communities comprising academia and the private sector.

2. eBRAIN-Health partners

- CHARITE – Universitaetsmedizin Berlin, Germany
- EBRAINS, Belgium
- Forschungszentrum Juelich GmbH, Germany
- Stichting Radboud Universiteit, Netherlands
- Universidad Pompeu Fabra, Spain
- OSLO Universitetssykehus, Norway
- tp21 GMBH, Germany
- Fraunhofer Gesellschaft zur Foerderung der Angewandten Forschung eV, Germany
- INDOC RESEARCH EUROPE gGmbH, Germany
- Universitaet Wien, Austria
- Universidad Complutense de Madrid, Spain
- EODYNE Systems SL, Spain
- ATHENA – Research and Innovation Center, Greece
- University of Oslo, Norway
- Universita degli Studi di Roma la Sapienza, Italy
- Alzheimer Europe, Luxembourg
- Institute National de Recherche en Informatique et Automatique, France
- Centre Hospitalier Universitaire Vaudois, Switzerland
- The University of Edinburgh, United Kingdom

[Find the partners on our website](#)

3. Introduction

Collaborative Innovation in WP8: WP8, led by Oslo University Hospital (OUH), is a central component of the eBRAIN-Health project, running from Month 1 to Month 48. Its main objective is to develop advanced machine learning (ML) and deep learning (DL) tools that integrate synthetic and real-world brain data to support early, accurate, and explainable dementia diagnostics. By leveraging multimodal datasets and high-performance computing, WP8 aims to identify robust biomarkers and construct predictive models that are both clinically meaningful and generalisable.

The work begins with **Task 8.1**, which strengthens the AI-Mind Connector algorithm by training it on a combination of empirical EEG/MEG data and synthetic datasets generated from eBRAIN-Health’s multiscale simulations. These simulated data capture a range of cognitive trajectories, including stable MCI and progression to dementia, and are used to improve the model’s performance and resilience.

Expanding on this, **Task 8.2** focuses on extracting domain-invariant features that characterize brain dysfunction, using both synthetic and biological data. It employs explainability techniques such as gradient-based methods to interpret the internal representations of DL models. These insights yield novel biomarkers and provide empirical validation for the simulation models, ensuring biological plausibility. In later phases, **Task 8.6** introduces scalable ML workflows tailored to the analysis of outputs from simulated brain models. These workflows enable the exploration of functional brain network changes and support hypothesis generation related to disease mechanisms. Concurrently, **Task 8.7** delivers a software suite for interpretable inference. This includes tools to statistically rank feature relevance and visualize predictions at both individual and cohort levels. The methods emphasize non-parametric modeling, using deep generative approaches and counterfactual simulations to support causal inference in digital twin analyses.

A key focus throughout WP8 is on **subgroup stratification** within the AI-Mind cohort, combining diverse data types—demographic factors (age, sex, education, alcohol use), blood biomarkers (pTau217, pTau181), APOE genotype, clinical scores (MoCA, MADRS), and digital cognitive testing (CANTAB). By applying unsupervised learning, WP8 identifies distinct patient profiles that enhance the accuracy and fairness of AI predictions, reduce potential bias, and improve clinical relevance.

WP8 culminates in four major deliverables of whom **the D8.1 (Month 36), Identification of biomarkers through explainable AI** is now approaching, and will lay the foundation for the upcoming

- **D8.2 (Month 42):** Algorithms for bias mitigation and model robustness

Altogether, WP8 delivers the analytical engine of eBRAIN-Health, bridging simulation science and clinical application. It ensures that digital brain twins evolve as interpretable, ethically grounded tools for precision medicine—enabling earlier, more personalized interventions for individuals at risk of dementia.

Synergy with [AI-Mind.eu](https://www.ai-mind.eu)

A central contributor to eBRAIN-Health is the AI-Mind project, which focuses on early identification of dementia risk in individuals with mild cognitive impairment (MCI). AI-Mind brings into eBRAIN-Health

a rich repository of real-world clinical data, including EEG and MEG recordings, cognitive assessments, genetic and blood-based biomarkers, and predictive machine learning models.

In return, eBRAIN-Health strengthens AI-Mind through the provision of synthetic data derived from brain simulations, enabling model augmentation and robustness testing. AI-Mind also functions as a federated satellite node within the eBRAIN-Health infrastructure, allowing secure, legally compliant data exchange while preserving institutional autonomy and data sovereignty.

AI-Mind's core innovation lies in two AI-driven tools:

- The AI-Mind Connector, which identifies dysfunctional brain network patterns using EEG/MEG.
- The AI-Mind Predictor, which estimates the individual's dementia risk by integrating neurophysiological signals, cognitive test scores, genotypic markers, and digital health data.

These tools are currently being validated through longitudinal clinical studies at four European hospitals, with data collected from 1,022 MCI patients. Of these, 594 high-quality datasets are used for AI development, and 725 are available for broader clinical research purposes. The study is now entering its final phase, where definitive ground-truth labels—such as stable MCI, cognitive decline, or conversion to dementia—are being established.

Deliverable of WP8 (D8.1) Identify Biomarkers, Using Explainable AI

This deliverable presents the outcomes of T8.1 and T8.2: a set of novel, explainable biomarkers derived from domain-adapted AI models trained on hybrid (synthetic + real) datasets. It validates that incorporating simulated knowledge improves biomarker discovery and model transparency. Although this is an ongoing process throughout the project lifetime. The principles of production and evaluation are finalised and reported in the Deliverable D8.1 which included two tasks:

Task 8.1 – Enriching AI-Mind Algorithms with Simulated Data

This task generated synthetic EEG and brain network data using eBRAIN-Health simulations, mimicking various trajectories of neurodegenerative decline. These synthetic datasets are used to train and enhance AI-Mind's algorithms, increasing their robustness and accuracy in predicting dementia-related changes, especially when real data are limited.

Task 8.2 – Extracting Biomarkers Using Explainable AI Methods on Models Trained in T8.1

This task focused on identifying **biologically meaningful and explainable biomarkers** by interpreting deep learning models trained in Task 8.1. Key steps include:

- Extracting latent code features from biological EEG data
- Interpret the association of extracted features with alpha peaks and overall PSD patterns
- Used extracted EEG features to create multiple possible trajectories of synthetic EEG
- Combining synthetic and biological (empirical) data using to train AI models
- Applying **explainability tools** to extract interpretable features:
 - **Gradient-based methods** to assess input-output sensitivity in predictions.

4. Partners involved

OUS - Ramesh Upreti, Lukas Gemein, Anis Yazidi, Ira Haraldsen

UNIRM1 - Claudio Babiloni

5. Description of work performed

Within AI-Mind, we have access to 594 subjects data from baseline (visit 1) for development and validation of AI algorithms out of total scientific available cases of 1022. All participants were registered by two sessions of eyes closed (EC) and eyes open (EO). Since EC signals contain less external non-brain related recording signals, e.g. due to eye movement or face muscle activity, we used the EC sessions as a basis to train our generative AI model to produce synthetic AI-Mind EEG data.

The overall workflow performed in this study is as follows: (1) train a deep learning-based generative adversarial network (GAN) model using EEG data from EC sessions, (2) extract features from each EEG (discretized latent features, also known as code features), (3) check the association between the extracted code/latent features and alpha peaks (one of the most prominent Power Spectral Density (PSD)-based features in dementia studies) to determine whether the extracted features capture frequency-based features, (4) use the extracted code features to create personalised diversifications by randomly selecting five or ten code features and replacing them with other code features based on a defined level of similarity, (5) use the decoder part of the pre-trained model from step one to invert from the code features to the EEG, (6) extract covariance features from biological and synthetic EEG using the MEEGLET package, (7) train the GREEN2 model based on the covariance model to predict sex and P-tau, as well as the deep learning-based (Shallow and EEGInception) models on continuous raw EEG, (9) evaluate the performance of the models trained over different data conditions, and (10) finally, use explainable approach to explain the the trained classifier model decision when the model is trained on biological and synthetic EGG to decode the decision boundary.

5.1 GAN

Generative Adversarial Networks (GANs), introduced by Ian Goodfellow in 2014 [11], represent a significant advancement in data generation technology. GANs operate through an adversarial relationship between two neural networks: a generator and a discriminator. The generator creates synthetic data while the discriminator evaluates authenticity by distinguishing between real and synthetic samples. Through an iterative feedback process, the generator continually improves its output quality to better deceive the discriminator, while the discriminator simultaneously enhances its ability to detect synthetic data. This competitive training continues until the generator produces high-fidelity synthetic data that the discriminator can no longer reliably identify as fake. At this ideal endpoint, the discriminator's accuracy approaches around 50% which is essentially random guessing which indicates that the synthetic data has become virtually indistinguishable from real data. The architecture of the GAN is presented in the below Figure 1.

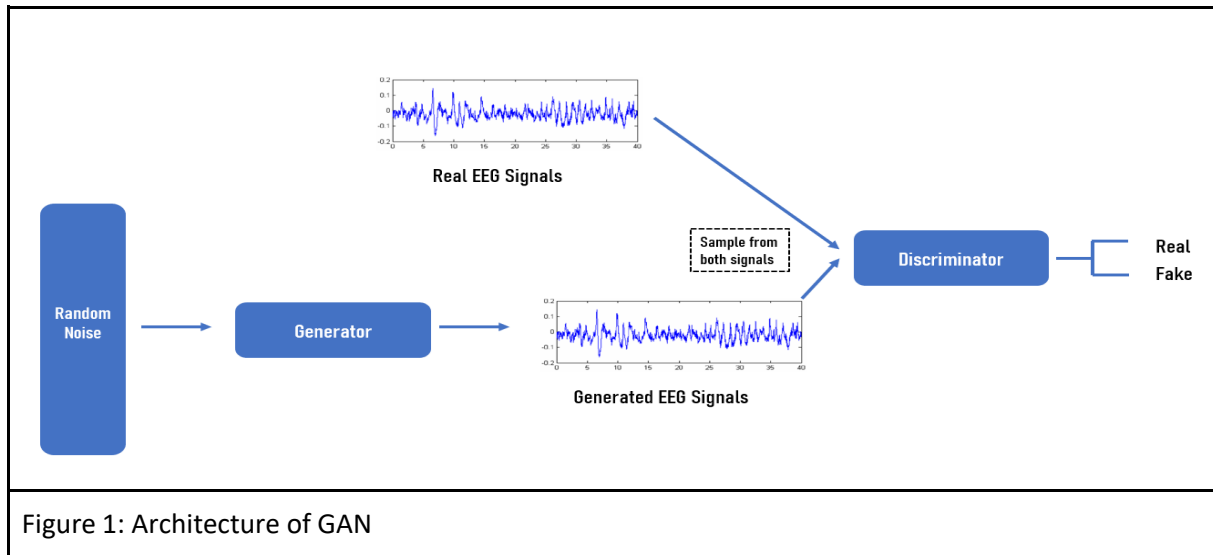


Figure 1: Architecture of GAN

The application of GANs for synthetic electroencephalogram (EEG) data generation has gained significant traction in recent years. Song et al. [22] introduced EEGGAN-Net, an innovative approach to EEG signal classification that effectively addresses the persistent challenges of limited data availability in medicine and more efficient feature extraction in brain-computer interface (BCI) applications [12]. The authors used conditional generative adversarial network (CGAN) for data augmentation which accepts random noise and class labels as input which helps generators to generate synthetic EEG samples while preserving class-specific characteristics. Similarly, the authors also used a cropped training strategy that segments EEG signals into overlapping windows to increase training data quantity and capture local patterns. The authors modified the last layer of the discriminator and made it a classifier for prediction. The authors performed rigorous evaluation on the BCI Competition IV-2a and IV-2b datasets. According to the result reported in the paper, the authors achieved highest classification accuracies of 81.3% and 90.3% on two datasets respectively. The authors claimed that their result outperformed four other contemporary CNN-based decoding models.

Another recent study addresses the critical challenge of limited biophysical data availability in medical applications by implementing a Wasserstein GAN with Gradient Penalty (WGAN-GP) to generate synthetic electroencephalography (EEG) and electrocardiogram (ECG) waveforms [13]. For EEG, the authors collected 60-second recordings across four channels (TP9, AF7, AF8, and TP10) from four individuals, yielding approximately 15,000 data points per subject per mental state (concentration and relaxation), while the ECG dataset comprised 1200 real samples with 140 data points each for normal and abnormal cardiac states. Authors performed some data preprocessing techniques including discrete wavelet transform, downsampling, and upsampling to enhance signal quality before feeding into the model. The authors reported classification accuracies of CNN based models over varying amounts of synthetic data in addition to real data. The authors reported the classifier yields 92% accuracy on real EEG data while the addition of 50% of synthetic data helped to achieve 98.48% of accuracy while for ECG data, the addition of synthetic samples improved random forest classifier accuracy from 97% to 98.40%. Statistical validation using the Wilcoxon signed-rank test confirmed the robustness of these improvements, demonstrating that WGAN-GP-based data augmentation offers a viable solution for overcoming data scarcity while maintaining privacy and addressing volunteer availability limitations in biophysical signal classification applications.

These recent studies demonstrate significant advancements in GAN-based EEG synthesis, addressing key challenges in data augmentation, clinically valid signal generation, spatial-temporal modeling, and privacy preservation. The innovative architectures and training techniques have enabled the generation of increasingly realistic EEG signals that preserve critical characteristics for both research and clinical applications. As these methods continue to evolve, synthetic EEG data generation using GANs holds promise for accelerating neuroscience research, improving diagnostic tools, and enabling more effective dementia studies while addressing data scarcity and privacy concerns.

5.2 Model Architecture

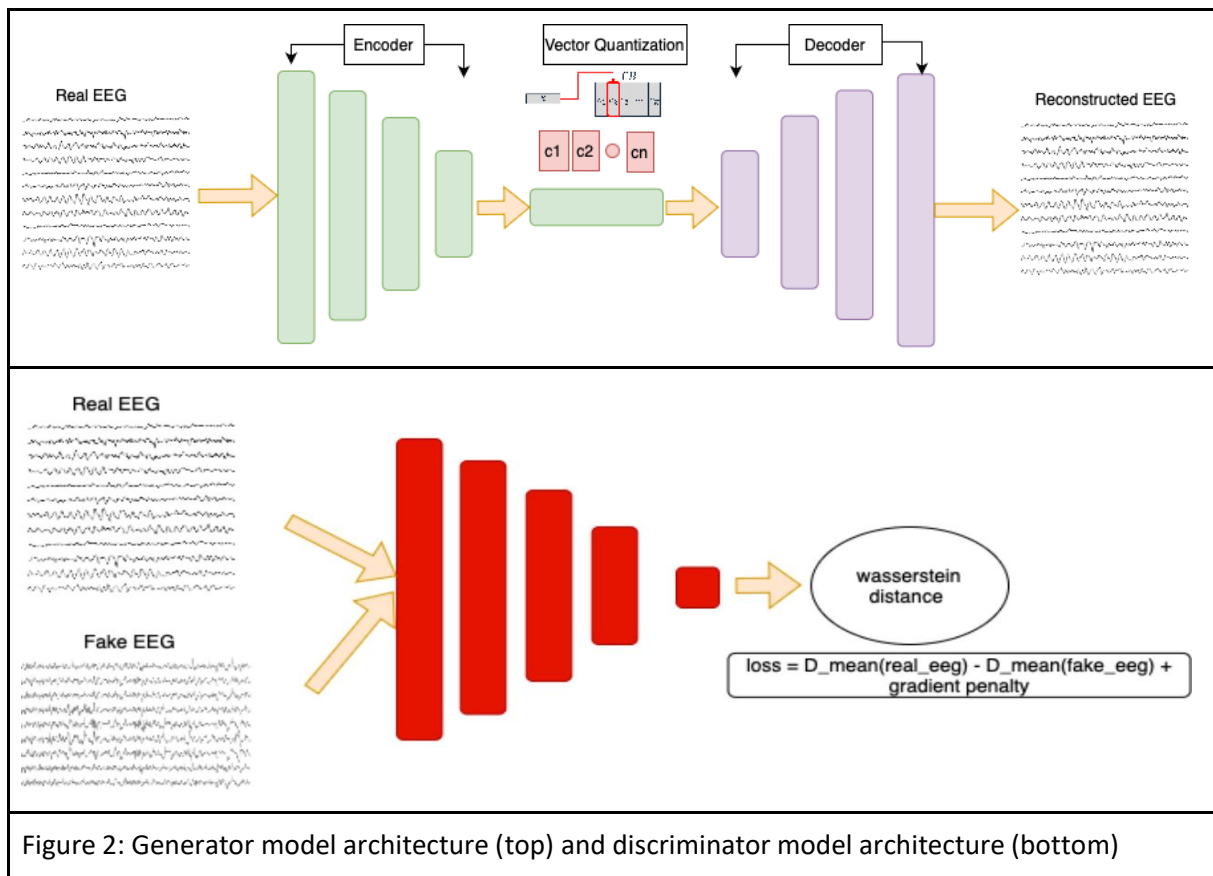


Figure 2 depicts the proposed architecture of generator and discriminator models used in this work. The generator model architecture consists of three different main blocks, namely encoder, vector quantization and decoder block. The encoder block transforms the input real EEG signals into a latent representation. This component compresses the high-dimensional time series EEG data into a more compact latent space, capturing the essential features of the brain activity patterns. The green blocks of varying heights represent different layers of the encoder network (combination of convolutional or attention layers) that progressively extract features from the raw EEG input.

The second block i.e vector quantization is a key differentiating component of this architecture. The Vector Quantization (VQ) module discretizes the continuous latent space into a finite set of vectors from a codebook (represented by c_1, c_2, \dots, c_n). The VQ learning process operates by mapping continuous latent vectors, produced by the encoder, to their closest discrete entries in a predefined

codebook. This quantization step introduces a bottleneck in the latent space, compelling the network to learn more compact and efficient representations of the input data. By enforcing discrete representation learning, vector quantization mitigates the problem of posterior collapse, a common issue in generative models where the latent variables fail to contribute meaningfully to the reconstruction. Consequently, this approach promotes the utilization of diverse and informative latent codes, which enhances both the expressiveness and stability of the generative model.

One important parameter in VQ is the number of codes (or codebook entries) used to represent the input data. A larger number of codes allows the model to capture greater diversity and finer details in the data but also increases the model's complexity and the risk of overfitting. Conversely, a smaller number of codes may lead to poor representation, as the limited capacity cannot capture the full variability of the input [14]. Therefore, it is crucial to select an appropriate number of codes that balances expressiveness and generalization. Based on insights from the literature and the need to represent the high variability in EEG signals, our model uses 256 codebook entries to effectively encode the input features [14]. Each input EEG segment is encoded into a sequence of 50 discrete code indices, with each index referring to one of the 256 available codes. This design imposes a powerful representational bottleneck that enforces efficiency and structure in the latent space.

The combinatorial capacity of this setup is exceptionally large. Specifically, the total number of unique EEG representations that can be formed is:

Total representations = n^r where $n = 256$ (total number of codes) and $r = 50$ (length of the code sequence) i.e. total representation = $256^{50} = (2^8)^{50} = 2^{400}$.

This astronomical number reflects the theoretical maximum number of different EEG patterns that the model can generate or represent using combinations of 50 code positions from a set of 256 code entries. Such a capacity is more than sufficient to represent billions or even trillions of distinct EEG patterns, making it highly suitable for large-scale EEG generation and modeling tasks. Additionally, this vast latent space enables the model to capture a rich diversity of EEG dynamics, while maintaining interpretability and compactness in the latent domain.

The last block of the generator is the decoder. The decoder takes the quantized representation and reconstructs it back into the EEG signal domain. The purple blocks represent the decoder network layers that progressively transform the quantized representation back into the full dimensionality of EEG signals. The decoder learns to generate realistic EEG patterns from the discrete codebook vectors.

Similarly, the discriminator has two inputs (one at a time) real EEGs and synthetic EEGs. The discriminator model architecture consists of the feature extraction network. The red blocks represent the discriminator's neural network layers. This is a set of convolution, normalization and activation layers which processes both real and fake EEG signals, progressively extracts and compresses features through multiple layers, and finally reduces the dimensionality until it reaches a feature representation that can be used for comparison.

The discriminator employs Wasserstein distance as its metric for evaluating the dissimilarity between real and generated EEG distributions, representing a crucial architectural decision in this model. Wasserstein distance, commonly referred to as Earth Mover's Distance, quantifies the minimum "work" required to transform one probability distribution into another, effectively measuring how

fundamentally different the distributions are in a mathematically rigorous way. This metric has proven particularly valuable for training GAN due to its provision of more stable and meaningful gradients compared to traditional GAN loss functions like Jensen-Shannon divergence [15]. The improved gradient properties of Wasserstein distance directly address two persistent challenges in GAN training: mode collapse, where generators produce limited varieties of outputs, and vanishing gradients, which can stall learning [15]. By implementing this distance metric, the EEG generation model benefits from more consistent training dynamics, allowing it to capture the complex, subtle patterns characteristic of neurophysiological data.

The VQ-GAN model incorporates PSD as a complementary loss function alongside the Wasserstein distance, creating a more comprehensive training mechanism for EEG generation. This dual-loss approach addresses a critical aspect of neurophysiological data that is the frequency characteristics that vary distinctly across different brain regions. By explicitly calculating loss based on PSD, the model ensures that generated EEG signals maintain appropriate frequency distributions that match those observed in authentic brain recordings. The elastic power spectral constraint encourages the generator to produce signals with regionally appropriate oscillatory patterns, capturing the nuanced differences in frequency components that exist between frontal, parietal, temporal, and occipital areas of the brain. This spectral awareness is particularly valuable for applications requiring physiologically accurate EEG synthesis, as many neurological conditions and cognitive states are characterized by specific alterations in the power spectrum across distinct brain regions. The addition of this spectral loss component represents a targeted solution to a common limitation in synthetic EEG generation, where time-domain similarities might be achieved while missing the crucial frequency-domain characteristics that neurologist researchers rely on for interpretation and analysis [16].

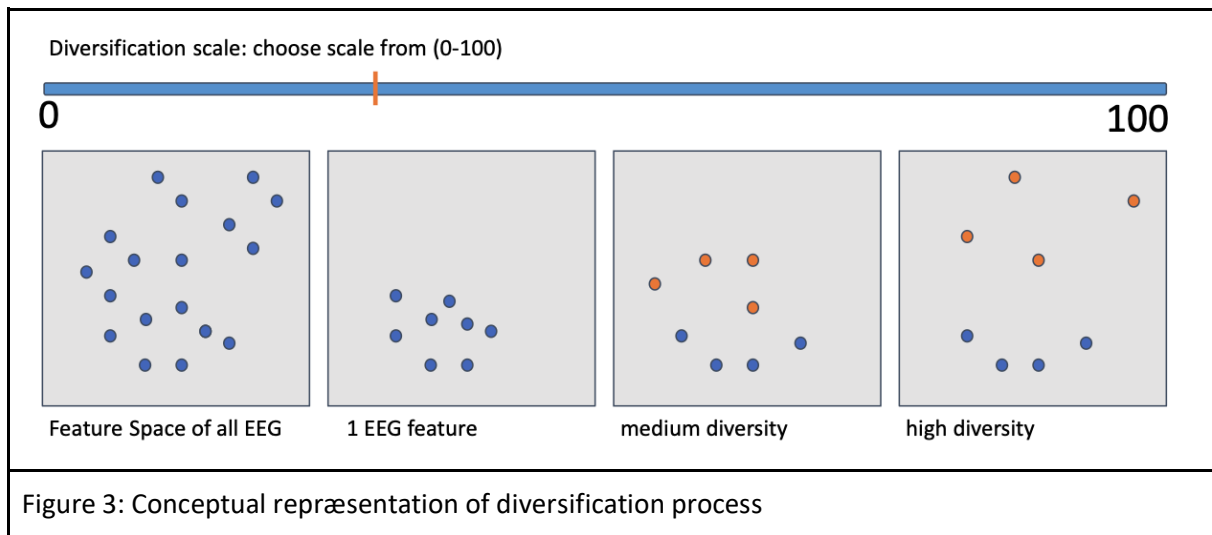
5.3 WP8's Personalized Variations

Our WP8 developed system leverages VQ to create controlled variations of EEG signals through a novel feature replacement approach. At its heart, this system allows us to generate a spectrum of EEG variations ranging from highly similar to significantly diverse, all stemming from a single original EEG recording.

When an original EEG signal is processed through the encoder component of the VQ-GAN, it produces a discrete representation consisting of 50 codebook indices (code features). These features capture the essential characteristics of the original EEG pattern. The proposed diversification system then strategically modifies this representation by selectively replacing certain codebook indices with alternative ones from the learned codebook space of 256 possible features.

The key innovation lies in how these replacements are managed:

1. **Selection Based on Similarity:** Code features aren't replaced randomly but according to a similarity metric, ensuring that replacements maintain physiological plausibility
2. **Controlled Diversification Scale:** By selecting varying degree of similarity, we can precisely control the degree of diversification applied
3. **Feature Space Targeting:** The replaced features create new combinations that exist within the valid manifold of EEG signals



The conceptual Figure 3 effectively visualizes the EEG diversification process through a progressive sequence of four panels that illustrate the transformation of signal features. Beginning with the "Feature Space of all EEG" panel, we see the complete distribution of possible EEG features (256 code) represented as blue dots scattered across a multidimensional space, providing context for the overall domain of neural signal patterns. Moving to the second panel labeled "1 EEG feature," the visualization narrows to focus on a specific region containing fewer blue dots, representing the particular feature cluster of a single EEG recording that serves as our starting point for diversification. The "Medium diversity" panel then introduces the first level of controlled variation, where several orange dots appear among the blue ones, signifying how selected original features have been strategically replaced with similar but distinct alternatives from the codebook, creating moderate signal diversification while maintaining neurophysiological plausibility. Finally, the "High diversity" rightmost panel demonstrates more extensive feature replacement evidenced by a greater proportion of orange dots, illustrating how the system can generate EEG variations that preserve fundamental physiological validity while exhibiting substantially different characteristics from the original signal, all controlled through the diversification scale shown above the panels.

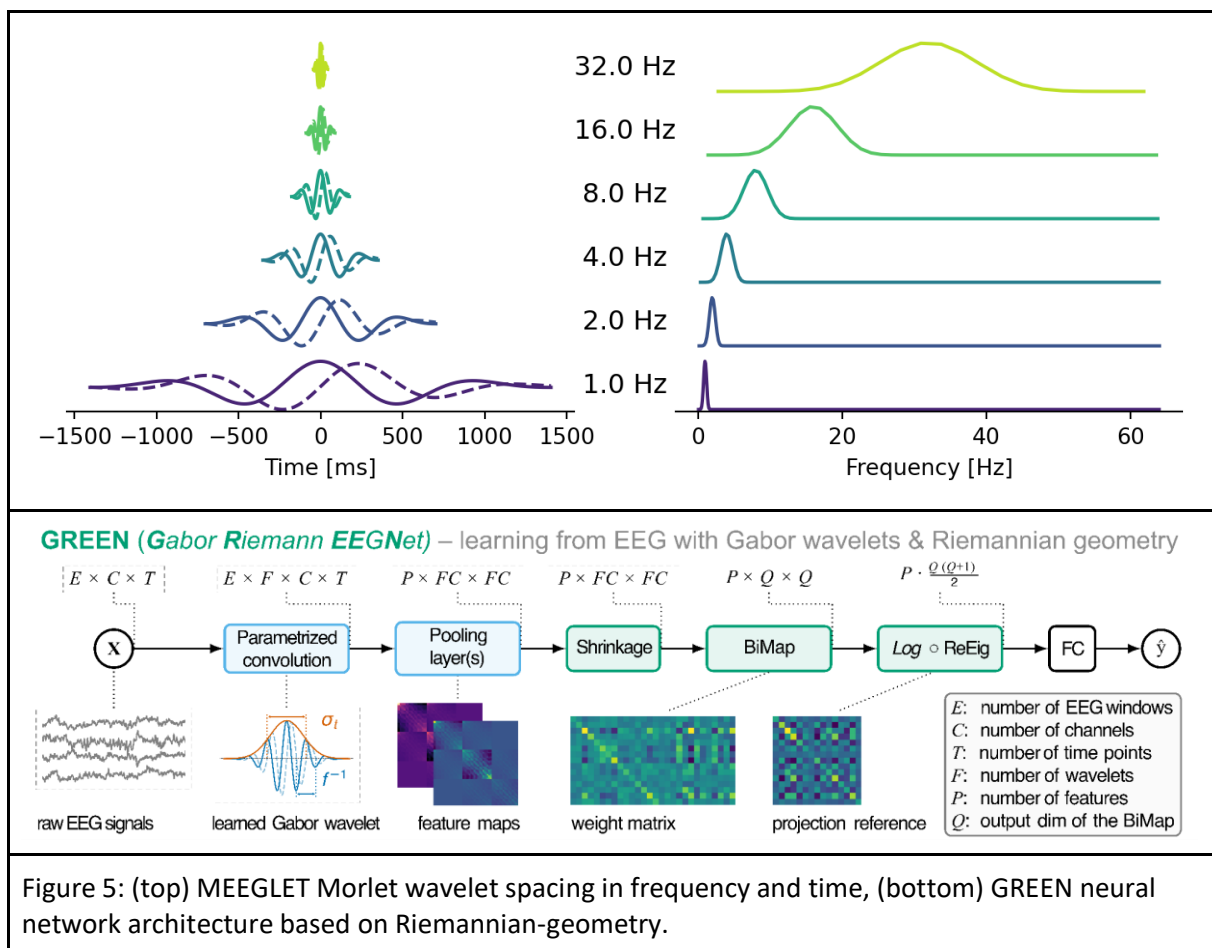
The personalized diversification system delivers four significant advantages that enhance EEG signal processing and analysis capabilities. By generating physiologically plausible EEG variations, the system serves as a powerful data augmentation tool that addresses the common challenge of limited datasets in machine learning applications, effectively expanding the available training data while maintaining neurological validity. The approach excels at creating subject-specific variations that preserve individual neurophysiological characteristics, enabling personalized simulations that account for the unique brain activity patterns of different subjects. Researchers benefit from the ability to systematically explore EEG variations across a precisely controlled similarity spectrum, facilitating controlled experimentation that can reveal insights about neural signal properties and their relationships to cognitive and pathological states. Perhaps most valuable, the system enables targeted feature manipulation, allowing for selective modification of specific EEG characteristics while preserving others, a capability that supports detailed hypothesis testing and the creation of specialized training datasets for developing more robust dementia analysis and diagnostic tools.

5.4 Machine-learning pipeline

From the biological as well as synthetic EEG, we computed covariance matrices using the neuro-meeplet [2] library. By convolution with complex-valued Morlet wavelets, MEEGLET filters the signals to predefined frequencies. By log-linear spacing in frequency space, MEEGLET puts more emphasis on the lower compared to the higher frequency components (Figure 5-top) accounting for the naturally occurring $1/f$ slope in the power spectral density (PSD). Covariance matrices live in a mathematical space where Euclidean metrics are only approximations and therefore require special handling. They are handy for multiple reasons:

1. Their dimension is considerably lower than the one of continuous EEG by collapsing the time dimension.
2. Their diagonal (the variance at each EEG electrode) represents signal power, which allows for convenient analysis of PSD and spatial distribution of power.
3. They serve directly as input to Riemannian-geometry based decoding methods. Recent works have shown that these approaches are very robust and strong contenders on a variety of applications [5, 6, 7, 9, 10]. Therefore, one direction of our research follows this direction.

The covariance matrices are fed to the GREEN2 network (Figure 5-bottom) from the neuro-green [1] library.



In another line of research, we attempt to directly decode properties from the continuous EEG. Therefore, we used the critic of the original GAN model that synthesized the EEG data, as well as two

well known models (Shallow, and Inception) from the braindecode [3] library, that have proved their good performance and applicability on several tasks and datasets [5, 8].

We performed two machine learning tasks: sex classification and p-tau regression. Binary sex classification can be considered a substantially easier task and serves both as a sanity check for our pipelines as well as a practicing target. As the extraction of the p-tau biomarker from EEG has not yet been explored in detail, and as its interactions with EEG are relatively unknown to date, we consider sex decoding a valid and good starting point.

5.5 Explainability

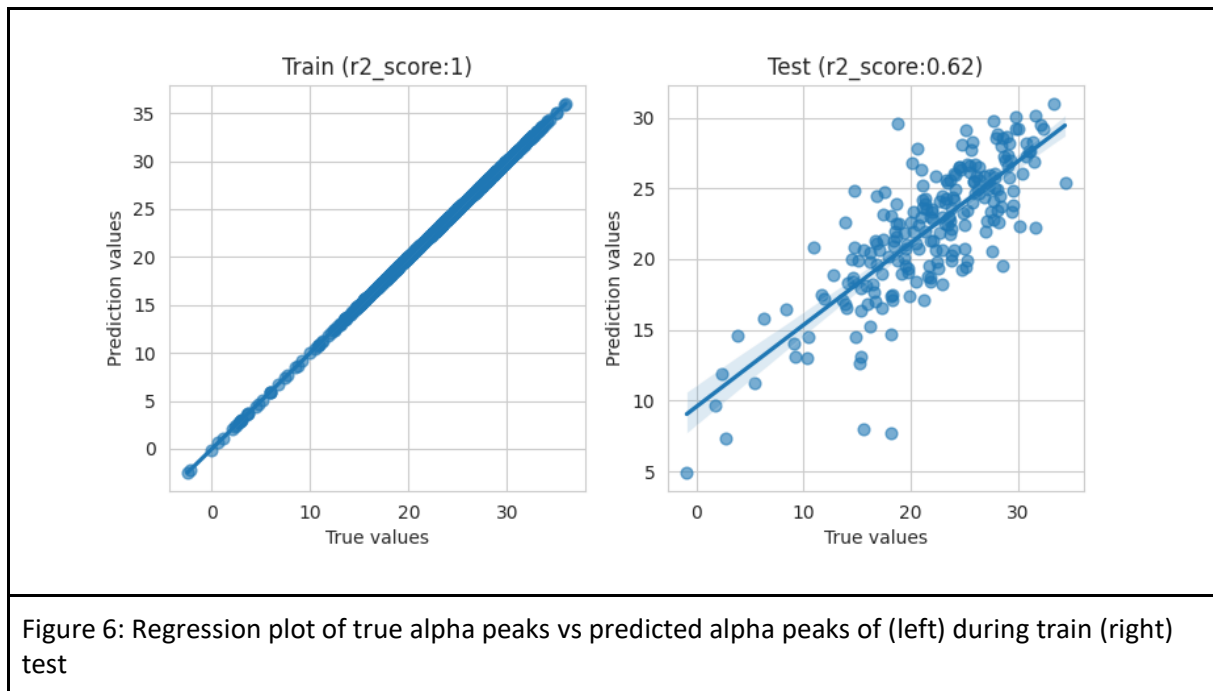
For the explainability of deep learning models, GradCAM (Gradient-weighted Class Activation Mapping) is one of the commonly used methods by researchers as it provides crucial interpretability for deep learning models [17, 18]. Within the domain of EEG classification tasks, it helps by visualizing which temporal and spectral features most influence the model's decisions. GradCAM generates heatmaps highlighting the specific time windows, frequency bands, and electrode locations that contribute most significantly to the classification outcome. This visualization capability transforms otherwise opaque neural network decisions into interpretable insights about which neurophysiological patterns drive the model's predictions [17].

In the context of EEG analysis, GradCAM offers unique advantages by bridging the gap between machine learning outputs and clinical interpretability. By highlighting activation patterns across the frequency and spatial dimension, GradCAM can reveal which frequencies, which brain regions and electrode positions contribute most to detection of specific conditions or cognitive states, aligning machine learning findings with neuroanatomical understanding [19]. This spatial-temporal localization capability makes GradCAM particularly valuable for neurologists and neuroscientists who need to connect model outputs with established knowledge about brain function.

When evaluating synthetic EEG data generated by VQ GAN with a diversification system, GradCAM provides essential validation by comparing the features prioritized in models trained on real versus synthetic data. If models trained on synthetic data produce GradCAM heatmaps that highlight similar spatial-temporal regions as those trained on real data, this provides strong evidence that synthetic signals maintain the physiologically relevant features necessary for classification. Conversely, if GradCAM reveals that models trained on synthetic data focus on different regions or patterns, this can identify specific aspects of the generation process that require refinement. Similarly, the technique can also evaluate how diversification scale affects feature preservation i.e. showing whether highly diversified signals maintain the same discriminative characteristics as the original EEG or introduce novel predictive patterns. This comparison framework makes GradCAM an invaluable tool for validating the fidelity and utility of synthetic EEG data, ensuring that models trained with augmented datasets remain focused on clinically relevant neural patterns rather than artifacts or generation biases.

6. Results

6.1 Extracted EEG code features association with alpha peaks



The regression plots shown in the above Figure 6 demonstrate the relationship between the predicted and true values of the alpha band frequency peak, using features extracted from the vector-quantized (VQ) latent codes of EEG signals. The left panel represents the model performance on the training set, achieving a perfect R^2 score of 1.0, indicating that the model can fully explain the variance in alpha peak frequency based on the learned latent representations during training. The right panel shows the model performance on the test set, where the R^2 score is 0.62, indicating a good but imperfect generalization.

This result provides strong evidence that the discrete code features extracted via vector quantization contain meaningful physiological information, specifically about alpha band characteristics. The model's ability to predict the highest alpha peak frequency is a critical marker in EEG spectral analysis which suggests that the quantized latent space retains relevant frequency-domain information. Even though some variance is lost in the test set (as seen from the scatter), the correlation between predicted and true values remains substantial.

This analysis strongly supports the argument that the 256-code latent representation learned through the VQ-GAN framework captures spectrally relevant EEG features, particularly those associated with alpha activity. This helps to enhance the trustability of extracted code features. This makes the VQ representation not just useful for signal reconstruction, but also a compact, informative feature space for downstream predictive tasks like alpha peak estimation.

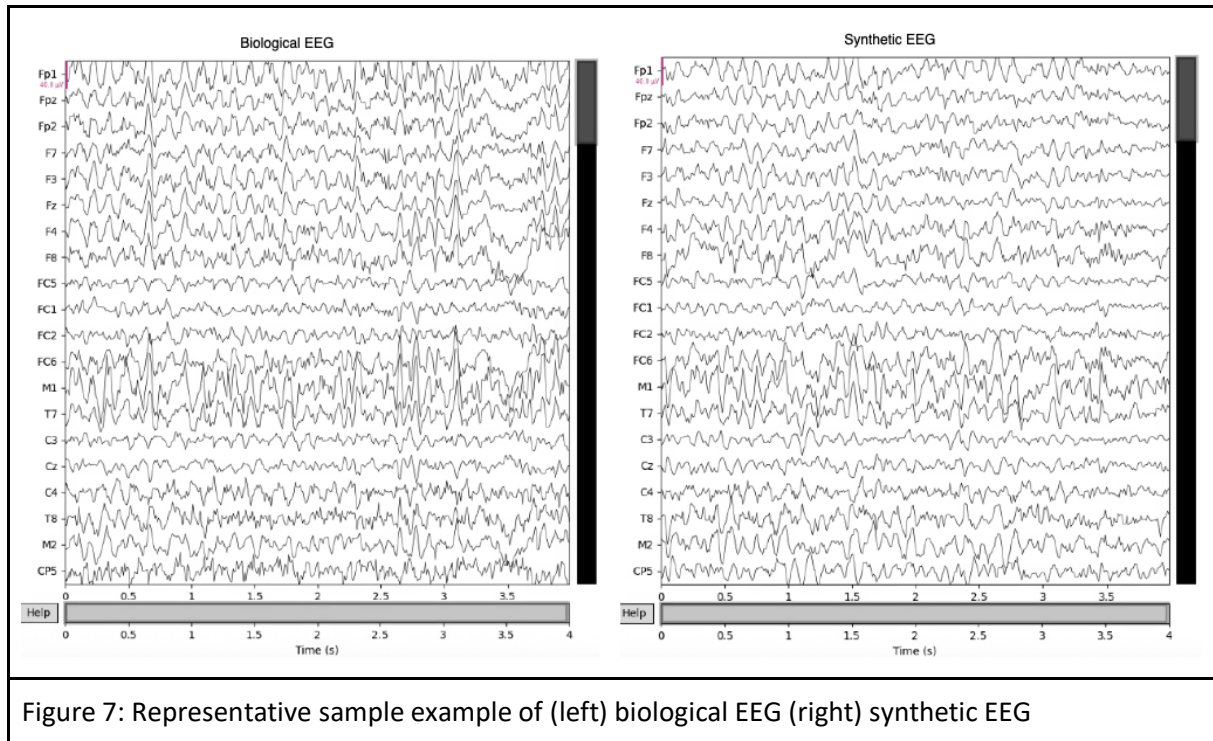
6.2 Personalized EEG variations

6.2.1 Synthetic EEG

We created a total of 48240 synthetic EEGs. They are available through TSD and are located at </ess/p1504/data/durable/AI-Mind-Synthetic-EEG/>. These are variations of the original biological EEG inputs to maintain biological and metadata information for supervised ML. By randomly selecting subsets of 5 to 10 code-level features extracted from each individual EEG recording, our algorithm performs feature replacement based on a predefined similarity-diversity criterion. This augmentation process generates 48,240 synthetic EEG samples, which complement the original dataset of 1,022

biological EEGs. The framework is designed to be scalable, enabling the generation of significantly more synthetic data with minimal manual intervention. While this approach cannot exhaustively capture the infinite variability inherent in all real EEG signal possibilities, it allows us to approximate a broader representation of the underlying feature space. As the volume and diversity of synthetic samples increase, the empirical distribution of our dataset progressively converges—within the bounds of our model assumptions—toward a more comprehensive coverage of plausible EEG variations.

Figure 7 represents one of the representative examples of biological and synthetic EEG plot.



6.2.2 PSD of generated variations

To measure the quality of generated synthetic EEG signals from a neuroscience perspective, we computed the PSD. The Figure 8 illustrates the PSD curves of synthetic EEG signals generated by systematically altering the latent code features in a vector-quantized generative model. These variations offer insight into how changes in the latent space influence spectral characteristics of the generated EEG.

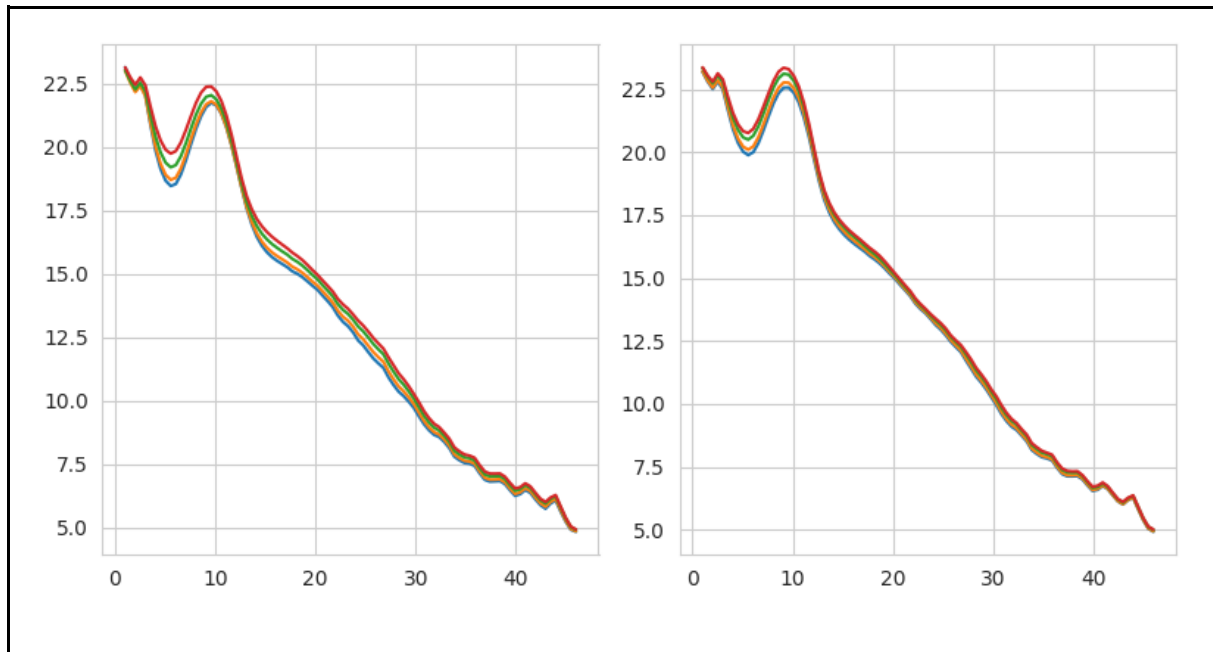


Figure 8: PSD of EEG signals generated after feature replacement. (left) shows the PSD after replacing five features, while (right) shows the PSD after replacing ten features. The x-axis represents frequency, and the y-axis represents PSD ($\log_{10} \times 10$). Different colors indicate variations corresponding to different levels of similarity: blue $\geq 75\%$ similarity, orange $\geq 50\%$ similarity, green $\geq 25\%$ similarity, red similarity $\geq 0\%$ in extracted features.

The left plot illustrates the PSDs of synthetic EEG signals generated by altering five discrete code indices in the latent space. Despite these modifications, the overall spectral structure remains well-preserved across the different signal variations. Notably, there are minor shifts observed in both the peak amplitude and the frequency location, particularly within the alpha frequency range (8–12 Hz), which is a key band of interest in EEG analysis. These variations, however, are relatively subtle, indicating that small perturbations in the code features lead to slight and controlled changes in the signal characteristics. Similarly, the right plot presents the PSDs of synthetic EEG signals generated by modifying ten discrete code indices in the latent space. While the general spectral shape remains consistent, the PSDs show noticeably greater variability in amplitude, especially within the alpha and beta frequency bands.

Compared to the left plot, the spectral peaks in the right plot appear more dispersed and display increased divergence, suggesting that a higher number of altered code features introduces greater diversity in the signal characteristics. This increased variability indicates that the model's latent space is sensitive to the number of modified codes. This allows for the generation of a broader range of realistic EEG patterns, whilst maintaining biologically plausible spectral structures.

6.2.3 PSD of generated variations extracted from MEEGLET

Since we used MEEGLET-based covariance features for the downstream application, it is meaningful to examine the PSD distribution based on MEEGLET. The Figure 9 below shows that the PSD based on the left-hand plot has high variations, especially at the delta and alpha peaks, while the PSD based on the right-hand plot looks very similar over varying degrees of similarity. However, an interesting pattern that we observed in the plots is that when the EEG is augmented by changing only five code features, the alpha peak declines slightly as the degree of similarity decreases. Conversely, the alpha peaks in the right plot consist of different levels of similarity.

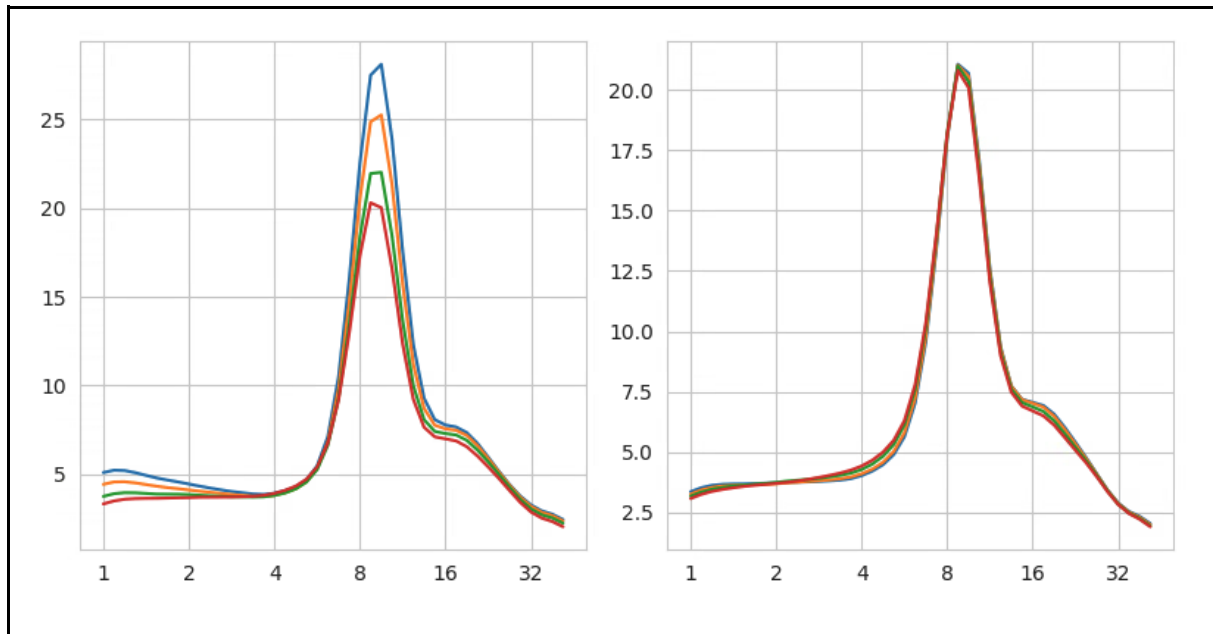


Figure 9: PSD distribution extracted from MEEGLET. The x-axis represents frequency, and the y-axis represents PSD (left) shows the PSD after modifying five code features, while (right) shows the PSD after modifying ten code features. Color variations indicate different levels of similarity: blue $\geq 75\%$ similarity, orange $\geq 50\%$ similarity, green $\geq 25\%$ similarity, red similarity $\geq 0\%$ in extracted features.

6.3 Machine learning prediction

In this final section, we present our machine learning results for two downstream tasks: i) sex classification (a well-researched area) and ii) P-tau level prediction (strongly correlated with dementia but understudied in relation to EEG). We utilized two distinct data modeling approaches: the GREEN2 model, which processes covariance features extracted from both biological and synthetic EEG data, and the Shallow and Inception architecture based deep neural network models, which directly takes continuous raw EEG inputs.

We defined the following 4 ML settings:

- training on biological EEG (80%) and validation on biological EEG (20%)
- training on synthetic EEG (one variant, $n=856$) and validation on biological EEG (same 20%)
- training on biological and synthetic EEG combined (one variant, $n=856$) and validation on biological EEG (same 20%)
- training on biological and synthetic EEG (tree variants, $n=3*856$) and validation on biological EEG (same 20%)

The overview of classifier models is presented in Table 1. The GREEN2 model demonstrates a strong baseline performance of 78.01% when trained exclusively on biological EEG data for the task of sex classification. This high accuracy indicates that the model effectively captures relevant features from real EEG signals. However, when trained solely on synthetic EEG data, the performance drops to 70.63%, suggesting that while the synthetic data retains discriminative patterns, it may not fully capture the subtle intricacies inherent in biological recordings.

Interestingly, combining biological and synthetic EEG data results in a performance improvement, reaching 80.79%, which surpasses the baseline achieved with biological data alone. This indicates that synthetic data provides complementary information that enhances the model's learning capacity. Furthermore, augmenting the dataset with a threefold increase in synthetic EEG samples (i.e., Biological + 3×Synthetic) leads to an even higher accuracy of 82.38%. This performance boost highlights the synthetic data's contribution to increasing variability and reducing overfitting, thereby improving the model's generalization. Overall, these results affirm the utility of synthetic EEG in augmenting limited biological datasets and enhancing classification performance.

Task	Model	Data	Performance (accuracy) %
Sex classification	GREEN2	Biological	78.01
		Synthetic	70.63
		Biological + Synthetic	80.79
		Biological + 3*Synthetic	82.38
	Shallow	Biological	74.85
		Synthetic	68.32
		Biological + Synthetic	75.06
		Biological + 3*Synthetic	76.41
	Inception	Biological	66.85
		Synthetic	73.10
		Biological + Synthetic	75.45
		Biological + 3*Synthetic	77.19

Table 1: Performance overview of three different models and different combinations of biological and synthetic EEG in sex classification.

Similarly, the Shallow model achieves an accuracy of 74.85% when trained on biological EEG data alone, serving as its baseline performance for sex classification. In this setup, when the model is trained only on synthetic data, the model performance is quite low. However, when synthetic data is combined with biological EEG, a modest improvement is observed, with accuracy increasing to 75.06%. Further augmentation using three times more synthetic data (Biological + 3×Synthetic) results in a higher accuracy of 76.41%. These findings suggest that even a relatively simple model can benefit from the inclusion of synthetic EEG, though the performance gains are less pronounced compared to those achieved by more complex architectures like GREEN2.

Finally the third model i.e. the Inception model demonstrates the lowest baseline performance among the evaluated architectures, achieving an accuracy of only 66.85% when trained solely on biological EEG data. This suboptimal result may reflect a mismatch between the model's complexity and the size or variability of the training data. Interestingly, training the model exclusively on synthetic EEG yields a significantly higher accuracy of 72.35%, suggesting that the generated data may possess more structured or noise-reduced features that align better with the Inception architecture. Although results for the combined biological and synthetic dataset are not explicitly reported, further augmentation with three times the synthetic data (Biological + 3×Synthetic) leads to a substantial performance increase, reaching 77.19%. This represents the largest gain across all models and data conditions, highlighting the potential of synthetic EEG to enhance classification performance, particularly for deeper models when biological data is limited or noisy.

Similarly, we summarized the result from the second task i.e. **predicting P-tau values from EEG and covariance features are presented in Table 2.**

Task	Model	Data	Performance (r2_score)
p-tau regression	GREEN2	Biological	-0.073
		Synthetic	-0.01
		Biological + Synthetic	0.0021
		Biological + 3*Synthetic	0.044
	Shallow	Biological	0.0063
		Synthetic	-0.028
		Biological + Synthetic	0.043
		Biological + 3*Synthetic	0.0253
	Inception	Biological	-0.005
		Synthetic	-0.01
		Biological + Synthetic	-0.007
		Biological + 3*Synthetic	0.0007

Table 2: Performance overview of three different models and different combinations of biological and synthetic EEG in p-tau regression.

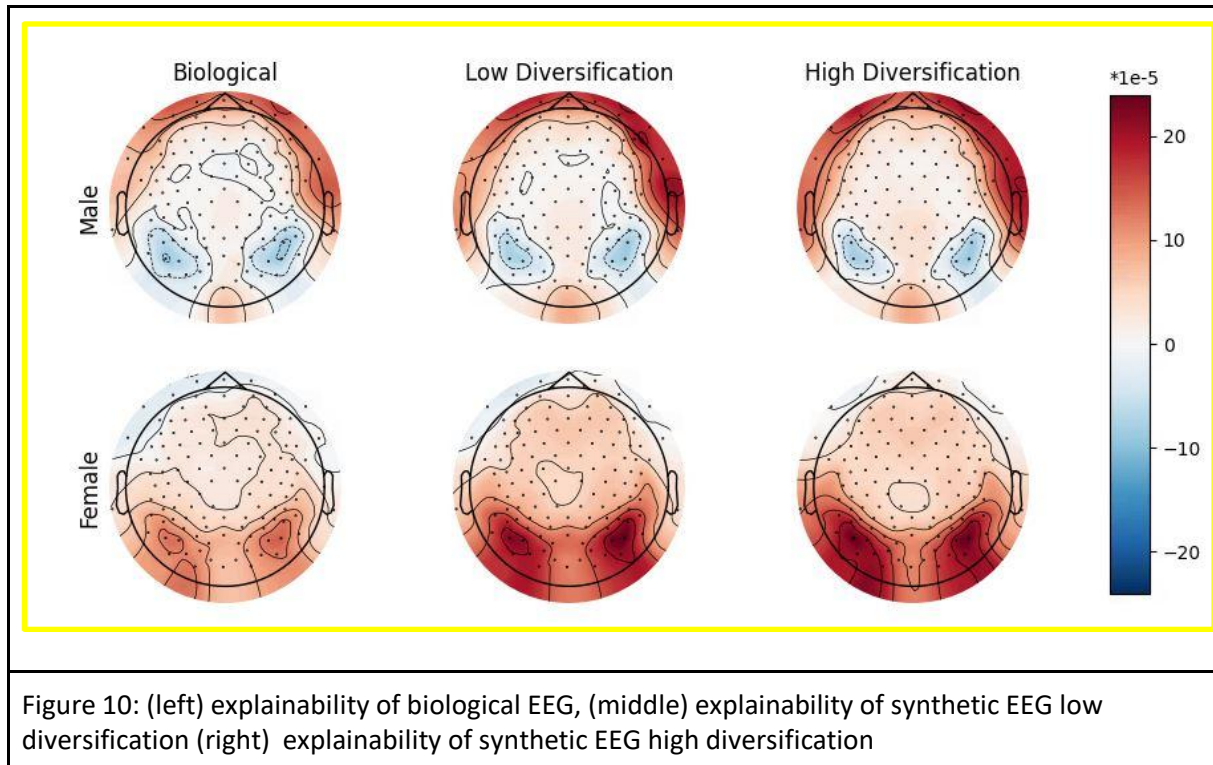
The results presented in Table 2 illustrate the performance of three neural network architectures (GREEN2, Shallow, and Inception) for P-tau level prediction using various combinations of biological and synthetic EEG data, with performance quantified by the R^2 score. Overall, the regression task yields substantially modest performance across all models and data configurations, with R^2 values ranging from negative (-0.073) to marginally positive (0.044), indicating limited predictive capability for P-tau levels from EEG signals. The GREEN2 model demonstrates the widest performance variance, yielding the lowest score with biological data alone (-0.073) but achieving the highest overall performance (0.044) when biological data is augmented with triple synthetic data. The Shallow model exhibits the most consistently positive performance, with all configurations yielding non-negative R^2 scores and achieving its peak (0.043) using the biological plus synthetic combination.

The Inception architecture demonstrates consistently poor performance across all data configurations for this regression task, with R^2 scores remaining near zero or negative throughout the experiment. Notably, all three models show at least some improvement when biological data is augmented with synthetic data, particularly at higher synthetic-to-biological ratios, suggesting that synthetic data may contribute meaningful signals despite the overall modest performance. The consistently low R^2 scores across all models indicate that EEG signals may have limited direct correlation with P-tau levels, or that more sophisticated modeling approaches may be necessary to extract the relevant patterns. These findings represent an initial exploration into the understudied relationship between EEG measurements and P-tau biomarkers, establishing a baseline for future research while suggesting that this remains a challenging domain for predictive modeling.

6.4 Model Explainability

The topographic maps presented in Figure 10 show GradCAM-based explanations of GREEN2 network attention patterns in EEG-based sex classification. These maps reveal distinct spatial patterns that appear to align with established neurophysiological differences between male and female brains.

In the male topographic map (left), there is pronounced activation (red coloration) in the frontal and anterior temporal regions, with notable blue areas (negative contribution) in bilateral occipito-parietal regions. Conversely, the female topographic map (right) displays an inverted pattern, with stronger activation in the posterior regions, particularly in bilateral parieto-occipital areas, while showing less pronounced activity in frontal regions.



7. Future work and limitations

The results presented in this deliverable demonstrate the potential of our approach, and we are eager to further refine and validate our findings in the upcoming months, in conjunction with the bias and mitigation tasks outlined in the forthcoming Deliverable at M42. While our current analysis has focused on the time dynamics and power characteristics of EEG signals, we recognize the importance of exploring additional aspects, such as phase and connectivity.

Although our initial results are promising, we acknowledge that the hyperparameters of our models have not been exhaustively optimized, and we anticipate that further tuning will lead to improved performance, particularly for the more complex models. The single evaluation runs on a 20% validation split of the AI-Mind data provide a foundation, but we plan to expand our investigation to include 5-fold cross-validation, which will enable us to more comprehensively assess the robustness of our results.

Furthermore, we recognize that the two pipelines developed for covariance matrices and continuous EEG data, although following the same logic, may not be fully harmonized. To be more specific, while they both use 20% of biological EEG for validation, the included participants in the validation sets may differ. To address this, we will prioritize the harmonization of these pipelines, ensuring that they are consistent in terms of data split and validation procedures. This refinement will enable us to make fair comparisons between the different approaches.

In addition to the EEG synthesis method presented here, which can be viewed as a form of data augmentation, we plan to explore an alternative approach, which generates unlabeled EEG signals from noise input. Although these signals will not be compatible with our supervised machine learning pipelines, we are interested in investigating the potential of pretraining models in an unsupervised fashion on these data, which may lead to novel insights and improved performance.

As we move forward, we will iteratively refine our EEG synthesis pipeline, incorporating new insights and findings from our ongoing research. Our goal is to create an even larger dataset of synthetic EEGs, which will enable us to explore a wider range of applications, including supervised and unsupervised machine learning. We are particularly interested in investigating the effect of synthetic data on cognition prediction, in addition to sex classification and p-tau regression.

Finally, we recognize the importance of interpretability in our methods and plan to explore further explainable AI techniques to identify brain regions and frequency bands associated with p-tau levels and cognitive scores. By elucidating the significance and meaning of the learned latent GAN features, we aim to provide a more comprehensive understanding of our approach and its potential applications.

7. Conclusion

Our efforts to generate synthetic EEGs have yielded a substantial dataset. Preliminary analysis of the synthetic data, including time series plots and PSDs, suggests that the generated EEGs exhibit similar characteristics to their biological counterparts, indicating a high degree of realism.

This is supported by our models' performance trained solely on synthetic data. While the sex decoding performance is substantially worse compared to the performance obtained using biological EEG only, it is still significantly better than chance level, implying that meaningful correlations in the synthetic EEG and the targets were preserved through synthesis.

Moreover, our results show that combining biological and synthetic EEGs for training purposes can lead to a general improvement in performance. With GREEN2 we observed a +2% increase in accuracy when using a combination of both data types. Furthermore, we found that adding three times the amount of synthetic EEG data can lead to an additional +2% improvement in performance, suggesting that the generated data can be a valuable resource for enhancing model accuracy. Consistently across all models, we observed the best accuracy when combining the biological EEG with all available synthetic EEG, hinting on potential benefits of generating even more synthetic EEG.

The results on p-tau regression leave a lot of room for improvement. However, given the novelty of the marker and the investigations regarding correlations between EEG and p-tau, we are not discouraged but challenged to find the underlying link. We are convinced that the observed performance is not a pure issue of synthetic data, but of the p-tau prediction task itself, which will require more advanced modeling and optimization.

Overall, our results demonstrate the potential of synthetic EEGs as a valuable tool for enhancing the accuracy and robustness of machine learning models, and we are excited to explore further applications and refinements of this technology in future research.

Acknowledgements

The authors acknowledge the use of OpenAI's ChatGPT-4o for assistance in language editing, which has been used in order to improve the clarity and readability of the manuscript. The content of this report lies in the sole responsibility of the authors.

References

- [1] Paillard, Joseph, Jörg F. Hipp, and Denis A. Engemann. "GREEN: A lightweight architecture using learnable wavelets and Riemannian geometry for biomarker exploration with EEG signals." *Patterns* (2025).
- [2] Bomatter, Philipp, et al. "Machine learning of brain-specific biomarkers from EEG." *EBioMedicine* 106 (2024).
- [3] Schirrneister, Robin Tibor, et al. "Deep learning with convolutional neural networks for EEG decoding and visualization." *Human brain mapping* 38.11 (2017): 5391-5420.
- [4] Gramfort, Alexandre, et al. "MEG and EEG data analysis with MNE-Python." *Frontiers in Neuroinformatics* 7 (2013): 267.
- [5] Gemein, Lukas AW, et al. "Machine-learning-based diagnostics of EEG pathology." *NeuroImage* 220 (2020): 117021.
- [6] Engemann, Denis A., et al. "A reusable benchmark of brain-age prediction from M/EEG resting-state signals." *Neuroimage* 262 (2022): 119521.
- [7] Barachant, Alexandre, et al. "Multiclass brain-computer interface classification by Riemannian geometry." *IEEE Transactions on Biomedical Engineering* 59.4 (2011): 920-928.
- [8] Gemein, Lukas AW, et al. "Brain age revisited: Investigating the state vs. trait hypotheses of EEG-derived brain-age dynamics with deep learning." *Imaging Neuroscience* 2 (2024): 1-22.
- [9] Wilson, Daniel, et al. "Deep Riemannian Networks for end-to-end EEG decoding." *Imaging Neuroscience* 3 (2025): imag_a_00511.
- [10] Sabbagh, David, et al. "Predictive regression modeling with MEG/EEG: from source power to signals and cognitive states." *NeuroImage* 222 (2020): 116893.
- [11] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.
- [12] Song, J., Zhai, Q., Wang, C., & Liu, J. (2024). EEGGAN-Net: enhancing EEG signal classification through data augmentation. *Frontiers in Human Neuroscience*, 18, 1430086.
- [13] Venugopal, A., & Resende Faria, D. (2024). Boosting EEG and ECG Classification with Synthetic Biophysical Data Generated via Generative Adversarial Networks. *Applied Sciences*, 14(23), 10818.
- [14] Van Den Oord, A., & Vinyals, O. (2017). Neural discrete representation learning. *Advances in neural information processing systems*, 30.
- [15] Arjovsky, M., Chintala, S., & Bottou, L. (2017, July). Wasserstein generative adversarial networks. In *International conference on machine learning* (pp. 214-223). PMLR.

- [16] Flores-Sandoval, A. A., Davila-Pérez, P., Buss, S. S., Donohoe, K., O'Connor, M., Shafi, M. M., ... & Fried, P. J. (2023). Spectral power ratio as a measure of EEG changes in mild cognitive impairment due to Alzheimer's disease: a case-control study. *Neurobiology of aging*, 130, 50-60.
- [17] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618-626).
- [18] Wang, S., & Zhang, Y. (2023). Grad-CAM: understanding AI models. *Comput. Mater. Contin*, 76(2), 1321-1324.
- [19] Li, Y., Yang, H., Li, J., Chen, D., & Du, M. (2020). EEG-based intention recognition with deep recurrent-convolution neural network: Performance and channel selection by Grad-CAM. *Neurocomputing*, 415, 225-233.

Disclaimer

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101058516. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or other granting authorities. Neither the European Union nor other granting authorities can be held responsible for them.