UNIVERSITÀ DEGLI STUDI DI VERONA

DEPARTMENT OF
COMPUTER SCIENCE

GRADUATE SCHOOL OF
NATURAL SCIENCE AND ENGINEERING

DOCTORAL PROGRAM IN
COMPUTER SCIENCE

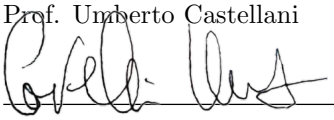
Cycle XXXVIII

TITLE OF THE DOCTORAL THESIS

Through the motions of human vigilance: geometry-aware generation and forecasting of EEG spectral dynamics

S.S.D. IBIO-01/A

Coordinator: Prof. Umberto Castellani

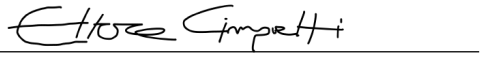
Signature 

Tutor: Prof. Silvia Francesca Storti

Signature 

Co-Tutors: Prof. Gloria Menegaz
Prof. Alberto Castellini

Doctoral Student: Dott. Ettore Cinquetti

Signature 

Ettore Cinquetti

Through the motions of human vigilance: geometry-aware generation and forecasting of EEG spectral dynamics

Ph.D. Thesis, XXXVIII cycle

Università degli Studi di Verona
Dipartimento di Informatica

Advisor:
Prof. Silvia Francesca Storti

Co-Advisors:
Prof. Gloria Menegaz
Prof. Alberto Castellini
Prof. Umberto Babiloni

Università di Verona
Dipartimento di Informatica
Strada le Grazie 15, 37134 Verona
Italy

Alla mia supervisor, Silvia Francesca Storti, nel ruolo di guida. A tutti i componenti e le componenti del BraiNavLab, e al gruppo di ricerca Potioc. Agli amici, vecchi e nuovi. Alla mia famiglia, e a Deborah, per avermi supportato e sopportato.

Summary

This thesis unfolds across eight chapters, detailing the importance of EEG in safety-critical workplace settings, possible methods for forecasting fluctuations in human vigilance, and a novel geometrical approach to data augmentation aimed at addressing the scarcity of biological signals.

The **Introduction** anticipates the thesis structure. Here, the intricacies of EEG and its practical applications, both in and outside the medical domain, are presented. Focus is put on the concept of vigilance, the human capability to sustain prolonged attention, and its implications for repetitive industrial tasks. To predict upcoming vigilance fluctuations, deep learning forecasting models are introduced, defining the *first aim* of this thesis: to anticipate upcoming α oscillations several seconds in advance. We postulate that successful predictions can be integrated in real-time monitoring systems to improve safety in critical workplace conditions - examples include industrial plant control, driving, or assembly-line supervision. To grant these models a proper informational basis, generative AI strategies are examined. In particular, the *second aim* of this thesis is stated: to develop a geometrically informed loss function for GANs to avoid mode collapse and generate diverse synthesized samples of envelope EEG data.

EEG-based Brain Computer Interfaces details the biomarker extracted from EEG; properties of the α, β, θ frequency bands are explained and contextualized from the perspective of human vigilance. α power, in particular, modulates along attentional engagement and is the most commonly used index of vigilance. EEG monitoring can improve safety in contexts where human sustained attention is critical to avoid dangerous mistakes; examples include driving, air traffic control, and Industry 4.0-5.0. After an overview of EEG-related safety methods in industrial settings, strong emphasis is placed on the effects of engagement, attention, and vigilance in a human-robot collaboration framework. Subsequently, a custom task designed to induce drowsiness and boredom is introduced: subjects monitor the movements of a robotic arm for up to 23 minutes across 120 trials. Results show that α fluctuates systematically during robotic arm operations, spiking seconds before each operation is completed and waning during brief relaxation between trials. The collected EEG is used as the basis for all the DL experiments reported in this thesis. Considerations follow regarding current challenges in EEG data acquisition, including the difficulties of acquisition in realistic, uncontrolled settings and the need to balance signal quality and subjects' comfort.

Multivariate Time Series Forecasting tackles the problem of predicting the non-stationary envelopes of EEG frequency bands. VAR models, despite their linear nature, offer a well-supported baseline to weigh against modern DL architectures. Among these, three classes were considered: Recurrent Neural Networks (RNNs), Convolutional Networks, and Transformers. RNNs are composed of a memory unit that gets progressively updated as new data streams in; improved architectures (such as GRU and LSTM) regulate the data flow through the memory cell, and were used in subsequent experiments. A specific RNN implementation in Encoder-Decoder form establishes a milestone in the DL literature. As a convolutional model, we used TCNs, which are specifically adapted to the temporal domain and leverage exponentially dilated filter banks to process long-term time series. Finally, the more recent Transformer PatchTST architecture offers extremely promising performance thanks to a powerful, channel-independent attention mechanism. Specifics for all these models, among other possible candidates, are reported throughout this chapter.

Experimental Results in Vigilance Forecasting reports a comparative evaluation of statistical

and Deep Learning architectures, namely VAR, TCN, GRU, Encoder-Decoder GRU (ED-GRU), and PatchTST, for forecasting vigilance-related α envelope oscillations. To isolate the trend of mental fatigue, the continuous α power was extracted via Continuous Wavelet Transform (CWT) and heavily decimated to 4 Hz. Centro-parietal channels were chosen (C3, C4, P3, P4), which notably exhibited a 0.09 Hz spectral peak matching the pacing of the robotic task.

Models were evaluated via leave-one-subject-out cross-validation over an 11-second (44-sample) forecasting horizon and exhibited vastly different predictive capabilities. The VAR baseline degraded rapidly beyond 1.5 seconds, while standard DL models like TCN and single-cell GRU failed to capture complex future dynamics, flatlining or naively repeating the last input. In contrast, the ED-GRU architecture demonstrated robust performance, accurately predicting α behavior up to 5.5 seconds into the future (half a robotic operation) with an R^2 of 0.726, while preserving the signal's dominant spectral frequencies.

Ultimately, the Transformer-based PatchTST model yielded state-of-the-art results. It significantly outperformed all recurrent and convolutional baselines, achieving near-perfect short-term accuracy and highly reliable mid-term forecasts (5.5 seconds, $R^2 = 0.8675$). While its performance eventually decays at extreme horizons (11 seconds), it clearly validates the superiority of attention mechanisms for the long-term forecasting of non-stationary cognitive states.

Generative Methods for EEG signals contextualizes the use of Generative Artificial Intelligence (GenAI) as a solution to the severe data scarcity, class imbalance, and privacy bottlenecks inherent to BCI research. While traditional data augmentation relies on deterministic, domain-agnostic manipulations (e.g., noise injection or window splicing), GANs employ a game-theoretic minimax optimization between a Generator and a Discriminator to explicitly learn and sample from the true high-dimensional data manifold. After tracing the evolution of these architectures, from standard image-based models to time-series specific variants (e.g., RCGAN, TimeGAN), the critical algorithmic vulnerabilities of GANs are reported, predominantly their susceptibility to training instability and mode collapse.

Since the fidelity of synthetic temporal data can not be captured by a single universal metric, evaluation is framed along three complementary axes: i) statistical fidelity (ACF and PDS consistency), ii) functional utility (downstream performance improvements), and iii) qualitative inspection via dimensionality deduction (t-SNE, PCA). A domain-specific adaptation of the Fréchet Inception Distance (FID) is introduced using a pre-trained ED-GRU network to extract high-level temporal embeddings and compute the Wasserstein-2 distance between the real and synthetic Gaussian distributions. The second involves functional utility, which is empirically measured by whether the generated data actively improves the performance of downstream supervised models, such as a forecaster. Finally, qualitative inspection can be accomplished using dimensionality reduction techniques (such as t-SNE and PCA) to visually assess manifold overlap. This heuristic inspection can confirm distributional diversity and visually diagnose failures such as mode collapse.

Joining these three perspectives provides sufficient quantitative benchmarking tools to validate the EEG generation models developed in the subsequent chapters.

Evaluation of Generative Models details the design, implementation, and validation of an RNN-CNN GAN framework for synthesizing time-variant EEG frequency band envelopes (α, β, θ). Validated across two distinct cognitive paradigms, a public resting-state dataset (SPIS) and a custom industrial Vigilance dataset, the architecture uses a bidirectional GRU Generator to model long-term sequential dependencies, paired with a 1D-CNN Discriminator to evaluate temporal authenticity. To address BCI data scarcity, the model's robustness is tested across progressively restricted data splits (100%, 50%, and 25% of available training data).

The synthetic data is evaluated using statistical, performance improvement, and embeddings assessments; several results follow. First, the GAN successfully replicates realistic amplitude ranges (L2 distance) and preserves complex inter-channel temporal correlations (Cross-Correlation Function). This structural fidelity degrades under severe data scarcity (25% split). Second, augmenting the training pool with GAN-generated samples yields statistically significant improvements in downstream forecasting tasks, measured via MAE and SMAPE. Specifically, the forecaster is an ED-GRU model. The exception concerns the 25% split of the Vigilance data, for which no significant improvement has been observed. On a final note, a naive baseline based on linear interpolation drastically reduced forecast-

ing accuracy, demonstrating that simple signal mixing is insufficient for complex biological time series. Finally, the study reveals a relationship between the domain-adapted forecaster FID and practical data utility. Linearly interpolated data achieves near-zero FID scores, but because it simply replicates already seen samples, it offers no novel diversity, rendering it useless for downstream training. Conversely, failed GAN generations (due to data scarcity) result in large FID outliers. The successful GAN augmentations achieve moderately low FID scores ($\sim$0.5–1.4), proving that optimal synthetic data must strike a delicate balance: it should remain statistically faithful to the physiological manifold while introducing enough novel variation to prevent model overfitting.

Ultimately, this chapter demonstrates that the proposed GAN framework can successfully synthesize physiologically valid and functionally useful EEG features. Furthermore, it establishes that a balanced, moderately low FID score serves as a reliable heuristic for identifying diverse synthetic data that can properly improve downstream BCI forecasting models.

Geometry-based Regularization of GANs for EEG Signal Generation establishes the mathematical and geometric foundations required to robustly model the complex spatio-temporal dynamics of EEG signals. Because neurophysiological data mapped to structured representations inherently reside on non-Euclidean manifolds, applying standard Euclidean metrics often results in severe geometric distortion. To address this, the chapter formulates a Riemannian approach, based on the Symmetric Positive-Definite (SPD) covariance matrix representation of EEG windows.

The rationale for this transformation is grounded in the Wiener-Khinchin theorem: by extracting the Sample Covariance Matrix (SCM), one effectively computes the zero-lag autocorrelation of the signals. Consequently, the diagonal elements capture the total spectral energy of individual channels, while the off-diagonal elements encode the zero-phase spatial synchronization between different brain regions.

Geometrically, the set of all $N \times N$ SPD matrices forms a differentiable convex cone. To perform operations that respect the curvature of this space, the Affine-Invariant Riemannian Metric (AIRM) is introduced as the optimal distance measure. The AIRM calculates the true geodesic distance and possesses the desirable property of affine invariance, making it resilient to common EEG artifacts such as sensor displacement. However, because computing the AIRM is computationally expensive and heavily penalizes matrices approaching singularity, alternative metrics such as the Log-Euclidean Metric (LEM), Cholesky Distance, and Stein Divergence are discussed as computationally efficient proxies.

While the standard SPD captures spatial connectivity, it discards the phase and temporal directionality of the frequency spectrum. To overcome this, the framework is extended to the Augmented Covariance Matrix (ACM). By incorporating time-lagged cross-correlations into a block-Toeplitz structure, the ACM recovers the temporal phase relationships lost in standard zero-lag covariance. Specifically, this extension recovers the linear phase relationships within the same frequency bands across channels, providing a more detailed time-domain representation of the cross-spectrum and reconstructing the inherent directionality of neural communication.

Still, there are computational trade-offs that should be made clear when using ACM. As the number of incorporated lags grows, the matrix dimensions expand quadratically. While downsampling the lag intervals and truncating the maximum lag (L) make the ACM computationally tractable, they inevitably introduce signal processing artifacts.

Building upon this geometric understanding, the chapter details the central methodological contribution of the thesis: a geometry-based generator regularization. To prevent mode collapse, the standard Generator cross-entropy loss is augmented with a penalty term based on the SPD manifold. Specifically, the loss function aims to maximize the minimum pairwise Riemannian distance between generated samples within a batch. This forces the Generator to synthesize distinct, well-distributed samples that cover a broader area of the Riemannian space.

To implement this efficiently, the AIRM is computed in PyTorch using a fully differentiable solver and eigenvalue decomposition, bypassing the numerical instability and computational cost of explicit matrix inversion. Finally, to handle this new multi-objective optimization problem, which risks being susceptible to gradient interference, Projected Conflicting Gradients (PCGrad) and dynamic magnitude normalization are considered. This gradient surgery ensures that conflicting update directions are orthogonalized, allowing the Generator to simultaneously improve signal realism and distributional diversity without one

objective overwhelming the other.

Finally, two methods for assessing generative variety are introduced. The first involves visual inspection via UMAP to embed the high-dimensional covariance matrices into a 2D/3D space. A successful Generator produces synthetic clusters that visually overlap with the real data distribution, whereas isolated clusters indicate mode collapse. The second quantitatively measures spread in the Riemannian space by partitioning the real SPD matrices using K-means clustering; synthetic samples are then assigned to these real centroids, and the Jensen-Shannon (JS) divergence between the resulting cluster membership distributions provides a normalized metric of global mode coverage.

The proposed geometry-regularized GAN is evaluated on the downsampled industrial α-band dataset to assess its efficacy under realistic acquisition conditions. This novel approach was compared against a standard, unregularized GAN baseline across qualitative, geometric, and functional results.

Visual inspection of the generated time series highlights the standard GAN's susceptibility to mode collapse. Conversely, the regularized Generator produces a variety of realistic oscillations. This diversity was quantitatively validated within the Augmented Covariance Matrix (ACM) space. UMAP embeddings revealed that while the standard GAN formed isolated, overspecialized clusters, the geometry-regularized synthetic data overlapped closely with the true data manifold. Furthermore, unsupervised Riemannian K-means clustering confirmed that the regularized approach drastically minimized the Kullback-Leibler (KL) and Jensen-Shannon (JS) divergences between real and synthetic distributions.

The functional utility of the generated data is proven by retraining the PatchTST forecasting model on the augmented dataset. Short-term prediction accuracy (11 steps) is slightly in favour of the augmented model. At long-term horizons (33 to 44 steps), the augmented model exhibited better MAE and R^2 values, in addition to lower error variance, on otherwise noisy channels (e.g., C3).

This framework represents a noteworthy departure from standard Euclidean generative models. By explicitly forcing the Generator to maximize the pairwise Affine-Invariant Riemannian Metric (AIRM) distance within generated batches, it provides a mathematical solution to mode collapse by considering the curvature of physiological data. In practice, this serves as an upstream, plug-and-play augmentation module for BCI pipelines, providing a way to expand small datasets, such as calibration data.

Currently, the methodology presents two primary limitations. First, as a self-supervised approach, the model interpolates excellently but cannot extrapolate out-of-distribution subject patterns. Second, the regularized loss imposes a steep computational bottleneck. Further research should consider both intersubject interpolation and computational improvements.

Discussion recaps the major points of this thesis, from the forecasting of EEG frequency envelopes to the geometrical augmentation of small datasets.

Transformer-based architectures are shown to successfully forecast α-band vigilance markers up to 5.5 seconds in advance, providing a predictive horizon that overcomes filtering latencies and establishing a viable foundation for preemptive interventions in safety-critical environments.

The proposed geometrically informed augmentation method offers a novel way to expand biological data, as confirmed by a brief survey of the ACM literature. Empirical evaluations confirmed that this geometric regularization successfully mitigates mode collapse and significantly enhances the robustness of downstream forecasting models.

While the exact computation of pairwise Riemannian distances imposes a steep computational bottleneck, this research outlines a clear roadmap for future scalability. By leveraging alternative proxy metrics, intelligent latent-space sampling, or more sophisticated topological decompositions, especially considering the Siegel Disk space, the proposed geometric generation paradigm can be optimized to support real time, patient-specific BCI calibration.

Finally, the proposed geometrical method to combat the scarcity of BCI training data is framed from the Geometry-Informed Deep Learning (GIDL) perspective, contextualizing it inside the broad research field of Deep Learning.

Sommario

Questa tesi si articola in otto capitoli, illustrando l'importanza dell'EEG in ambienti di lavoro critici per la sicurezza, i possibili metodi per prevedere fluttuazioni della vigilanza umana, e un approccio geometrico innovativo per la generazione di dati volto ad affrontare la carenza di segnali biologici.

L'**Introduction** anticipa la struttura della tesi. Qui vengono presentate le complessità proprie dell'EEG e le sue applicazioni pratiche, sia all'interno che all'esterno di realtà mediche. Si introduce il concetto di vigilanza, la capacità umana di sostenere attenzione per periodi di tempo prolungati, e le sue implicazioni in attività industriali ripetitive. Per prevedere imminenti fluttuazioni della vigilanza, vengono introdotti modelli di previsione basati sul deep learning, definendo il *primo obiettivo* di questa tesi: anticipare le future oscillazioni della banda α con diversi secondi di anticipo. Ottenere previsioni accurate può favorire sistemi di monitoraggio in tempo reale per migliorare la sicurezza in condizioni di lavoro critiche; esempi includono il controllo di impianti industriali, la guida o la supervisione di catene di montaggio. Per fornire a questi modelli un'adeguata base di informazioni, vengono esaminate strategie di intelligenza artificiale generativa. In particolare, viene dichiarato il *secondo obiettivo* di questa tesi: sviluppare una funzione di perdita (loss function) geometricamente informata per le GAN, al fine di evitare il collasso dei modi (mode collapse) e generare campioni sintetici diversificati di feature spettrali EEG.

EEG-based Brain Computer Interfaces descrive in dettaglio i biomarcatori estratti dall'EEG; le proprietà delle bande di frequenza α, β, θ sono spiegate e contestualizzate nella prospettiva della vigilanza umana. La potenza di α, in particolare, è modulata dal coinvolgimento attentivo ed è l'indice di vigilanza più comunemente utilizzato. Il monitoraggio dell'EEG può migliorare la sicurezza in contesti in sostenere attenzione è critico per evitare errori pericolosi; esempi includono la guida, il controllo del traffico aereo e l'Industria 4.0-5.0. Dopo una panoramica sui metodi di sicurezza legati all'EEG negli ambienti industriali, viene posta forte enfasi sugli effetti del coinvolgimento, dell'attenzione e della vigilanza in un framework di collaborazione uomo-robot. Successivamente, viene introdotto un task sperimentale progettato per indurre sonnolenza e noia: i soggetti monitorano i movimenti di un braccio robotico per un massimo di 23 minuti attraverso 120 trial. I risultati mostrano che la banda α fluttua sistematicamente durante le operazioni del braccio robotico, con picchi di ampiezza pochi secondi prima del completamento di ogni operazione e cali durante i brevi momenti di rilassamento tra i trial. L'EEG così raccolto funge da base per tutti gli esperimenti di DL riportati in questa tesi. Seguono considerazioni sulle sfide attuali nell'acquisizione dei dati EEG, comprese le difficoltà operative in contesti realistici e non controllati, e la necessità di bilanciare la qualità del segnale con il comfort dei soggetti.

Multivariate Time Series Forecasting affronta il problema della previsione degli inviluppi non stazionari delle bande di frequenza EEG. I modelli VAR, nonostante la loro natura lineare, offrono una solida baseline da confrontare con le moderne architetture DL. Tra queste, sono state considerate tre classi: Reti Neurali Ricorrenti (RNN), Reti Convoluzionali e Transformer. Le RNN sono composte da un'unità di memoria che si aggiorna progressivamente man mano che arrivano nuovi dati; evoluzioni delle suddette, come GRU e LSTM, regolano il flusso di dati attraverso la cella di memoria, e compaiono negli esperimenti successivi. Una specifica implementazione RNN in forma Encoder-Decoder costituisce una pietra miliare nella letteratura del DL. Come modello convoluzionale, abbiamo utilizzato le TCN, che sono specificamente adattate al dominio temporale e sfruttano banchi di filtri dilatati esponenzialmente per elaborare serie storiche a lungo termine. Infine, la più recente architettura Transformer PatchTST

offre prestazioni estremamente promettenti grazie a un potente meccanismo di attenzione indipendente fra canali. Le specifiche per tutti questi modelli, assieme ad altri possibili candidati, sono riportate in tutto il capitolo.

Experimental Results in Vigilance Forecasting riporta una valutazione comparativa di architetture statistiche e di Deep Learning, vale a dire VAR, TCN, GRU, Encoder-Decoder GRU (ED-GRU) e PatchTST, per la previsione delle oscillazioni dell'inviluppo α legate alla vigilanza. Per isolare il trend dell'affaticamento mentale, la potenza α continua è stata estratta tramite la Trasformata Wavelet Continua (CWT) e pesantemente decimata a 4 Hz. Sono stati scelti i canali centro-parietali (C3, C4, P3, P4), i quali hanno mostrato in modo evidente un picco spettrale a 0.09 Hz corrispondente al ritmo del task robotico.

I modelli sono stati valutati tramite cross-validation leave-one-subject-out su un orizzonte di previsione di 11 secondi (44 campioni) e hanno mostrato capacità predittive profondamente diverse. La baseline VAR si è degradata rapidamente oltre gli 1,5 secondi, mentre i modelli DL standard come TCN e GRU a singola cella non sono riusciti a catturare le complesse dinamiche future, appiattendosi o ripetendo ingenuamente l'ultimo input. Al contrario, l'architettura ED-GRU ha dimostrato prestazioni robuste, prevedendo accuratamente il comportamento α fino a 5,5 secondi nel futuro (metà di un'operazione robotica) con un R^2 di 0.726, preservando le frequenze spettrali dominanti del segnale.

In definitiva, il modello PatchTST basato su Transformer ha prodotto risultati allo stato dell'arte. Ha superato significativamente tutte le baseline ricorrenti e convoluzionali, raggiungendo un'accuratezza a breve termine quasi perfetta e previsioni a medio termine (5,5 secondi, $R^2 = 0.9$) altamente affidabili. Sebbene le sue prestazioni alla fine decadano per orizzonti temporali estremi (11 secondi), esso convalida chiaramente la superiorità dei meccanismi di attenzione per la previsione a lungo termine degli stati cognitivi non stazionari.

Generative Methods for EEG signals contestualizza l'uso dell'Intelligenza Artificiale Generativa (GenAI) come soluzione alla grave scarsità di dati, allo squilibrio delle classi e ai colli di bottiglia legati alla privacy, tipici della ricerca sulle BCI. Mentre la data augmentation tradizionale si basa su manipolazioni deterministiche e agnostiche rispetto al dominio (ad es. iniezione di rumore o ricombinazione di finestre temporali), le GAN impiegano un'ottimizzazione massimo-minimo, ispirata alla teoria dei giochi, tra un Generatore e un Discriminatore, per apprendere esplicitamente e campionare dal vero manifold dei dati ad alta dimensionalità. Dopo aver tracciato l'evoluzione di queste architetture, dai modelli standard basati sulle immagini alle varianti specifiche per le serie storiche (ad es. RCGAN, TimeGAN), vengono riportate le vulnerabilità algoritmiche critiche delle GAN, prevalentemente la loro suscettibilità all'instabilità in fase di addestramento e al collasso dei modi.

Poiché la fedeltà dei dati temporali sintetici non può essere catturata da una singola metrica universale, la valutazione è strutturata lungo tre assi complementari: i) fedeltà statistica (coerenza ACF e PSD), ii) utilità funzionale (miglioramento delle prestazioni dei modelli a valle) e iii) ispezione qualitativa tramite riduzione della dimensionalità (t-SNE, PCA). Viene introdotta una versione adattata al dominio della Fréchet Inception Distance (FID) utilizzando una rete ED-GRU pre-addestrata per estrarre embedding temporali ad alto livello e calcolare la distanza di Wasserstein-2 tra le distribuzioni gaussiane reali e sintetiche. Il secondo asse riguarda l'utilità funzionale, che viene misurata empiricamente valutando se i dati generati migliorano attivamente le prestazioni di modelli supervisionati a valle, come un previsore. Infine, l'ispezione qualitativa può essere realizzata utilizzando tecniche di riduzione della dimensionalità per valutare visivamente la sovrapposizione dei manifold. Questa ispezione euristica può confermare la diversità distribuzionale e diagnosticare visivamente fallimenti come il collasso dei modi.

L'unione di queste tre prospettive fornisce strumenti di benchmarking quantitativi sufficienti per convalidare i modelli di generazione EEG sviluppati nei capitoli successivi.

Evaluation of Generative Models descrive in dettaglio la progettazione, l'implementazione e la convalida di un framework GAN RNN-CNN per la sintesi di inviluppi tempo-varianti di bande di frequenza EEG (α, β, θ). Convalidata attraverso due distinti paradigmi cognitivi, un dataset pubblico di resting-state (SPIS) e un dataset industriale personalizzato (Vigilance), l'architettura utilizza un Generatore GRU bidirezionale per modellare le dipendenze sequenziali a lungo termine, abbinato a un Discriminatore CNN 1D per valutare l'autenticità temporale. Per affrontare la scarsità di dati BCI, la

robustezza del modello è stata testata su porzioni di dati progressivamente più ristrette (100%, 50% e 25% dei dati di addestramento disponibili).

I dati sintetici sono stati valutati mediante metriche statistiche, misurazioni del miglioramento delle prestazioni e analisi degli embedding; ne conseguono diversi risultati. In primo luogo, la GAN replica con successo range di ampiezza realistici (distanza L2) e preserva complesse correlazioni temporali inter-canale (Cross-Correlation Function). Questa fedeltà strutturale degrada in condizioni di grave scarsità di dati (split al 25%). In secondo luogo, arricchire il set di addestramento con campioni generati dalla GAN produce miglioramenti statisticamente significativi nei task di previsione a valle (nel nostro caso un modello ED-GRU), misurati tramite MAE e SMAPE. L'eccezione riguarda la porzione del 25% dei dati Vigilance, per la quale non è stato osservato alcun miglioramento significativo. Come nota finale, una baseline naive basata sull'interpolazione lineare ha drasticamente ridotto l'accuratezza della previsione, dimostrando che il semplice mescolamento dei segnali è insufficiente per serie storiche biologiche complesse. Infine, lo studio rivela una relazione tra la metrica FID adattata al previsore e l'effettiva utilità dei dati. I dati interpolati linearmente raggiungono punteggi FID vicini allo zero, ma poiché replicano semplicemente campioni già visti, non offrono alcuna nuova diversità, rendendoli inutili per l'addestramento a valle. Al contrario, le generazioni GAN fallite si traducono in grandi outlier FID. Le generazioni GAN riuscite ottengono punteggi FID moderatamente bassi ($\sim$0.5–1.4), a dimostrazione che i dati sintetici ottimali devono trovare un delicato equilibrio: dovrebbero rimanere statisticamente fedeli al manifold fisiologico introducendo al contempo una variazione sufficiente a prevenire l'overfitting del modello.

In definitiva, questo capitolo dimostra che il framework GAN proposto può sintetizzare con successo caratteristiche EEG fisiologicamente valide e funzionalmente utili. Inoltre, stabilisce che un punteggio FID bilanciato e moderatamente basso funge da euristica affidabile per identificare dati sintetici diversificati in grado di migliorare concretamente i modelli predittivi BCI a valle.

Geometry-based Regularization of GANs for EEG Signal Generation stabilisce le basi matematiche e geometriche necessarie per modellare in modo robusto le complesse dinamiche spazio-temporali dei segnali EEG. Poiché i dati neurofisiologici mappati su rappresentazioni strutturate risiedono intrinsecamente su manifold non euclidei, l'applicazione di metriche euclidee standard si traduce spesso in gravi distorsioni geometriche. Per affrontare questo problema, il capitolo formula un approccio riemanniano basato sulla rappresentazione a matrice di covarianza Simmetrica Definita Positiva (SPD) delle finestre EEG.

La logica di questa trasformazione si fonda sul teorema di Wiener-Khinchin: estraendo la Matrice di Covarianza Campionaria (SCM), si calcola effettivamente l'autocorrelazione a lag zero dei segnali. Di conseguenza, gli elementi sulla diagonale catturano l'energia spettrale totale dei singoli canali, mentre gli elementi extra-diagonali codificano la sincronizzazione spaziale a fase zero tra diverse regioni cerebrali.

Geometricamente, l'insieme di tutte le matrici SPD $N \times N$ forma un cono convesso differenziabile. Per eseguire operazioni che rispettino la curvatura di questo spazio, viene introdotta l'Affine-Invariant Riemannian Metric (AIRM) come misura di distanza ottimale. L'AIRM calcola la vera distanza geodetica e possiede la proprietà dell'invarianza affine, rendendola resiliente ai comuni artefatti EEG come lo spostamento dei sensori. Tuttavia, poiché il calcolo dell'AIRM è computazionalmente costoso e penalizza pesantemente le matrici che si avvicinano alla singolarità, metriche alternative computazionalmente efficienti, come la Log-Euclidean Metric (LEM), la Distanza di Cholesky e la Divergenza di Stein, vengono discusse come validi proxy.

Mentre l'SPD standard cattura la connettività spaziale, essa scarta le informazioni di fase e la direzionalità temporale dello spettro di frequenza. Per superare questo limite, il framework viene esteso alla Matrice di Covarianza Aumentata (ACM). Incorporando le cross-correlazioni ritardate nel tempo in una struttura a blocchi di Toeplitz, l'ACM recupera le relazioni di fase temporali perse nella covarianza standard a lag zero. Nello specifico, questa estensione recupera le relazioni di fase lineari all'interno delle stesse bande di frequenza tra i canali, fornendo una rappresentazione nel dominio del tempo più dettagliata e ricostruendo la direzionalità intrinseca della comunicazione neurale.

Tuttavia, l'utilizzo dell'ACM comporta dei compromessi computazionali. Con l'aumentare del numero di ritardi (lags) incorporati, le dimensioni della matrice crescono in modo quadratico. Sebbene il

sottocampionamento degli intervalli di lag e il troncamento del lag massimo (L) rendano l'ACM computazionalmente trattabile, introducono inevitabilmente artefatti di elaborazione del segnale.

Basandosi su questa prospettiva geometrica, il capitolo illustra in dettaglio il contributo metodologico centrale della tesi: una regolarizzazione del generatore basata sulla geometria. Per prevenire il collasso dei modi, la perdita (loss) di entropia incrociata standard del Generatore viene aumentata con un termine di penalità basato sul manifold SPD. Nello specifico, la funzione di perdita mira a massimizzare la distanza riemanniana minima a coppie tra i campioni generati all'interno di un batch. Ciò costringe il Generatore a sintetizzare campioni distinti e ben distribuiti che coprono un'area più ampia dello spazio riemanniano.

Per implementare ciò in modo efficiente, l'AIRM viene calcolata in PyTorch utilizzando un risolutore completamente differenziabile e la decomposizione agli autovalori, aggirando l'instabilità numerica e il costo computazionale dell'inversione esplicita della matrice. Infine, per gestire questo nuovo problema di ottimizzazione multi-obiettivo, che rischia di essere suscettibile all'interferenza del gradiente, vengono considerati i gradienti conflittuali proiettati (Projected Conflicting Gradients, PCGrad) e la normalizzazione dinamica della magnitudo. Questa chirurgia del gradiente (gradient surgery) assicura che le direzioni di aggiornamento conflittuali vengano ortogonalizzate, permettendo al Generatore di migliorare contemporaneamente il realismo del segnale e la diversità distribuzionale senza che un obiettivo prevarichi l'altro.

Infine, vengono introdotti due metodi per valutare la varietà generativa. Il primo prevede l'ispezione visiva tramite UMAP per incorporare le matrici di covarianza ad alta dimensionalità in uno spazio 2D/3D. Un Generatore di successo produce cluster sintetici che si sovrappongono visivamente alla distribuzione dei dati reali, mentre cluster isolati indicano un collasso dei modi. Il secondo misura quantitativamente la dispersione nello spazio riemanniano partizionando le matrici SPD reali utilizzando il clustering K-means; i campioni sintetici vengono poi assegnati a questi centroidi reali, e la divergenza di Jensen-Shannon (JS) tra le distribuzioni di appartenenza ai cluster fornisce una metrica normalizzata della copertura globale.

La GAN regolarizzata geometricamente è stata valutata sul dataset industriale della banda α decimata per valutarne l'efficacia in condizioni di acquisizione realistiche. Questo approccio innovativo è stato confrontato con una baseline GAN (non regolarizzata) attraverso risultati qualitativi, geometrici e funzionali.

L'ispezione visiva delle serie storiche generate evidenzia la suscettibilità della GAN standard al collasso dei modi. Al contrario, il Generatore regolarizzato produce una varietà di oscillazioni realistiche. Questa diversità è stata convalidata quantitativamente all'interno dello spazio della Matrice di Covarianza Aumentata (ACM). Gli embedding UMAP hanno rivelato che, mentre la GAN standard formava cluster isolati e iperspecializzati, i dati sintetici regolarizzati geometricamente si sovrapponevano fedelmente al vero manifold dei dati. Inoltre, il clustering riemanniano non supervisionato K-means ha confermato che l'approccio regolarizzato minimizza drasticamente le divergenze di Kullback-Leibler (KL) e Jensen-Shannon (JS) tra le distribuzioni reali e sintetiche.

L'utilità funzionale dei dati generati è provata riaddestrando il modello di previsione PatchTST sul dataset aumentato. L'accuratezza predittiva a breve termine (11 step) è leggermente a favore del modello aumentato. Su orizzonti a lungo termine (da 33 a 44 step), il modello aumentato ha mostrato valori inferiori di MAE e R^2, oltre a una minore varianza dell'errore, specialmente su canali altrimenti rumorosi (ad es. C3).

Questo framework rappresenta un notevole distacco dai modelli generativi euclidei standard. Forzando esplicitamente il Generatore a massimizzare la distanza AIRM a coppie all'interno dei batch generati, fornisce una soluzione matematica al collasso dei modi considerando la curvatura dei dati fisiologici. Nella pratica, ciò funge da modulo di data augmentation a monte per le pipeline BCI, fornendo un modo per espandere piccoli dataset (ad es. dati di calibrazione).

Attualmente, la metodologia presenta due limitazioni principali. In primo luogo, essendo un approccio auto-supervisionato, il modello interpola in modo eccellente ma non può estrapolare pattern di soggetti fuori distribuzione (out-of-distribution). In secondo luogo, la loss regolarizzata impone un pesante collo di bottiglia computazionale. La ricerca futura dovrebbe considerare sia l'interpolazione inter-soggetto che i miglioramenti computazionali.

La **Discussion** ricapitola i punti principali di questa tesi, dalla previsione degli inviluppi di frequenza

EEG all'augmentation geometrica di piccoli dataset.

È dimostrato che le architetture basate su Transformer possono prevedere con successo i marker di vigilanza della banda α fino a 5,5 secondi in anticipo, fornendo un orizzonte predittivo che supera le latenze di filtraggio e stabilendo una base praticabile per interventi preventivi in ambienti critici per la sicurezza.

Il metodo di augmentation geometricamente informato proposto offre un modo nuovo per espandere i dati biologici, come confermato da una breve rassegna della letteratura sull'ACM. Le valutazioni empiriche hanno dimostrato che questa regolarizzazione geometrica mitiga con successo il collasso dei modi e migliora significativamente la robustezza dei modelli predittivi a valle.

Sebbene il calcolo esatto delle distanze riemanniane imponga un grave limite computazionale, questa ricerca delinea una chiara roadmap per la scalabilità futura. Sfruttando metriche proxy alternative, un campionamento intelligente dello spazio latente o scomposizioni topologiche più sofisticate (in particolare considerando lo spazio del Disco di Siegel), il paradigma di generazione geometrica proposto può essere ottimizzato per supportare la calibrazione BCI paziente-specifica in tempo reale.

Infine, il metodo geometrico proposto per combattere la scarsità di dati di addestramento BCI è inquadrato dalla prospettiva del Geometry-Informed Deep Learning (GIDL), contestualizzandolo all'interno del più ampio campo di ricerca del Deep Learning.

Contents

Chapter 1

Introduction

1.1 The Electroencephalography

1.1.1 Brain-Computer Interfaces and Electroencephalography

A Brain-Computer Interface (BCI) constitutes a direct communication pathway between the human brain and an external device, reading the brain's electrical activity as the input to a system. Functioning as a closed-loop system, BCI reads neural states to control external outputs and provide the user with feedback [1]. Historically, the application of BCI systems has been concentrated in the clinical domain, offering solutions for communication in speech-impaired patients, motor rehabilitation, and neural restoration [2].

Within the broader spectrum of BCI technologies, Passive BCIs are devices used for monitoring rather than for active control. These systems analyze long-term spontaneous brain activity to detect implicit changes in cognitive states, mental workload, or emotional arousal. While their role is well established in monitoring neurological conditions such as epilepsy or sleep disorders [3], their deployment requires acquisition hardware that balances signal quality with usability. In this landscape, Electroencephalography (EEG) emerges as the preferred modality. Offering a cost-effective, non-invasive, and portable solution, EEG enables flexible data acquisition via electrode-covered headsets [4]. Although this setup presents specific signal-processing challenges—discussed in Section 2.3—its versatility enables the deployment of neurotechnology outside strictly controlled laboratory settings.

From Clinical to Industrial Applications

The portability of modern EEG solutions facilitated the migration of BCI technology from clinical wards to human-centered operational contexts. This transition benefited domains where the continuous monitoring of human attention is critical for safety and efficiency. Prominent examples include automotive driving [5], air traffic control [6, 7], and the evolving paradigms of Industry 4.0-5.0 [8].

In particular, the ability to monitor *vigilance* is paramount. Vigilance can be defined as the capacity to sustain attention over extended periods under monotonous stimuli [9], and has been causally linked to workplace accidents [10]. Tracking this mental quality thus provides a proactive layer of safety in potentially dangerous workplaces.

Neurophysiological Markers of Vigilance

To quantify vigilance via EEG, research has extensively focused on the power spectrum of specific frequency bands, with the α rhythm (8-12 Hz) serving as a primary biomarker. α activity is a robust indicator of mental fatigue; as individuals perform repetitive tasks for extended periods, a reduction in vigilance is distinctively marked by a progressive increase in the power of α oscillations [11]. While α activity is central to vigilance monitoring, complementary information is encoded in adjacent frequency

bands. Increased power in the θ band (4-8 Hz), particularly over frontal mesial regions, is frequently associated with the onset of mental fatigue and reduced alertness [12, 13]. Meanwhile, β activity (12-30 Hz) over frontal and central areas typically reflects active concentration and is a marker of rapid cognitive intervention in safety-critical scenarios [14].

AIM: Building a Vigilance Dataset to Study Human Attention

This thesis stems from the objective of analyzing human attention during prolonged tasks and exploring applications to detect and possibly anticipate mental fatigue. Consequently, a monitoring experiment was designed to simulate a real-world industrial environment, thus obtaining a custom vigilance dataset.

In particular, 10 participants were tasked with observing a robotic arm performing repetitive pick-and-place operations for approximately 23 minutes, while their reactions were recorded based on placement accuracy. The experimental protocol successfully induced measurable states of waning vigilance and mental fatigue, a thesis confirmed by subjective questionnaires reporting increased effort and frustration. EEG analysis revealed robust physiological markers of these cognitive states: specifically, a statistically significant ($p < 0.05$) progressive increase in α-band power was observed throughout the session, particularly in centro-parietal regions. Furthermore, short-term sub-trial analysis confirmed dynamic modulation of α consistent with ERS mechanisms. These findings validate the experimental setup as a reliable source of EEG data captured during a situation stressing human vigilance and attention, forming the basis for the subsequent forecasting and generative modeling chapters.

1.2 Deep Learning for Neural Time Series

The slow-changing nature of the power-spectrum biomarkers extracted from our vigilance experiment makes them ideal to predict over long temporal horizons. The rationale for forecasting EEG rhythms, rather than merely detecting them, lies in the possibility of defining proactive safety protocols. Traditional BCI systems identify a decrement in cognitive state—such as the onset of microsleep or a lapse in vigilance—only after the physiological pattern has fully manifested. In safety-critical environments like industrial human-robot collaboration or aviation, this detection latency, compounded by the system's processing time, may render the intervention too late to prevent an accident. By forecasting the trajectories of relevant frequency bands, especially α power, up to seconds in advance, we establish a predictive margin to allow the system to initiate preemptive countermeasures—such as engaging automated failsafes, slowing down machinery, or alerting the operator—before the cognitive failures translate into a behavioral error. Thus, forecasting gives an anticipatory safeguard, compensating for both computational delays and human reaction times.

Unfortunately, the prevailing body of literature regarding EEG analysis shares a common methodological denominator: the reliance on discriminative modeling rather than on signal prediction. In fact, most existing work focuses on post-hoc assessment of collected data or on real-time classification of mental states to detect participants' current vigilance level. This scarcity of predictive works is not unique to neuroscience but permeates the broader domain of biological signal processing, as noted in [15]. A notable exception to this trend is found in the field of Continuous Glucose Monitoring (CGM), where clinical forecasting has achieved significant milestones [16, 17]. The success in CGM is largely attributable to the signal's intrinsic properties: glucose levels fluctuate gradually over minutes or hours, presenting a smooth trajectory that is amenable to long-horizon prediction. In contrast, EEG signals exhibit high-frequency volatility, making long-term prediction exponentially more challenging.

EEG and Deep Learning

To address these challenges, the methodological strategies employed in EEG classification can be adapted for forecasting. Given the erratic and highly non-linear nature of neural signals, Deep Learning (DL) has emerged as the standard computational framework. Unlike classical autoregressive methods (e.g.,

ARIMA), DL models do not require strict stationarity assumptions and can learn complex, non-linear mappings between past and future states.

Historically, two architectures have dominated this landscape. Convolutional Neural Networks (CNNs), originally designed for image processing, excel at extracting local, shift-invariant features from the signal, such as specific waveform morphologies or spectral patterns [18]. Conversely, Recurrent Neural Networks (RNNs) are explicitly designed for sequential data, maintaining an internal memory-like state to capture temporal dependencies. Over time, these foundational blocks have been refined into hybrid architectures and specialized variants [19].

Within the few works that handle EEG forecasting, existing efforts have been confined to extremely short temporal scales. A prominent example is the work by Pankka et al. [20], who adapted WaveNet—a staple convolutional architecture in audio generation—to forecast specific band-passed (θ, α) EEG oscillations. While WaveNet demonstrated superior performance compared to classical Autoregressive (AR) models, the forecasting horizon was limited to just 100 ms. Similarly, studies predicting human motor intentions for safety applications [21] focus on immediate, millisecond-scale reactions. These horizons differ fundamentally from the temporal scales required to monitor cognitive phenomena such as vigilance, which evolve slowly over seconds or minutes.

To bridge the gap between the microdynamics of raw EEG and the macrodynamics of cognitive states, there remains a substantial array of unexplored methodologies. We focus on three distinct architectural paradigms that have shown success both within and beyond the domain of biological time series. First, gated RNNs, born to overcome the standard RNNs' vanishing gradient problem, which limits their ability to learn long-term dependencies. Variants like Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) introduce gating mechanisms that regulate the flow of information, allowing the model to retain relevant history over longer sequences [22]. Temporal Convolutional Networks (TCNs) represent a convolutional approach to sequence modeling instead. By utilizing dilated causal convolutions, they can expand their receptive field exponentially with network depth [23]. This allows them to capture long-range historical context without the sequential processing bottleneck of RNNs. Finally, Transformers represent the current state-of-the-art. These models rely on the self-attention mechanism to weigh the significance of different time steps globally, regardless of their distance [24]. This capability makes them particularly effective at non-linear long-range dependencies that recurrent or convolutional local operations might miss.

AIM: Forecasting EEG Envelope

By investigating and comparing these baseline architectures, this thesis aims to establish a framework for the long-term forecasting of cognitive states, moving beyond the millisecond scale to clinically and industrially relevant horizons.

The custom vigilance dataset was used as a reference for this analysis. To push the forecasting horizon forward, signals were smoothed: the trend of the α envelope is used as the baseline for the forecasting accomplishments in Chap. 4, where the trend of α instantaneous power is accurately predicted up to 5.5 s into the future. In addition, forecasting architectures can be used to extract signal representations, a property subsequently used for comparing groups of different time series in Chap. 6. In this chapter, signals are composed of multi-channel envelopes of α, β, and θ. Here, the second-level trend was not extracted, and signals were instead sampled at 256 Hz.

1.3 Data Augmentation in EEG

While modern DL architectures are theoretically sophisticated enough to model the captured cognitive dynamics, their practical efficacy is predicated on the availability of expansive training datasets. Modern neural networks, particularly deep sequence models and Transformers, are highly parameterized, comprising millions of trainable weights. To effectively map the underlying data distribution, they require a correspondingly large volume of diverse data. Otherwise, during training, they risk overfitting—a common scenario in which models merely adapt their weights to memorize the training data, failing to

generalize to unseen subjects or conditions. The hunger for high-quality data poses a significant bottleneck in the neurophysiological domain. Unlike text or image data, which can be harvested in abundance, acquiring clean, continuous EEG recordings is inherently costly and intensive, further constrained by stringent privacy regulations and the need for specialized hardware.

Given the inherent limitations of empirical data collection, the paradigm of synthetic data augmentation via Generative Artificial Intelligence (GenAI) has been adopted in the BCI domain as a promising solution. Simplistic transformations, such as geometric scaling, noise addition, or time-shifting, often fail to preserve the complex temporal and spatial dynamics of brain activity, risking confounding channel phase information or interpolating meaningless samples. GenAI models can instead learn the underlying probability distribution of physiological signals; architectures such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), originally designed for the image domain, can be adapted to model this high-dimensional manifold [25]. These can sample entirely new, artificial data points that are statistically consistent with real recordings and thus provide the volume and diversity needed by highly parameterized forecasting networks, effectively mitigating the risk of overfitting. Furthermore, because these synthetic samples are artificially constructed and not linked to any real individuals, this approach bypasses the ethical and privacy constraints associated with sharing biological data, offering a scalable and secure pathway to robust BCI development.

Generative Adversarial Networks for Biological Signals

GANs have emerged as one of the most prominent frameworks for biological data augmentation. Introduced by Goodfellow et al. in 2014 [26], a GAN operates under a minimax game-theoretic paradigm, with two distinct neural networks trained simultaneously: a Generator and a Discriminator. The Generator samples and transforms noisy vectors to approximate the true underlying probability distribution of the training data. Conversely, the Discriminator is trained to distinguish between genuine empirical samples and the artificial outputs produced by the Generator. Through iterative adversarial training, the Generator becomes increasingly proficient at synthesizing highly realistic data, while the Discriminator becomes a stricter judge. This ideally drives the system toward a Nash equilibrium in which the synthetic data is indistinguishable from the real data.

In the context of biological and neurophysiological signals, including EEG, Electrocardiography (ECG), and Electromyography (EMG), GANs have been extensively employed to synthesize realistic bio-signals, augmenting scarce clinical datasets and improving the generalization capabilities of deep learning classifiers in downstream tasks. These may range from neurological condition detection to cognitive state monitoring to denoising [27]. We contribute by exploring the efficacy of GANs for generating EEG band envelopes, reporting the work published in [28]. In particular, we perform statistical analysis on the embeddings of a forecaster trained on both real and augmented data to capture differences between fake and real samples. Moreover, differences in forecasting performance are reported to assess the benefits brought by data augmentation.

However, while standard GAN architectures have achieved remarkable success in the image domain, their direct translation to neurophysiological time series poses significant challenges that require targeted architectural interventions.

AIM: Overcoming GAN Limitations with Differential Geometry

Despite the theoretical promise of GANs for augmenting EEG datasets, their practical implementation is characterized by instability during training [29]. Traditional generative models implicitly assume that the data support lies on a flat manifold embedded in Euclidean space, and they use standard metrics, such as Mean Squared Error or binary cross-entropy, to measure the divergence between the real and synthetic distributions. However, neurophysiological signals do not naturally reside in a Euclidean space. Applying standard image-generation techniques to EEG time-series often leads to mode collapse, a state where the Generator completely stops learning - thus completely failing to replicate the non-stationary spatio-temporal correlations between electrodes.

To overcome these architectural limitations, this thesis proposes a novel, geometry-informed generative framework: rather than relying solely on the Generator-side BCE loss, we project EEG signals into the domain of Augmented Covariance Matrices (ACMs) and enforce a spread inside produced batches.

ACM matrices are descriptors that encapsulate both the spatial and temporal correlations of the underlying brain rhythms, and naturally reside on the non-Euclidean Riemannian manifold of Symmetric Positive-Definite (SPD) matrices. In light of this, the core contribution of this work is the integration of Riemannian geometry directly into the generative optimization process. Chapter 7 goes in depth by specifying the mathematics behind this approach; for the moment, it will suffice to say that we implemented a metric between ACM matrices in a differentiable way, and employed it to regularize the Generator loss function. The metric in question is the widely used Affine Invariant Riemannian Metric (AIRM), which allows for the extraction of distances invariant to affine transformations.

This structure-informed approach ensures that the generated synthetic matrices are forced apart in the Riemannian space, avoiding mode collapse. This framework is tested for the generation of decimated *alpha* envelope signals, starting from an extremely small cohort of data points (200 completely different temporal window samples, if overlaps are not considered), the same one used for the forecasting results in Sec.4. Ultimately, this thesis demonstrates that geometrically regularized GANs can synthesize highly realistic and diverse EEG states. This allows overcoming data bottlenecks, as represented in this thesis by the 10-subject vigilance dataset, and improves the performance and robustness of downstream models. The latter point is demonstrated by comparing the accuracy of previously used forecasting models after they are trained on augmented data.

Chapter 2

EEG-based Brain Computer Interfaces

2.1 Brain Computer Interfaces and Electroencephalography

A brain computer interface (BCI) is a system that acts as a direct communication pathway between the brain and an external device, exploiting the brain's electrical activity. BCI forms a closed loop by providing the user with feedback and ensuring a bidirectional information flow by reading and sending information to and from the brain [1]. BCI systems' applications span diverse domains, including the clinical field for communication [30], rehabilitation and restoration [31, 32]. Among these systems, passive BCIs analyze long periods of spontaneous brain activity to detect changes in cognitive states or emotional arousal. Their role in the medical field is well-established and includes monitoring cognitive and neurological conditions, like epilepsy, Parkinson's disease, or dyslexia [33, 34, 35]. In particular, electroencephalography (EEG) offers a cost-effective, non-invasive, and portable implementation [36]. EEG acquisitions are performed using an electrode-covered headset connected to an amplifier; this setup is available in multiple configurations, each presenting its own challenges (see Sec. 2.3), but offers flexibility that allows for its use even outside controlled laboratory settings.

Traditionally employed in the clinical field, as in numerous epilepsy studies [37], EEG has been successfully ported to human-centered contexts such as education or entertainment, with a particular focus for domains where monitoring of human attention is critical, such as in car driving [38, 39], air traffic control and aviation [40], and modern industrial paradigms [41]. A sizable portion of the literature focuses on EEG applications in the construction industry [42], where the impact of visual attention, the effects and implications of mental workload, and the role of emotional states have been recognized as affecting hazard detection and decision-making [43]. In such risky settings, EEG allows for monitoring human vigilance, the ability to sustain attention over a long period of time under monotonous stimulus [36]; vigilance absence has been proven to lead to accidents [10], making it a crucial mental quality to track in potentially dangerous workplaces.

In EEG-based vigilance detection, the power spectrum of different frequency bands and their ratios have been used extensively, with major focus put on α (8-12 Hz). Indeed, the α power spectrum can decrease or increase in response to a stimulus, with increased activity called event-related synchronization (ERS) and decreased activity called event-related desynchronization (ERD) [44]. ERD in the α band is associated with visual stimulation, and is related to activation of cortical areas related to information processing, selective attention, and motor preparation. On the other end, ERS is associated with inhibition of brain activity and may indicate a resting or idling cortical area in which no information is currently processed [45]. The α band activity is also influenced by increased mental fatigue [46]. Mental fatigue refers to a state that occurs when individuals perform repetitive and monotonous tasks for an extended period, which is characterized by a loss of vigilance. Growing mental fatigue, and thus vigilance

reduction, is characterised by an increase in the power of α oscillations [36].

In addition to α activity, other frequency bands carry complementary information: increased θ (4-8 Hz) power, over frontal mesial regions, is often associated with mental fatigue and reduced alertness [12], while β (12 - 30 Hz) activity, often observed over frontal and central areas, typically reflects active concentration and rapid interventions in safety critical environments [14].

2.2 The Industrial Frontier

Novel experiments bringing EEG to the Industrial field show promising results in terms of both monitoring and improving human factors performance. After an introduction on recent strategies, we define the data acquisition paradigm upon which the majority of this thesis's experiments were based.

2.2.1 Human safety in Industry 5.0

Industry 5.0 [47] aims to create a symbiotic relationship between humans and robots that boosts productivity while ensuring human well-being and safety. In this advanced industrial paradigm, the integration of collaborative robots, or cobots, necessitates a sophisticated approach to monitor and manage human factors that could lead to safety hazards [48], especially since a human-robot collaboration (HRC) cell often lacks physical barriers. This aspect greatly increases work flexibility but stresses the need for further safety measures [49]. A cell is equipped with various sensors that monitor and adjust human behavior, nudging toward ergonomics postures to avoid musculoskeletal disorders [50], and regulate the robot movements according to torque, proximity, or touch to detect obstacles and react properly [51]. An additional layer of safety is accomplished by preventing dangerous collisions through human detection and pose estimation. Human position is obtained either through wearables, such as inertial measurement unit (IMU) sensors, or by cameras, in a scenario which may or may not include the presence of markers [52]. The human pose is often extracted from videos using convolutional neural networks (CNN) [53] and compared with the robot trajectory to modify its behavior [54]. These safety methods, however, do not account for subjective factors such as stress, drowsiness, or excessive workload; in fact, a particular factor of risk is the decreased awareness caused by performing repetitive and monotonous tasks for long hours [55]. Thus, human safety in Industry 5.0 concerns not only physical but also cognitive and psychological factors that necessitate continuous assessment.

Real-time monitoring strategies for improving this safety aspect in the human-cobot framework have been extensively studied in recent years, with applications that aim to reduce user stress [56] or mental workload [57].

2.2.2 EEG Applications in the Industrial Field

In the human-cobot framework, understanding human psycho-physiological behavior during task execution is crucial to guarantee operators' mental well-being and avoid dangerous workplace accidents. Monitoring adversarial mental effects such as stress, fatigue, or drowsiness made EEG-based approaches flourish in the Industry 5.0 scenarios, allowing researchers to investigate mental fatigue onset, predict cognitive overload, and assess the impact of different assembly times on operator mental wellbeing [58]. Furthermore, recent works highlight the importance and feasibility of monitoring mental states in HRC. Workers' EEG can be implemented in human-robot cooperation either as a control signal, to regulate the robot kinetic functions for safety purposes, or to detect anomalies in the assembly procedures, when not used to directly control a robotic appendage [21]. Other studies focus directly on the human factor: for instance, how trust in robots, physical proximity to them or varying levels of automation affect the operator mental health. Indeed, EEG can sensitively track the level of perceived stress [59], which, if left unchecked, may have adverse effects on assembly workers' motivation and decisions [60]. This assessment frequently involves the EEG power spectrum, in the α (8-12 Hz) and β (12-30 Hz) frequency bands, and can become even more precise if other biometrics, such as electrocardiogram or facial expressions, are considered [61]. In fact, EEG features have been confirmed as precise biomarkers of adverse mental

effects, such as excessive mental workload or lack of vigilance, in various sectors, ranging from plant control simulation [62] to the construction industry [63]. The ratio of α and β power bands is also used as a biomarker for mental workload, a measure that excess denotes frustration in operating along the robotic arm, and lack of alertness about workers detaching from the task at hand, a state known as the "out-of-the-loop" phenomenon [57]. In this context, the importance of human engagement is a factor that is becoming increasingly considered in the field of HRC. Engagement is strictly related to vigilance: the EEG engagement index is expressed as the ratio of β and $\alpha + \theta$, which changes proportionally with an operator's attention state. Inside the industrial field, the relevance of modulating robot movements using BCI in robot-assisted tasks has been proven: changes in robot behavior can temporarily restore user engagement when distracted [64]. Similarly, [65] demonstrated how extracting spectral and entropy features from θ, α, β frequency bands could drive a cognitive-fatigue detection system. The system redistributes tasks to prevent overwork using an LSTM-based model, which correctly detects fatigued conditions combining EEG information and the participants' video stream, and a task-reallocation function based on reinforcement learning and a genetic algorithm. The authors conclude that this approach is a way to avoid overworking conditions. Engagement is positively fostered when synchronizing with a cobot to avoid idle times that can generate detachment and boredom: this is confirmed in [66], although only for extremely short working shifts (3 minutes) - introducing a way to modulate task intensity during longer working hours could help to avoid excessive mental burden on the operator. The importance of monitoring engagement and the negative effects of uncontrolled task automation are further emphasized in [67], where the authors develop an interface to monitor levels of involvement in the current assembly task. As a final example, heart rate variability and galvanic skin response have been combined with EEG to obtain a reliable biomarker for human trust in robotic appendages; human trust is indispensable to guarantee proper collaboration and is characterized by power fluctuations of α and β in the frontal brain cortex [68]. These works mark how crucial it is to consider the temporal development of EEG to better improve the human experience along cobots, considering workers' engagement and confidence in the task at hand as essential to improving working conditions.

2.2.3 Experimental setup and paradigm

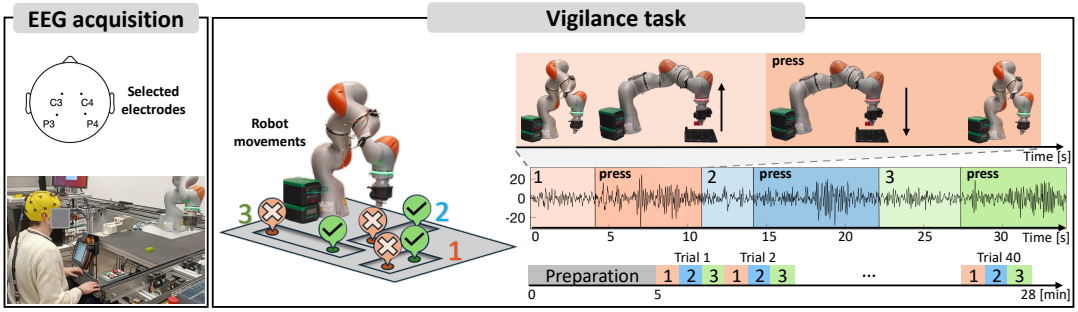


Figure 2.1: Industrial laboratory setup in which EEG signals were acquired and a schematic representation of the vigilance protocol. The subject observed the KUKA robot while it moved a small block in 3 different positions 40 times (i.e., from position 1 to 2, from 2 to 3, from 3 to 1, etc.). The block could be placed in the correct (tick in green) or the wrong sub-position (cross in orange). The timing schedule of the protocol is reported in the right part and an example of the acquired EEG signal is represented for a trial. The 'press' marker roughly denotes the temporal instant the subject pressed the keyboard.

This thesis is grounded in human vigilance data collected during a custom task, hereinafter referred to. The process mimics an industrial cobot workstation and sees participants observing and recording robotic movements.

A total of 10 volunteers (average age 27.3 $\pm$ 3.6 years, all males) participated in the data acquisition. Volunteers were either university students or personnel with no previous experience in the task. They

were asked to give informed consent before starting the experiment. Data were collected according to the protocol approved by the local ethics committee of the University of Verona (Comitato di approvazione della ricerca sulla persona [CARP]: protocol n. 2.R1/2022). The study design and data collection occurred in an industrial research facility laboratory, the ICE Laboratory of Verona (`https://www.icelab.di.univr.it/`). This laboratory simulates a real-world industry setting with several advanced technological machines and systems which are interconnected with a logistics system consisting of a transport line. Once the task was completed, the participants were asked to respond to a subset of questions adapted from the NASA task-load index [69]. Questions pertained to the perceived pace of the task (labeled as "Temporal Demand"), the amount of effort the task required (labeled as "Effort"), and their level of frustration with the task (labeled as "Frustration"), all rated on a scale from 0 to 20 - low to high. Additionally, participants were asked whether they felt their attention waning during the task, and binary answers were recorded.

Recordings were made using a portable amplifier (BeMicro, EBNeuro S.p.A., Firenze, Italy) and a cap providing 21 Ag/AgCl electrodes positioned according to the standard 10/20 system. Reference was placed anteriorly to Fz, ground was posterior to Fz. Conductive gel was applied to maintain an impedance below 20kΩ. A custom-made layer of insulating material and shielding material was applied over the amplifier and the connecting cable. The cap was protected by wearing a similarly made headset. EEG data were acquired at a sampling frequency of 256 Hz. Participants sat in front of a workbench and were instructed to limit body movement. They observed the movement of a robotic arm (KUKA LBR iiwa 14 R820), which is composed of seven joints, a flange, and a GIMATIC electromechanical gripper at the end, used to grasp a small object. Fig. 2.1 shows the whole setup. EEG signals were recorded during the proposed vigilance task. This protocol consisted of 40 trials and lasted a total of 23 minutes. Each trial lasted between 33 and 35 s and comprised a sequence of 3 pick-and-place operations of the KUKA robot as depicted in Fig. 2.1. Each individual robot operation lasted approximately 11 s. The movement began with the robot picking a small block of 2x4 cm from position 1 and moving it to position 2. Then, the robot moved the object from position 2 to 3. Between each movement, the robot returned to its resting position. At each iteration, the block could be placed in two different sub-positions, marked as a tick or a cross in Fig. 2.1. The different distances between positions was the main factor accounting for the changing trial durations; specifically, movements involving position 3 (see Fig. 2.1) took about half a second longer than those involving other positions. On a smaller scale, the same principle applied to the right and wrong sub-positions. Additionally, each operation was transmitted individually to the robotic arm through the laboratory's internal network, with slightly varying reception times. The probability of success for the correct sub-position was set to 80%, while it was 20% for the wrong one. Participants assessed KUKA's operations by pressing the space bar button (for the "tick" sub-position) or the Enter button (for the "cross" sub-position). Participants reacted immediately upon evaluating the position with a minimal index finger press. A video was recorded to capture any failures, complications, or movements.

2.2.4 Data preprocessing

EEG data were preprocessed using Matlab R2022a (MathWorks, Natick, MA, USA) and EEGLAB (`https://sccn.ucsd.edu/eeglab/index.php`). A 3rd-order Butterworth filter removed frequencies outside the 4-30 Hz range. Data were re-referenced to common average reference (CAR) and baseline corrected. Then, data were filtered using the CleanRawData (CRD) EEGLAB plugin [70], which, through the automated subspace removal (ASR) algorithm, salvages background EEG activity submerged by high-amplitude noise by comparing the EEG to artifact-free reference data. Finally, independent component analysis (ICA) was used to remove residual artifacts (i.e., eye movements, cardiac activity, and scalp muscle contraction). The sequence of CRD and ICA applications follows the CRD authors' recommendations found at `https://github.com/sccn/clean_rawdata/wiki`. The pre-processed data was used as a starting point for all the experiments reported in this thesis.

2.2.5 Experiment Vigilance Correlates

To assess whether the experiment induced the expected EEG frequency behavior, data were analyzed in Matlab (R2022a) using EEGLAB (`https://sccn.ucsd.edu/eeglab/index.php`), and results were first reported in [71]. The Continuous Wavelet Transform (CWT) was employed to extract patterns from EEG signals [72], utilizing the Morlet wavelet, a widely used method in EEG analysis. We estimated the power spectrum over time for both rest and task conditions from all EEG recordings on the θ and α frequency ranges.

The task is composed of 120 sub-trials. Data analysis was performed in two different ways: on each sub-trial (i.e., 11 s) and on the complete EEG signal (i.e., 23 min). In the first case, the analysis focused on time-variant power spectrum behavior along different phases of a sub-trial [Fig. 2.2b]. The power signal trend following RT was obtained using a linear regression model. In the second case, the time-variant α power spectrum of each channel was considered in order to evaluate changes over the entire task recording. The obtained power spectrum was smoothed using an exponentially weighted moving average (EWMA) filter to reduce variability within the sub-trials. A first-order linear regression model was then used to evaluate increments or decrements over time.

Sub-trial analysis

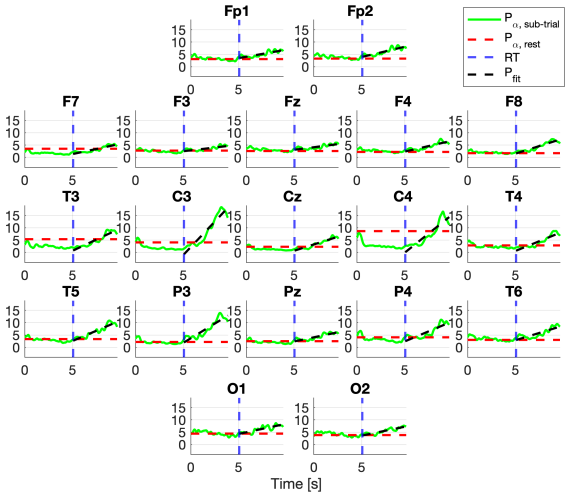
The sub-trial analysis for a representative subject (no. 1) is presented in Fig. 2.2b and in Fig. 2.2a. Fig. 2.2b illustrates the average of the power spectrum over sub-trial in the α frequency band for the channel P3, as well as the power calculated during a resting EEG and the average RT. As shown, the average power decreases when the robot grasps the small object, for at least 3 s, and then increases after the RT. The power for each EEG channel is reported in Fig. 2.2a. The fitted line, obtained with the linear regression model starting from RT, is also represented. A positive trend is confirmed, consistent with the ERS behaviour triggered by the subject's relaxation. It can be observed that the slope is higher for channels C3, C4, P3, and P4. This trend is confirmed for the other subjects. In fact, slopes of the fitted line referred to each sub-trial were firstly averaged over each sub-trial and then over subjects. These values are shown in Table 2.1. These results confirm that α power spectrum increments affect the centro-parietal brain areas more. In fact, the highest values were equal to 1.14 (C3), 0.98 (C4), 0.82 (P4), 0.78 (P3), 0.71 (T5), and 0.62 (Pz). Grouping the results by brain regions, and by averaging them, central and parietal areas are the ones experiencing the highest slope, being at 0.93 and 0.74.

Table 2.1: *Sub-trial analysis*: The slopes of sub-trials computed following RT, averaged over subjects. The higher six values are reported. Despite high SD values, no single subject reports negative slopes. P-value of slope parameter is significant for all subjects ($p < 0.05$).

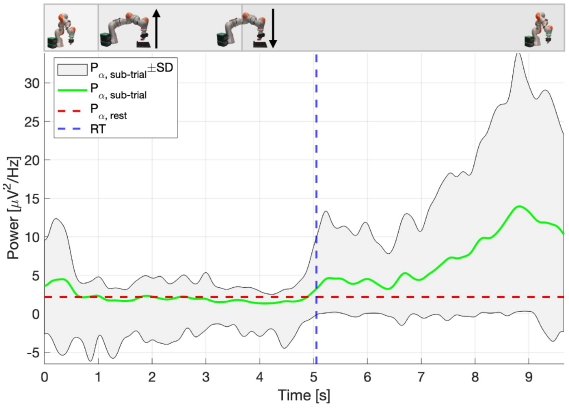
Ch	*Mean*	*SD*	R^2
C3	1.25	±1.46	0.63
P3	1.00	±0.77	0.66
T5	0.89	±0.59	0.61
C4	1.07	±1.05	0.62
P4	0.68	±1.09	0.60
PZ	0.60	±0.76	0.59

Overall EEG signal analysis

The results of the experiment indicate that mental fatigue increased in all subjects. This is first reported in questionnaires, the results of which are shown in Tab. 2.2. High value of either mental fatigue or frustration were reported; outside the questionnaires, all subjects confirmed they felt waning attention



(a) *Sub-trial analysis*: Time-variant average power spectrum in the α frequency band across sub-trials of all channels for subject 5 (average $P_{\alpha,sub-trial}$ [green], average $P_{\alpha,rest}$ [red] and the average RT [blue]). Data fitted starting from RT (black).



(b) *Sub-trial analysis*: Time-variant average power spectrum in the α frequency band across sub-trials of a parietal channel ($P3$) for subject 5 (average $P_{\alpha,sub-trial} \pm SD$ [light gray region], average $P_{\alpha,sub-trial}$ [green], average $P_{\alpha,rest}$ [red] and the average RT [blue]).

when asked directly. An example is illustrated in Fig. 2.3 for the α frequency band. For each channel, the linear regression shows a growing trend, especially in centro-parietal areas. The 95% confidence interval is also reported. A significant relationship between time and the increase of $P_{\alpha,task}$ (F-statistic: $p < 0.05$) is confirmed.

Additionally, Fig. 2.4 shows the slopes of the fitted line over $P_{\theta,task}$ and $P_{\alpha,task}$, both θ and α frequency bands. For all the brain areas of interest, we confirm a positive trend in the α band, with the exception of subject 10, channel P4. θ perform more erratically, possibly denoting the presence of noise even after the filtering pipeline. Subject 5, with the highest α slope, is also the one with the highest combination of effort and frustration, as shown in Table 2.2, indicating that the task was particularly tiresome for them. The opposite verifies for subjects 1, and 4; they both reported the minimum degree of frustration possible, and their computed slopes are the lowest two. Still, the easiness of the task leave doesn't leave room for a more in-depth interpretation - even when exceedingly tired, the participants were able to perform the assignment correctly.

Table 2.2: Subjects' mistakes and self-evaluations.

Subject	Error Total	Temporal Demand	Effort	Frustration
1	no errors	2	15	1
2	2	3	3	19
3	no errors	6	1	14
4	1	1	4	1
5	2	3	15	14
6	no errors	4	2	16
7	1	2	1	17
8	1	4	13	2
9	no errors	1	1	20
10	no errors	3	14	15

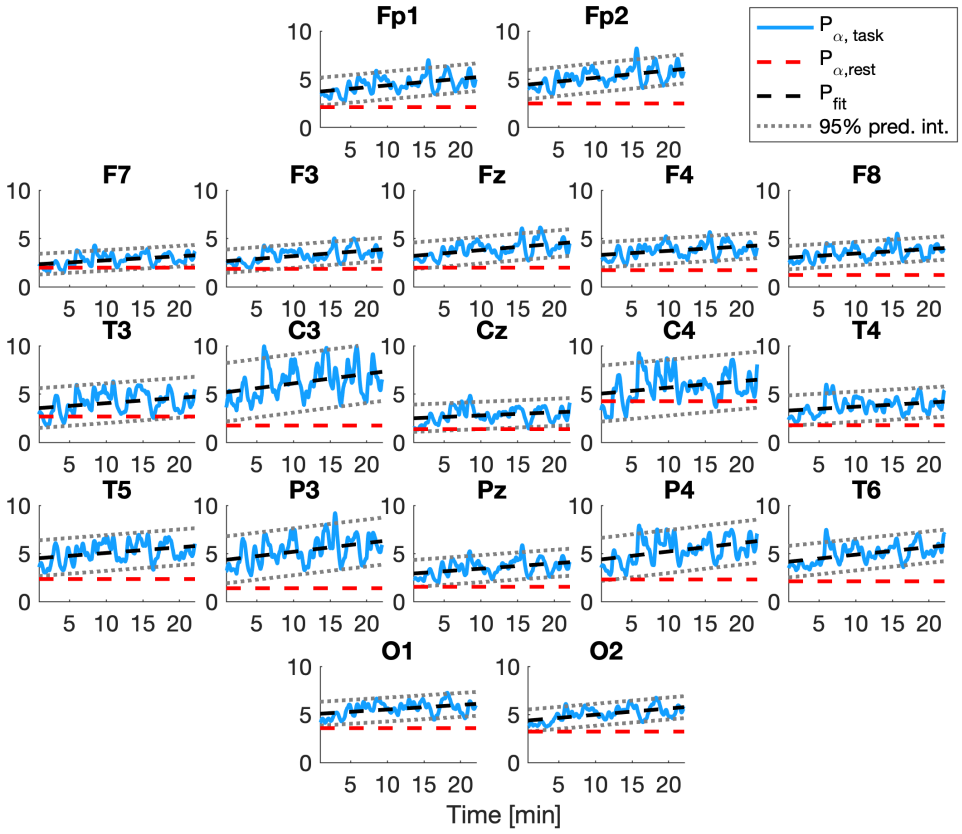


Figure 2.3: *Overall EEG signal analysis*: Power spectrum in the α frequency bands over the entire acquisition for subject 5 (smoothed $P_{\alpha,task}$ [light blue], average $P_{\alpha,rest}$ [red], fitted line [black] and the 95% prediction interval [gray]).

2.3 Current Challenges in EEG-based BCIs

EEG acquisitions may be affected by multiple issues that can impact data quality and overall availability. First of all, EEG data is captured in the order of micro-volt (μV). This makes it susceptible to both internal and external artefacts. On the one hand, the subject's muscular activity is detected by the setup and manifests itself as brief burst disturbances, especially in cases involving eye movement and jaw clenching. These two situations illustrate well-established waveforms that have been extensively studied in the literature in order to filter them out. It is also important to consider that EEG captures all cortical activity; this means that singular factors, such as human fatigue or movement intent, aren't trivial to isolate, even with knowledge of their spatial location and frequency footprints. When bringing EEG out of controlled settings, precautions must be taken to shield the acquisition gear from electromagnetic and luminous stimuli and to avoid unnecessary movements. Furthermore, if users are not supposed to stay still, having a wireless amplifier becomes a necessity: cabled connections would perturb the acquisition if the user were to move around.

Another major point involves setup times and the user's comfort. Higher-quality data can be recorded if conductive gel is applied to each individual electrode, thereby improving signal reception from the human scalp. To ensure an even cleaner signal, exfoliating the skin is recommended. However, such a lengthy procedure can not be afforded in practical scenarios, especially if involving partially impaired individuals, such as during rehabilitative tasks. Moreover, the need to reapply the gel over hours-long acquisitions is a further hindrance. An alternative exists in the form of saline-based solutions, in which

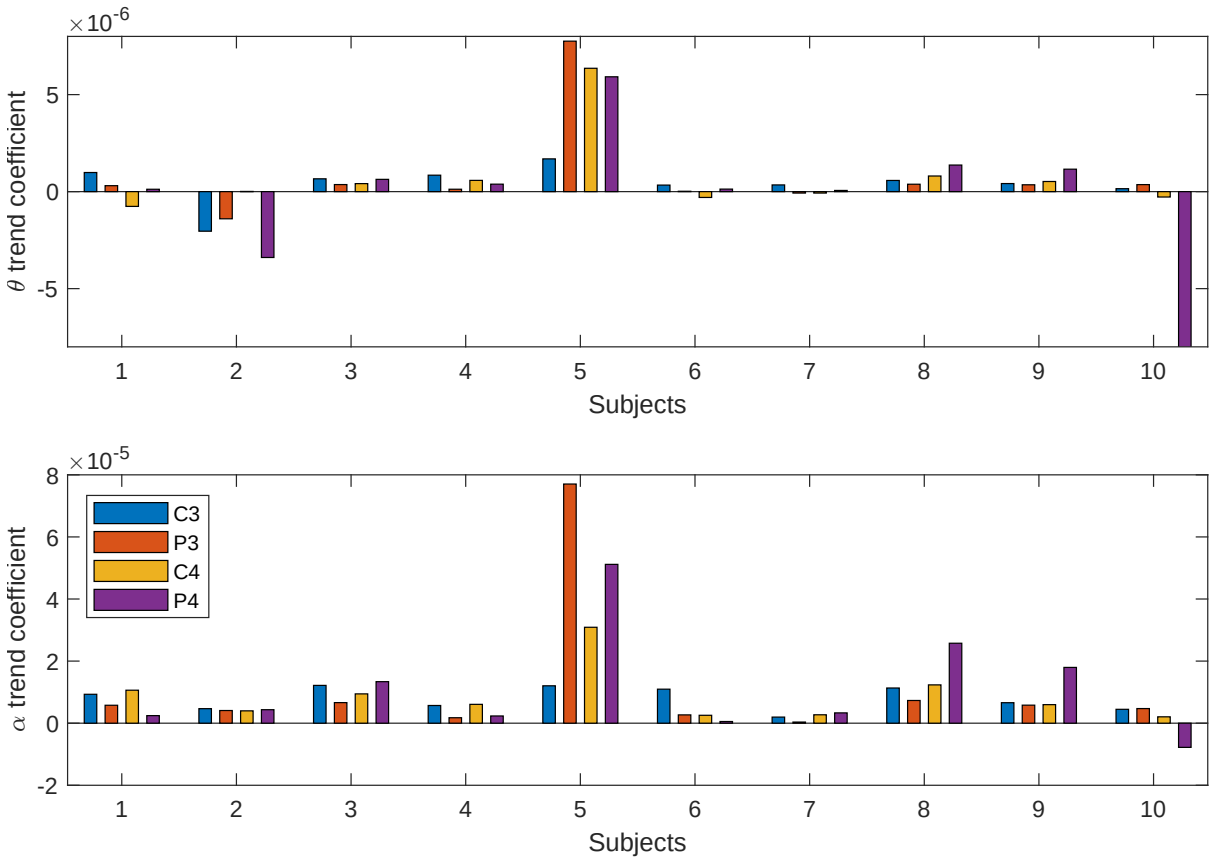


Figure 2.4: *Overall EEG signal analysis*: Slopes of the fitted line over time-variant power, computed for each subject, for each of the four centro-parietal channels of interest. θ is reported above, α below.

the electrodes can be collectively bathed, or solution-free electrodes, which often use smoothed prongs to better reach the scalp among the mass of hair. Both approaches, of course, come at the cost of a worse signal-to-noise ratio.

The hardships of collecting data are evident in the overall lack of public datasets and in small-scale experiments that rarely exceed tens of subjects. Taking this into consideration, Sec. 5.1 offers data-augmenting solutions to circumvent this last issue.

Chapter 3

Multivariate Time Series Forecasting

3.1 Problem formulation

Multivariate time series analysis is the study of multiple time-dependent variables observed simultaneously over a specified period. A multivariate time series can be formally represented as a sequence of random vectors, $\{\mathbf{X}_t\}$, where each vector $\mathbf{X}_t = [X_{1t}, X_{2t}, \ldots, X_{Nt}]^T$ contains the observations of N different variables at time t. This general formulation applies well to EEG data, in which multiple channels are recorded simultaneously.

Along this thesis, high importance is given to the forecasting of future variables; formally, the objective of forecasting is to determine an optimal prediction, denoted $\hat{\mathbf{X}}_{T+h}$, for the value of the series at a future time $T+h$, where $h \geq 1$ is the *forecast horizon*. The forecast horizon defines how far into the future the prediction is made; it can be a one-step-ahead forecast (h=1) or a multi-step-ahead forecast (h>1). An optimal forecast is one that minimizes the expected value of a specified loss function, which quantifies the discrepancy between the actual future value $\mathbf{X}_{T+h}$ and the forecast $\hat{\mathbf{X}}_{T+h}$.

The choice of loss function implies different desired outcomes and sensitivities. In subsequent experiments, model performance is primarily evaluated using three standard metrics: Mean Squared Error (MSE), Mean Absolute Error (MAE), and Symmetric Mean Absolute Percentage Error (sMAPE). MSE, defined as $\frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{X}_i)^2$, optimizes for the conditional expectation of the future value [73]. It offers a smooth, convex optimization landscape but is highly sensitive to outliers due to its quadratic nature. Conversely, MAE, defined as $\frac{1}{n}\sum_{i=1}^{n}|X_i - \hat{X}_i|$, targets the conditional median [73], providing greater robustness against extreme data deviations. Finally, sMAPE, formulated here as $\frac{100}{n}\sum_{i=1}^{n}\frac{|X_i - \hat{X}_i|}{|X_i| + |\hat{X}_i|}$, provides a bounded (0–100%), scale-independent measure. This makes it particularly advantageous for directly comparing forecast accuracy across multiple EEG channels or series with varying amplitudes.

3.1.1 Model choiche and evaluation

As previously established, effective forecasting requires applying a predictive model to historical data; unfortunately, no catch-all solution exists. While classical autoregressive methods provide optimal representations for wide-sense stationary time series exhibiting linear correlations, these assumptions are rarely met in real-world physiological signals. EEG data, in particular, is characterized by inherent non-stationarity, noise levels, and non-linear dynamics that persist even after filtering and dimensionality reduction. Consequently, traditional statistical approaches often fail to capture the underlying generative processes of brain activity.

The advent of deep temporal neural networks has provided a robust alternative, offering multiple architectures capable of capturing these sequential dependencies. Yet, this abundance presents its own challenge: with available models ranging from simple recurrent units to complex transformer-based systems, finding the most adequate architecture necessitates both a deep understanding of the data's topological structure and extensive empirical validation.

Furthermore, the data-driven nature of DL theory may fall short in the face of the high intra-subject variability of EEG: different people exhibit waveforms with distinct spectral distributions, and a perfectly working model may not generalize well when presented with entirely new data from subjects not in the training set.

In this thesis, three distinct classes of deep learning architectures were rigorously evaluated: Recurrent Neural Networks (RNNs) for their sequential memory capabilities, Convolutional Neural Networks (CNNs) for their efficiency in local feature extraction, and attention-based Transformers for their ability to capture long-range dependencies. A theoretical summary of these architectures is detailed in the upcoming section, while a comparative analysis of their efficacy in forecasting α-band activity is presented in the Forecasting Results chapter 4.

3.2 Vector Autoregressive Models

The **Vector Autoregressive (VAR)** model is a stochastic process model used for analyzing the joint dynamics of multivariate time series. Modelling linear dependencies with autoregressive models has been proved effective even in complex contexts such as the study of resting-state brain activity [74]. Used in the neurological field to perform source connectivity analysis [75], VAR processes forecast multidimensional vectors of future values with a dimensionality of d. As a direct generalization of the univariate autoregressive (AR) model, a VAR model formulates future observations as a linear combination of multiple prior observations, effectively encapsulating the (linear) dependencies within the data.

3.2.1 Formal Representation

A VAR model of order p, denoted VAR(p), for a set of N time series variables is represented by a system of N linear equations. Let $\mathbf{X}_t = [X_{1t}, X_{2t}, \ldots, X_{Nt}]^T$ be the $N \times 1$ vector of variables at time t. The VAR(p) process is formally defined as:

$$\mathbf{X}_t = A_1\mathbf{X}_{t-1} + A_2\mathbf{X}_{t-2} + \cdots + A_p\mathbf{X}_{t-p} + \epsilon_t$$

or more compactly:

$$\mathbf{X}_t = \sum_{i=1}^{p} A_i\mathbf{X}_{t-i} + \epsilon_t$$

where:

- A_i are $N \times N$ coefficient matrices for each lag $i = 1, \ldots, p$. These matrices capture the linear interdependencies within the system.
- ϵ_t is an $N \times 1$ vector of white noise error terms, called innovations or shocks, that satisfies:
  1. $\mathbb{E}[\epsilon_t] = \mathbf{0}$ for all t
  2. $\mathbb{E}[\epsilon_t\epsilon_s^T] = \mathbf{0}$ for all $t \neq s$
  3. $\mathbb{E}[\epsilon_t\epsilon_t^T] = \Sigma_\epsilon$, where Σ_ϵ is the $N \times N$ positive definite covariance matrix of the error terms

3.2.2 Lag Order Selection

The choice of the lag order p is a critical step in VAR modeling. An insufficient lag order (p too small) may lead to model misspecification and autocorrelated residuals, while an excessive lag order (p too large) results in the estimation of unnecessary parameters, leading to a loss of degrees of freedom and model overspecialization. Lag selection is typically performed by minimizing a model selection criterion; two of the most used are the Akaike Information Criterion (AIC) [76] and the Bayesian Information Criterion (BIC) [77].

3.2.3 Limitations

Despite its utility, the VAR model has limitations. Its primary drawback is the curse of dimensionality. The number of parameters to be estimated is $N(1+pN)$, which grows quadratically with the number of variables N, leading to overfitting for even moderately large systems. Moreover, applying VAR to EEG data violates its fundamental assumption of wide-sense stationarity. This assumption posits that the first two orders statistical properties of the signals— their mean, variance, and autocorrelation — are constant over time. However, EEG recordings are inherently non-stationary, as they reflect the brain's dynamic transitions between different cognitive and physiological states. Events like task-switching, the onset of drowsiness, or the presence of event-related potentials and artifacts all violate the stationarity assumption by altering the underlying statistical structure of the data. A common workaround in practice is to apply VAR analysis to very short, quasi-stationary epochs, where the brain state can be reasonably assumed to be stable.

3.3 Convolutional Neural Networks

A Convolutional Neural Network (CNN) is composed of a cascade of convolutional layers, characterized by their kernel matrix. Weights of the kernel matrix are learned during the training process thanks to the backpropagation algorithm. For the 2D case, each convolutional layer takes as input a B x N_i x M_i x C_i tensor, with B being the batch dimension, N and M the current image size, and C being the channel dimension, usually starting from 3, as the RGB color channels of a picture. The layer will output a B x N_o x M_o x C_o tensor: channels can shrink or increase in this passage, while the image dimension will either remain the same or contract if using a dilated filter.

Convolutional layers are usually interspersed with activation layers, pooling layers, and normalization layers. Activation layers introduce non-linearities in the network, pooling layers are required for feature aggregation, and normalization layers stabilize the training process. The initial layers of the architecture extract localized features that become increasingly general as the network depth increases. The convolutional operator is by itself covariant with respect to data translations; the entire network can be considered invariant to translations due to the pooling operations [78].

CNNs have shown remarkable performances in the field of adversarial mental state classification, with architectures obtaining SoA results in detecting vigilance for the sectors of driving [18] and construction [79]. Domain-specific architectures are seldom limited to a convolutional approach, and often involve the combination with recurrent neural networks or attentional mechanisms [80, 81]. In contrast, the TCN, first introduced in 2018 [82], is a standalone variation of a CNN designed for sequential data. TCN can be viewed as a 1D fully convolutional network that employs only causal convolutions, meaning there is no information "leakage" from the future to the past. Additionally, the architecture can take a sequence of any length and map it to an output sequence of the same length. For the network to access information from the far past, a combination of profound network depth (augmented with residual layers) and dilated convolutions is employed. Specifically, the dilation factor d of the kernel increases exponentially as the network becomes deeper. TCN has been shown to obtain SOA performance on both classification and forecasting problems [83].

3.4 Recurrent Neural Networks

RNNs are known for their effectiveness in time-series analysis. Using RNNs has produced noteworthy results in EEG studies [84]. The RNN architecture, containing an input layer (I_n), recurrent hidden layers (h), and an output layer (O_n), has been extensively explored as a DL methodology in the domain of time series analysis and forecasting, as its internal structure theoretically allows the retention of previously seen information. Practical challenges have traditionally hindered the application of RNNs to signals with long-term dependencies; "vanishing error" occurs when the back-propagated error signal in an RNN converges to zero during training, while the opposite can also happen, causing the error value

to "blow up". This is an effect caused by the exponential dependency of the error signal on the size of the weights [85]. Gated architectures such as GRU [86] and the LSTM [85] mitigate the vanishing gradient problem and are recognized for their ability to capture significant information over long data sequences [87], and have been used extensively in contexts ranging from natural language processing [88] to motor imagery classification [89]. Gated architectures address the vanishing error problem by expanding the atomic memory block with gating units, whose purpose is to control the flow of input and output signals through the memory block, and to bridge the error signal without disruptions over long temporal intervals. GRUs, distinguished by their lack of a dedicated memory cell, instead incorporate two gating units: the "reset" gate and the "update" gate, which regulate error propagation through the unit. GRUs have been shown to perform comparably to LSTM models while requiring fewer parameters to tune [90], and for this reason, were chosen as the recurrent cell used in subsequent analysis. The architectural configuration of GRUs is reported in Fig. 3.1. The update gate plays a crucial role in weighing a proposed result, known as the candidate activation, against the previous unit output. It is possible to stack multiple GRU cells so that they sequentially process the outputs derived from the previous one [91].

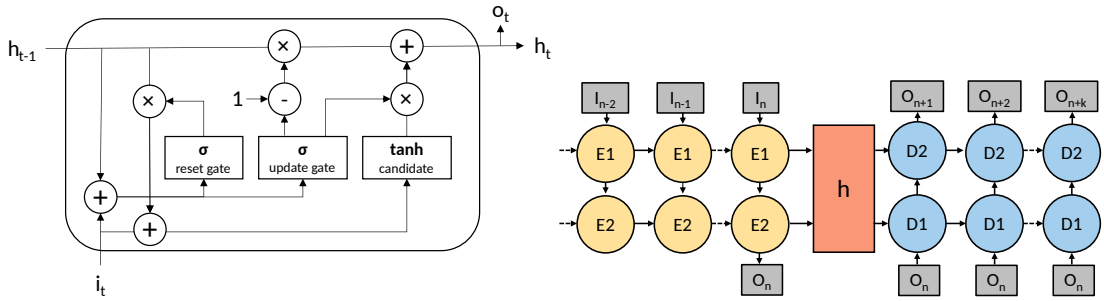


Figure 3.1: Left. Architectural diagram of a GRU. Right. Architectural diagram of the proposed ED architecture.

Encoder-Decoder GRU architecture (ED-GRU)

The ED architecture (Fig. 3.1) consists of two main components: the encoder and the decoder, both of which are composed of a stacked sequence of RNN cells [88]. The role of the encoder is to compress the information read from the input into a hidden state. This hidden state is used to initialize the decoder's one, which then, after receiving a sequence of inputs, generates the forecast series. The architectural configuration of the ED is reported in Fig. 3.1.

3.5 Transformers

The Transformer architecture, initially developed for natural language processing, represents a paradigm shift in deep learning sequence modeling. Unlike Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks that process data sequentially, the Transformer relies entirely on self-attention mechanisms to weigh the importance of all elements in an input sequence simultaneously when encoding a single element. This parallelization capability allows the model to capture long-range dependencies effectively, mitigating the gradient degradation issues often observed in recurrent architectures.

3.5.1 The Self-Attention Mechanism

The core of the Transformer is the self-attention mechanism, which determines how strongly a specific time step (or token) should correlate with other time steps in the sequence. This process operates with three main vectors derived from the linear projection of input embeddings: queries (Q), keys (K), and

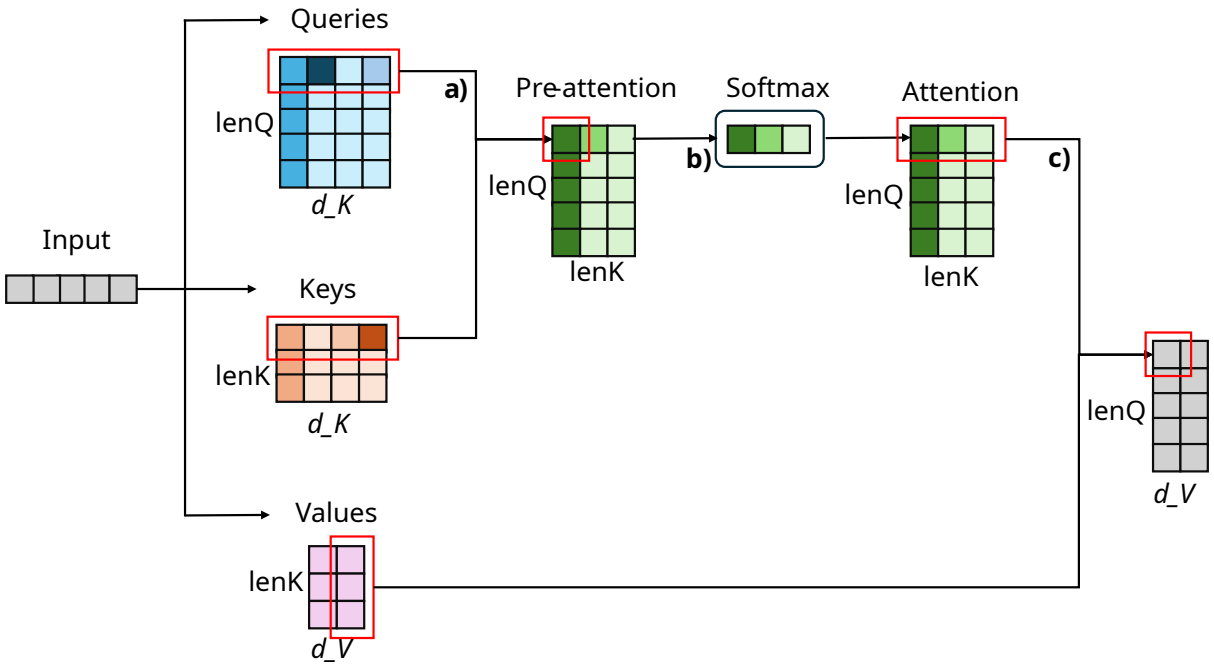


Figure 3.2: Stylized scheme of the self-attention mechanism. **a)** Matrix product between *queries* and *keys*, yielding a pre-attentional matrix. **b)** Softmax is then applied to each row. **c)** Finally, *values* are multiplied with the attention matrix, in a process similar to a weighted average.

values (V). Conceptually, queries represent the element currently being processed, while keys represent the elements in the sequence against which the query is matched. These two vectors encode different representations of the sequence elements and are optimized for comparison. The values (V) encode the semantic content of the element, from which the final output is aggregated.

Mathematically, scores are obtained by computing the dot product of the query with all keys. These scores are scaled and passed through a softmax function to obtain weights, which are then applied to the respective values and summed. The final result is a weighted vector reinterpreting the input sequence. The operation is formally defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3.1}$$

where d_k is the dimension of the key vectors, acting as a scaling factor to stabilize gradients. See Fig. 3.2 for a representation of the self-attention mechanism.

To enhance the model's expressivity, the Transformer typically employs *Multi-Head Attention*. This involves performing the attention function in parallel across h different "heads," allowing the model to jointly attend to information from different representation subspaces at different positions. Furthermore, since the self-attention mechanism is permutation-invariant—meaning it treats the input as a set rather than a sequence—the Transformer requires explicit *Positional Encodings*. These are added to the input embeddings to inject information about the relative or absolute position of the tokens in the sequence, ensuring that the temporal order is preserved. See

3.5.2 Adaptations for Time Series Forecasting

While the vanilla Transformer offers significant advantages in capturing global dependencies, its direct application to time series forecasting presents unique challenges, primarily due to the quadratic computational complexity ($O(L^2)$) of the self-attention mechanism with respect to sequence length. To address

this, several specialized transformer architectures have been proposed in the time-series domain, modifying the attention mechanism and restructuring the overall architecture to accommodate continuous data and forecasting tasks.

One prominent solution to the efficiency bottleneck is the Informer model. It introduces a *ProbSparse* self-attention mechanism, which is based on the observation that the attention probability distribution is often sparse, meaning only a few keys contribute significantly to the query. By selectively focusing on the most prominent queries using a logarithmic complexity ($O(L \log L)$), the Informer significantly reduces memory usage and computational time, making it feasible to predict over much longer horizons.

Departing from the point-wise attention focus, the Autoformer introduces a decomposition architecture that explicitly separates the series into trend-cyclical and seasonal components. It utilizes an *Auto-Correlation* mechanism as a substitute for the standard self-attention. Instead of calculating point-wise relationships, the Auto-Correlation block discovers period-based dependencies by aggregating similar sub-series based on time delays. This approach provides a stronger inductive bias for periodic time series, allowing the model to capture complex temporal patterns more effectively than canonical attention.

More recently, the PatchTST (Patch Time Series Transformer) has shifted the modeling paradigm by treating time series as a set of non-overlapping temporal patches rather than individual time steps. In this architecture, each patch is flattened and linearly projected into a vector of a higher dimension, serving as the input embedding (or token). for a standard Transformer encoder. This method offers dual benefits: it drastically reduces the effective sequence length, thereby lowering computational costs, and it preserves local semantic information within each patch. Notably, PatchTST employs a channel-independent strategy, where a single Transformer backbone is shared across all variable channels. This allows the model to learn universal temporal patterns rather than overfitting to specific channel correlations. A final linear layer serves as the prediction head, projecting the encoded representations into the future forecast horizon.

Chapter 4

Experimental Results in Vigilance Forecasting

Forecasting EEG features aligns with emerging needs in monitoring mental states over time and could be integrated with existing multimodal approaches by anticipating future α-frequency band oscillations. This forecasting approach, even over a short temporal horizon, could enable systems to proactively adjust task parameters or robotic behavior, thereby maintaining optimal operator engagement and reducing fatigue. For example, authors of [59] use both EEG spectral features and functional near-infrared spectroscopy to predict missed beeps (users' reactions) in the defined task, authors of [61] denote the importance of EEG power spectrum to predict user stress, while authors of [66] affirm their importance as a predictor of perceived engagement. The possibility of anticipating the future power spectrum, combined with the intrinsic cyclicity of HRC tasks, could lend insight into upcoming human factors.

4.1 Encoder-Decoder GRU Results

This section reports the first noteworthy results published in [92]. A significant accomplishment came from comparing VAR, TCN, GRU, and ED-GRU in forecasting upcoming vigilance oscillations. After performing an aggressive data filtering, the comparison showed the ED-GRU architecture capable of predicting α behavior up to 5.5 s into the future with satisfying accuracy, corresponding to half of a robot operation. Data came from the custom vigilance dataset detailed in Sec. 2.2.3 and was initially preprocessed following the guidelines described in Sec. 2.2.4. The entire procedure, from feature extraction to model specifics, is summarized by Fig. 4.1.

4.1.1 Features engineering

The wavelet transform can be used to extract activities from specific EEG bands, as it reveals temporal variations in frequency content with high temporal and frequency resolution. An overcomplete representation of the signal can be obtained using the continuous wavelet transform (CWT), which provides a sample-by-sample representation of signal variations across frequency bands. To capture a significant time-variant index of vigilance activity, the α band amplitude component was extracted through the CWT [93] from the preprocessed EEG data, specifically using the Morse wavelet. Given the mother wavelet $\psi(t)$, a family of scaled and shifted wavelets $\psi_{a,b}(t)$ captures time-variant frequency information over different scales. This information is encapsulated by the wavelet coefficients $CWT(a,b)$, expressed as:

$$CWT(a,b) = \int_{-\infty}^{\infty} y(t)\psi^*_{a,b}(t)\ dt \tag{4.1}$$

where $*$ is the complex conjugate. Parameter a determines the temporal position parameter and b is the wavelet scale.
The Morse wavelet is expressed in the frequency domain as:

$$\Psi_{P,\gamma}(\omega) = U(\omega)\alpha_{P,\gamma}\omega^{\frac{P^2}{\gamma}}e^{-\omega^\gamma} \tag{4.2}$$

where $U(\omega)$ is the frequency unit step, $\alpha_{P,\gamma}$ is a normalizing constant, parameter γ controls the symmetry of the wavelet in time, parameter P is the duration or inverse bandwidth and defines the number of oscillations in the wavelet [94]. The default Matlab parameters were kept (γ set to 3, P set to $\sqrt{60}$). Notably, $\gamma = 3$ corresponds to a Morse wavelet having minimum Heisenberg area and zero skewness. The squared absolute modulus of the 8-12 Hz coefficients was averaged to extract a time-frequency signal of α band power, represented by $\mathbb{E}[|CWT_\alpha|^2]$. These resulting signals were filtered (low-pass 4th-order Butterworth filter, cutoff frequency at 0.12 Hz) and decimated by a factor 64 to extrapolate the underlying trend, resulting in a 4 Hz sampling frequency (Fig. 4.1, "CWT" yellow block). This aggressive smoothing yields a tractable signal that allows predictions up to seconds into the future, without removing the main dominant frequency that modulates amplitude with respect to the task pacing (see Section 4.2.1).

4.1.2 Channel Selection

To identify channels exhibiting a significant increase in α power, a 1st-order polynomial was fitted by the least-squares method on channel data. Then, the estimated β_1 model coefficient was normalized to the range [0, 1] for each subject, along each channel, and averaged across all subjects. Centro-parietal channels (C3, C4, P3, and P4) were found to have a greater slope, aligning with findings in existing literature [95, 14]. Consequently, these four channels were selected as the primary focus of prediction.

4.1.3 VAR specifics

For the VAR model implementation, we relied on the statsmodel library [96], using the VARMAX and VARMAXResults classes (`https://www.statsmodels.org/stable/generated/statsmodels.tsa.statespace.varmax.VARMAX.html`), which allows for the model to dynamically add new samples of data to formulate a prediction, potentially updating its internal parameters, through the usage of the Kalman filter.

To reduce the amplitude variance in the signal, evaluations were performed using the log-transformed data of each individual. Before applying the logarithm, data were shifted as to be ≥ 1. Following this transformation, each subject was treated independently. To make a parsimonious choice among the candidate VAR models having different orders, the Akaike Information Criteria (AIC) was used [76]. A VAR model with parameters $p = 6$ (6^{th} order AR component) was selected for all subjects. For each subject, the VAR was first fitted on the initial two minutes of the acquisitions, comprising a total of 480 data points. As new samples were presented, they were integrated into the model using the *extend()* functionality of statsmodels. At each new sample, the subsequent 44 samples (11 s) were forecast. The model was periodically refitted over the newly presented data, encompassing a total of 480 data points. A representation of the entire process can be found in Fig. 4.1, specifically in the "VAR MODEL" red block.

4.1.4 DL Model Training

The DL models were implemented using Keras (https://keras.io/) and evaluated using a leave-one-subject-out cross-validation strategy. Specifically, for each subject, data was trained on the other nine to properly capture the model's cross-subject capabilities. Each model was trained on the complete temporal data (roughly 5,500 samples) from 9 of 10 subjects and tested on the remaining subject. The data were normalized to the range [0, 1] using min-max normalization. The data were subdivided into overlapping windows, shifted from sample to sample. The total number of training windows was roughly

50,000 for each subject-specific model, while the total number of testing windows was roughly 5,000. Models were thus trained to forecast future trends after being presented with a segment of fixed length. Model performance was affected by different window sizes; for both GRU and ED-GRU models, the optimal window size resulted in 44 samples, whereas for the TCN model, it was 132 samples. The window size was selected after performing a grid search on the significant time windows of 44, 88, and 132 samples, corresponding to 1 (i.e. 11 s), 2 (i.e. 22 s), and 3 (i.e. 33 s) pick-and-place operations in the past, respectively.

DL models were trained using the Adam optimizer with a learning rate of 0.001 over 30 epochs and a batch size of 32. Exponential decay (decay rate of 0.95) was used as a learning rate scheduler. The mean absolute error (MAE) was used as the loss function: out-of-sample points were compared against the future data for the ED-GRU and TCN, while for the GRU the difference between current and future data points was used, a strategy inspired by [97].

4.1.5 Performance Evaluation

For both the VAR and DL model evaluations, out-of-sample forecasts were computed at each time step, spanning up to 44 samples in the future. The selection of the 44^{th} forecast was based on its correspondence with the duration of a robot pick-and-place operation, i.e. 11 s. The term "k-ahead prediction" or forecast refers to predicting a sample k steps ahead in the future. Performance evaluations were calculated for each k-ahead prediction, using two quantitative metrics: the MAE and the coefficient of determination (R^2). The MAE for every k-ahead horizon, $1 \leq k \leq 44$, was computed as follows:

$$\mathrm{MAE}_k = \frac{1}{N} \sum_{i=1}^{N} |\hat{y}_{i,\,k} - y_{i,\,k}| \tag{4.3}$$

where N is the total length of the signal, $\hat{y}$ is the forecast data k-ahead and y is the actual data k-ahead. In the same way, the R^2 was computed as:

$$R^2{}_k = 1 - \frac{\sum_{i=1}^{N} (\hat{y}_{i,\,k} - y_{i,\,k})^2}{\sum_{i=1}^{N} (y_{i,\,k} - \mu(y_{i,\,k}))^2} \tag{4.4}$$

with $\mu(y)$ being the mean of y. Before metric evaluation, data predicted by the VAR were normalized using the same minimum and maximum values to ensure consistent MAE estimation.

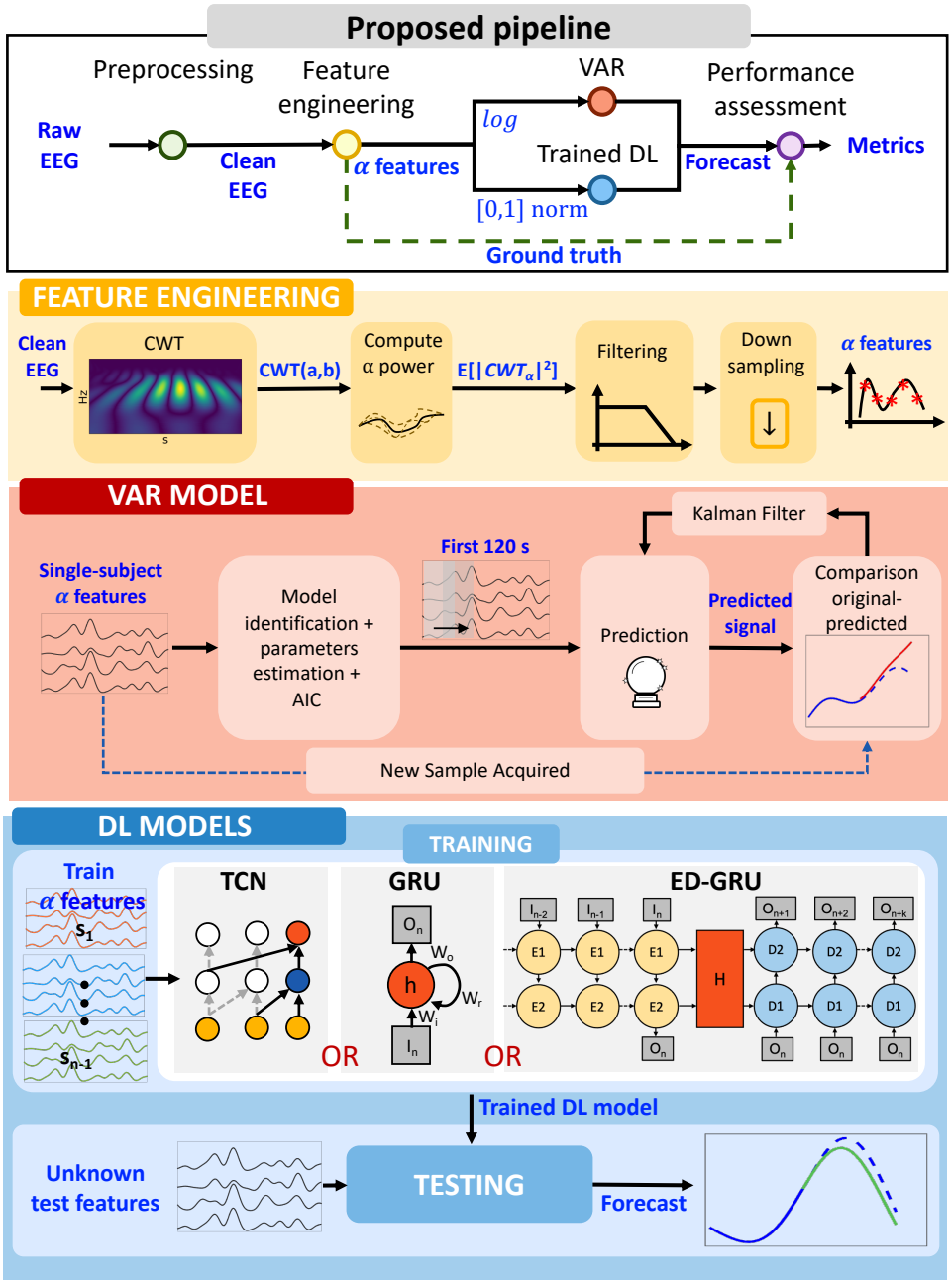


Figure 4.1: Overview of the vigilance forecasting pipeline. Raw EEG is preprocessed via CWT to extract power. Future values are predicted using either a statistical VAR model with iterative Kalman updates or Deep Learning architectures (TCN, GRU, ED-GRU) evaluated on left-out subjects.

4.2 ED-GRU Results

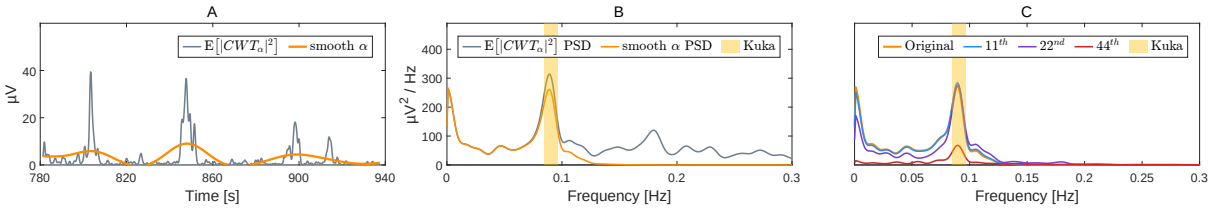


Figure 4.2: A) In gray, the squared absolute modulus of the 8-12 Hz coefficients averaged to extract a temporal representation of α band power (α features) for subject 2. Electrode P4 is shown as representative of the 4 channels showing similar behavior. Overlayed in orange, the resulting filtered signal (low-pass filter with cutoff at 0.12 Hz). B). The power spectral density of the entire signal is represented. The range (mean value ± sd) of frequencies of the Kuka pick and place operations executed during this specific task is reported as a yellow area. C) Power spectral density of the original signal and the predicted signals at different forecast horizons (11, 22, 44 samples into the future), for the ED-GRU model.

4.2.1 Results

After analyzing the behavior of the decimated α power signal, the performance of each proposed model is reported.

α oscillations

The extracted α features are smooth and slowly drifting, and in particular, subjects showed increased amplitude oscillations in the latter part of the task (after 16 min, as seen for two subjects in Fig. 4.4, third column). Interestingly, by observing the amplitude spectrum of the signal, an evident peak is denoted around the 0.09 Hz frequency for all subjects, as shown in Fig. 4.2. The peak is preserved by the filtering step, as shown in Figs. 4.2 A and B. This peak matches with the frequencies of the Kuka pick-and-place operations, which fluctuate around 0.09 Hz, as exemplified by the yellow area in Fig. 4.2, B. Small changes in the location of this peak can be explained by considering that the duration of an operation is not fixed, as the robot may perform a shorter path in some movements, and the transmission of the pick-and-place command inherently introduces a variable delay.

Model performances

We compared four methods for forecasting the extracted α features, demonstrating the promising performance of the ED-GRU model in achieving good results for predicting 22 samples (5.5 s) in the future. The predictive power of the algorithms in forecasting EEG signals is directly linked to MAE and R^2 values. Fig. 4.3 reports results for the VAR model, the TCN, the single GRU cell model, and the ED-GRU architecture. Each square represents the averaged MAE and R^2 across subjects and channels for a given forecast horizon ranging from 1 to 44 samples (11 s). The whisker lengths are determined by the standard deviation across subjects. Whiskers are omitted for the VAR and TCN models to improve graph readability. The right part of Fig. 4.3 focuses on model performance at each subject's 11^{th} horizon, averaged over the four channels. Fig. 4.4 visually represents the ground truth (in black) and forecasting signals for subject 4 in three different time intervals of the task. In particular, four different time forecast horizons are shown: 6^{th} (1.5 s), 11^{th} (2.75 s), 22^{nd} (5.5 s) and 44^{th} (11 s). VAR performance drops drastically after the 6^{th} temporal horizon when looking at the R^2 metric, and the unreported variance indicates substantial variability between subjects. ED-GRU performance gradually worsens but remains consistent among subjects, yielding good outcomes (MAE of 0.048, R^2 of 0.726) up to the 22^{nd} horizon (5.5 s), as shown in Fig. 4.3. GRU results show the same behavior, degrading as the

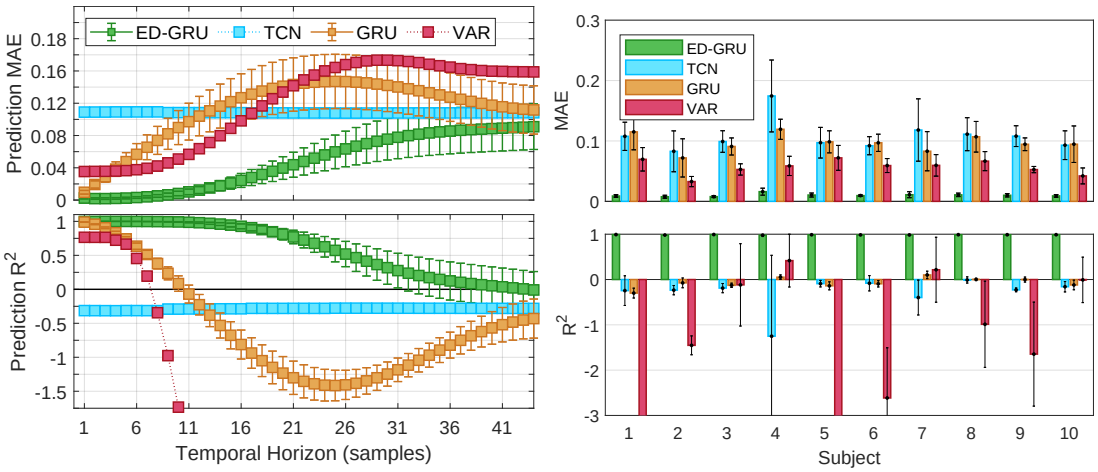


Figure 4.3: Comparison of the forecast performance of the different models. Left. Grand average MAE and R^2 values for the four models for each temporal horizon ranging from 1 to 44. Right. Mean performance, at the 11^{th} (i.e. 2.75 s) forecast horizon, for the four models, for each subject. Error lines represent the standard deviation across channels. VAR results are cropped out for clarity.

forecast horizon grows and consistently performing worse than the ED-GRU. This is because the GRU model tends to replicate the last possible input value, as seen in Fig. 4.4. TCN does not perform well: this model outputs a signal that does not change much across forecast horizons, as seen in Fig. 4.4.

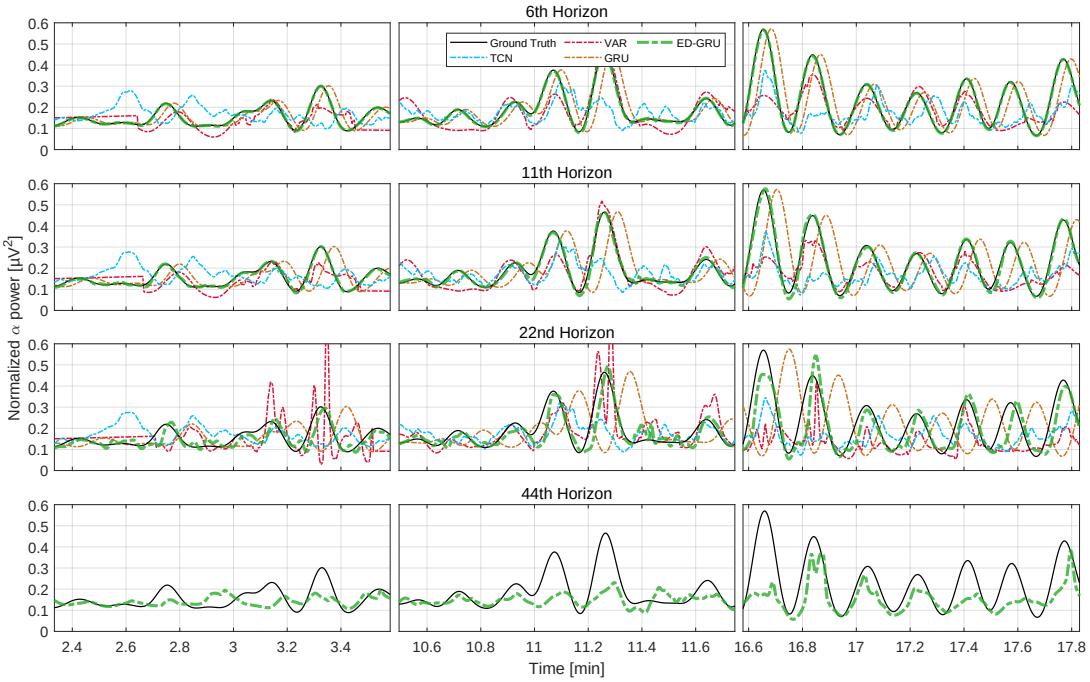


Figure 4.4: Ground truth (in black) and estimated signals at the 6^{th} (1.5 s), 11^{th} (2.75 s), 22^{nd} (5.5 s) and 44^{th} (11 s) forecast horizon, for all the four models at channel P3, for a representative subject (n. 4). Three windows from different time intervals of the task are reported (i.e. 2.3-3.6, 10.4-11.8, and 16.6-17.8 minutes). For the 44^{th} horizon, only the best performing model (i.e. ED-GRU) is reported.

Horizon	1	6	11	22	33	44
MAE	0.0026	0.0023	0.0021	0.0162	0.0366	0.0459
R^2	0.9978	0.9983	0.9981	0.9013	0.4980	0.2097

Table 4.1: Aggregated MAE, R^2 for all subjects, all channels, using the specified PatchTST architecture.

The plot of the 44^{th} forecast for ED-GRU, as seen in Fig. 4.4, shows a hint of cyclicity still captured by the model, albeit with strongly reduced amplitude. Interestingly, Fig. 4.2, C, shows that for the first forecast horizons, up to 11, the spectral amplitude is similar. For the 22^{nd}, the main peaks overlap, but the energy at different frequencies is slightly reduced. At the 44^{th}, the spectral energy has been significantly dampened; however, the predominant peak is still found at 0.09 Hz. This behavior is consistently shared among 9 of the 10 subjects.

4.3 Transformer Results

Using a Transformer-based model, PatchTST, allows for even more precise forecasting. Even if model accuracy rapidly decays for long-term horizons, performance is greatly improved at the 22nd time step. Both these facts can be seen in Tab. 4.1, which reports MAE and R^2 at specific time steps. For a broader understanding of how performance changes with time, see Fig. 4.5. To visualize how the forecast behaves at specific thresholds, see Fig. 4.7 and Fig. 4.6.

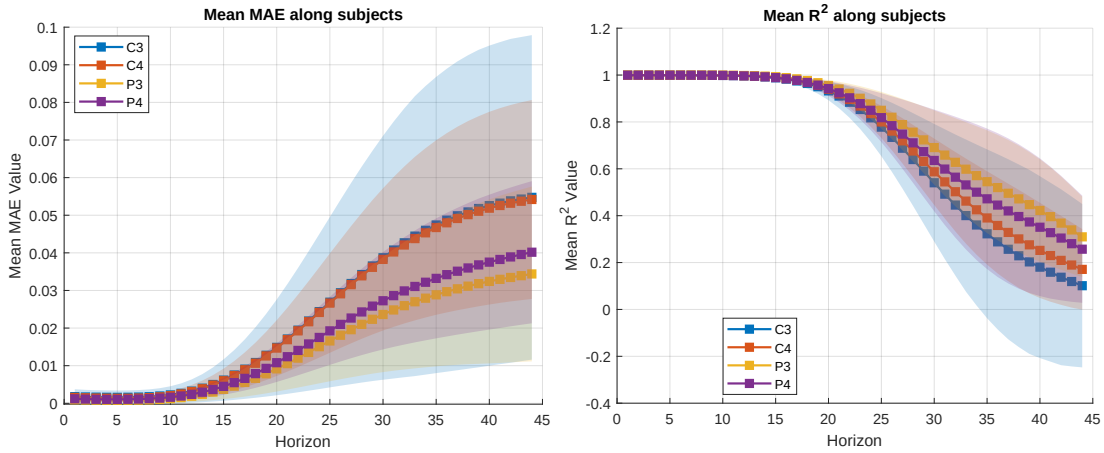


Figure 4.5: Comparison of the forecast performance of the PatchTST model. Grand average MAE and R^2 for the four channels, for temporal horizons for each temporal horizon ranging from 1 to 44. The shaded area represents the standard deviation across subjects.

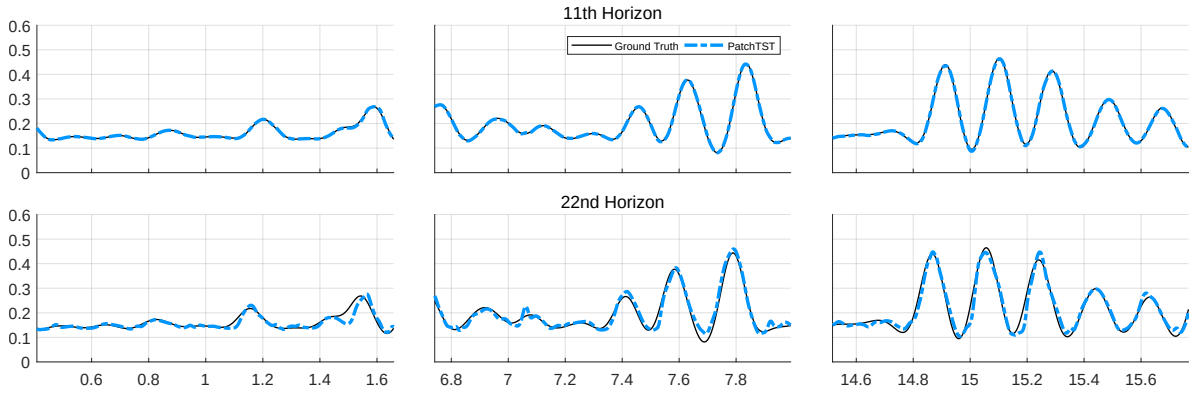


Figure 4.6: Ground truth (in black) and PatchTST estimated signal (light blue) at the 11^{th} (2.75 s), and 22^{nd} (5.5 s) forecast horizons, at channel P3, for a representative subject (n. 4). Three windows from different time intervals of the task are reported (i.e. 2.3-3.6, 10.4-11.8, and 16.6-17.8 minutes). Notice the extreme accuracy at low forecast horizons.

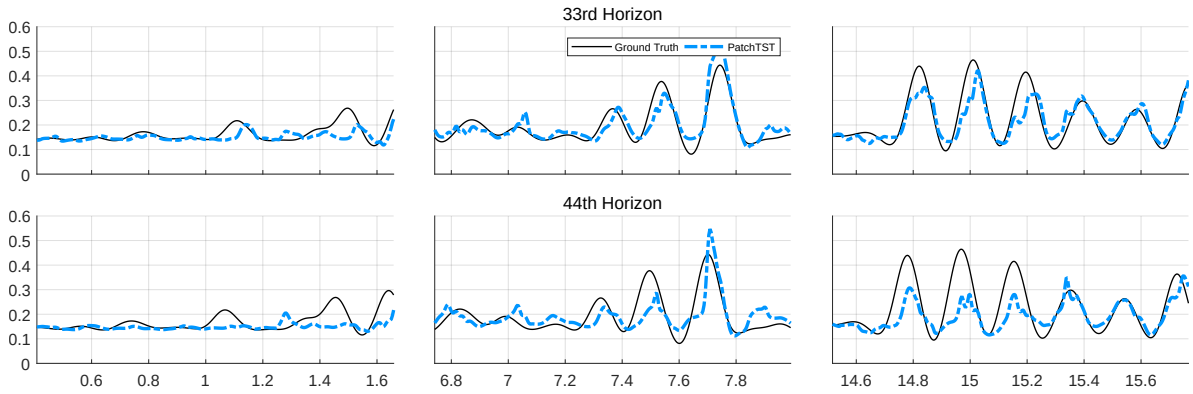


Figure 4.7: Ground truth (in black) and PatchTST estimated signal (light blue) at the 33^{rd} (8.25 s), and 44^{th} (11 s) forecast horizons, at channel P3, for a representative subject (n. 4). Notice how the model can still anticipate some fluctuations for the 33rd horizon.

Chapter 5

Generative Methods for EEG signals

5.1 Generative Strategies for EEG

Deep Learning (DL) research within the Brain-Computer Interface (BCI) domain is fundamentally hindered by the scarcity of large-scale, high-quality, and generalized datasets. Unlike computer vision or natural language processing, where massive repositories exist, BCI data acquisition is subject to severe constraints. Publicly available datasets are frequently domain-specific, strongly biased by experimental protocols, and dependent on the specific hardware configurations used. Furthermore, significant inter- and intra-subject variability—caused by non-stationarity in brain signals and anatomical differences—must be rigorously accounted for to achieve generalization [98]. The acquisition process itself is labor-intensive, time-consuming, and requires constant maintenance, particularly when utilizing wet electrodes that rely on conductive gels or saline solutions, which are prone to impedance degradation over time [99, 100].

Given these inherent limitations, enriching existing repositories with synthetic data presents a compelling solution. Synthetic data augmentation offers a pathway to overcome critical hurdles such as data scarcity and the severe class imbalance often observed in pathological datasets or rare event paradigms. Moreover, it facilitates privacy-preserving data sharing through synthetic anonymization, allowing researchers to disseminate realistic datasets without compromising subject confidentiality.

The literature addresses the data scarcity problem through a spectrum of solutions, ranging from deterministic signal manipulation to stochastic generative models. Traditional data augmentation strategies typically rely on geometric or temporal transformations. These include window splicing [101], affine deformations and noise injection [102], or the interpolation of existing signal vectors [103]. While effective to a degree, these methods are limited by their reliance on existing samples; they perturb known data points rather than learning the underlying manifold of the data distribution, often failing to introduce sufficient diversity for robust model training.

Consequently, there has been a paradigm shift toward Generative Artificial Intelligence (GenAI), with a predominant focus on Generative Adversarial Networks (GANs). First introduced in the seminal work of Goodfellow et al. [26], a GAN operates on a game-theoretic minimax principle involving two competing neural networks: a *Generator* (G), which synthesizes data candidates from a latent noise vector, and a *Discriminator* (D), which evaluates them against real samples. Theoretically, training converges to a Nash equilibrium where the generator reproduces the true data distribution so accurately that the discriminator serves as a random guesser.

In the specific context of time-series data, architectures such as TimeGAN have extended this framework to capture temporal dynamics, enforcing stepwise correlations alongside global feature distributions. Despite their potential, GANs are notoriously difficult to stabilize; they frequently suffer from non-convergence, vanishing gradients, and mode collapse, a failure state where the generator outputs a limited variety of samples, neglecting significant portions of the target distribution. For a detailed discussion on architectural modifications to mitigate mode collapse, refer to Chapter 7.

Following their introduction, GAN architectures were rapidly adapted to the BCI domain. Initial experiments demonstrated the feasibility of generating realistic raw EEG samples [104] as well as EEG-derived features, such as differential entropy maps [105, 106] or wavelet energy vectors [107]. Since these early implementations, numerous GAN variants have been proposed to address specific BCI paradigms, including Steady-State Visual Evoked Potentials (SSVEP), Motor Imagery (MI), and emotion recognition tasks [108, 109]. Building upon this body of work, we recently introduced a preliminary framework for EEG rhythm generation via GANs specifically tailored for passive BCIs [110], which constitutes the foundational architecture for the developments presented in this thesis.

5.2 Generative Adversarial Networks

Generative Adversarial Networks (GANs), first introduced by Goodfellow et al. in 2014 [26], represent a pivotal advancement in generative modeling. They offer a compelling solution to the chronic shortage of high-quality real-world data in Machine Learning tasks by learning to synthesize artificial samples that faithfully mimic the underlying data distribution.

The GAN architecture is founded on game-theoretic principles, specifically a zero-sum minimax game played between two competing neural networks: a *Generator* (G) and a *Discriminator* (D). The Discriminator acts as a binary classifier, tasked with distinguishing between genuine data samples drawn from the real distribution p_{data} and synthetic samples produced by the Generator. Conversely, the Generator takes a random noise vector z (sampled from a prior distribution p_z, typically Gaussian or Uniform) and maps it to the data space, aiming to create samples $G(z)$ that are indistinguishable from real data.

The training process is adversarial: D strives to maximize the probability of correctly assigning the label "real" to actual data and "fake" to generated data, while G strives to minimize the probability that D correctly identifies its outputs as fake. This dynamic creates a competitive optimization landscape where the improvement of one network drives the evolution of the other. The standard value function $V(G, D)$ for this minimax game is defined as:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{\text{data}}(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [\log (1 - D(G(z)))] \tag{5.1}$$

Ideally, if the training process remains stable, the system converges to a Nash equilibrium where the Generator recovers the true data distribution ($p_g = p_{\text{data}}$) and the Discriminator is reduced to random guessing ($D(x) = 0.5$ for all inputs). Once training is complete, the Generator can be detached and deployed independently to synthesize novel data samples [111].

Notable implementations for the temporal domain include the RCGAN (2017) [112] and the TimeGAN (2019) [113]. RCGAN incorporates Recurrent Neural Networks (RNNs), such as LSTMs or GRUs [114], to natively model temporal dependencies and, as a conditional model, generate sequences based on specific external labels. TimeGAN involves a hybrid architecture, combining an autoencoder, an adversarial network, and a supervised loss. The autoencoder learns a compact latent representation of the data, the adversarial network operates within this latent space to capture the temporal dynamics, and the loss explicitly enforces the step-by-step temporal relationships. Subsequent inspired works leverage a similar multi-component architecture to preserve anonymity in biomedical data by introducing perturbations [115], or to expand stock volatility and power-consumption datasets [116, 117]. More recent, specific models worth citing, especially in the context of finance, are Tail-GAN and the Feature-Enriched GAN (FE-GAN). The Tail-GAN architecture modifies the discriminator's objective by forcing the identification of discrepancies in the tail quantiles of the original distribution, shifting more importance to extreme events [118]. FE-GAN injects an additional input sequence, derived from preceding data (such as moving averages, RSI, or volatility indicators), into the generator. This provides guidance on a specific market regime and enables explorations of the model's latent space [119]. Despite their flexibility, GANs' training dynamics are prone to instabilities, namely mode collapse, a critical failure state where the generator produces only a limited subset of high-probability samples. Furthermore, evaluating GAN performance - by assessing the fidelity and diversity of generated temporal data - remains difficult and often relies on indirect, qualitative, or downstream task-based metrics.

5.3 Evaluation Criteria for Synthetic EEG

Evaluating the output of generative models, particularly within the complex domain of temporal data, constitutes a significant challenge for which no singular, universally accepted metric exists. In contrast to supervised learning paradigms, where performance can be summarily quantified by accuracy or F1-scores, the assessment of generative models necessitates the quantification of more abstract concepts, including output quality, diversity, and fidelity [111]. Consequently, a robust evaluative framework cannot rely on a single score but must instead be predicated upon a suite of complementary methodologies. The literature typically categorizes these methodologies into three non-exclusive strategic classifications: statistical fidelity, functional utility, and qualitative inspection.

Employing a multi-faceted benchmarking approach is standard practice in the field. For instance, Qian et al. [120] performed both quantitative and qualitative analyses over diffusion-generated temporal data, while Cinquetti et al. [121, 28] used a combination of autocorrelation analysis, Fréchet Inception Distance, and forecasting performance to validate EEG-like data. Similar multi-metric benchmarks are also requisite in other domains, such as textual data generation [122].

Statistical Fidelity Assessment

This strategic classification assesses the congruence between the statistical properties of the synthetic data and those inherent in the real-world reference dataset. A primary method involves comparing marginal distributions via histograms or Kernel Density Estimates applied to individual features. While necessary, this approach is insufficient on its own as it demonstrably disregards temporal interdependencies.

For time-series data, preserving temporal dynamics is critical. This is commonly evaluated through a comparative analysis of the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF). A high-fidelity model yields sequences whose ACF/PACF profiles are statistically indistinguishable from the original data. In the spectral domain, consistency is often checked by comparing Power Spectral Density (PSD) estimates [104].

To provide a single, comprehensive score of distributional similarity, we contributed by adapting the **Fréchet Inception Distance (FID)** to the time-series domain. Originally designed as a gold standard for synthetic images [123], the FID measures the distance between feature representations of real and generated data. In our adaptation, the standard Inception v3 network is replaced with a domain-specific feature extractor—specifically, a 2-layer deep ED-GRU architecture trained as a forecaster. The activations are captured from the penultimate pooling layer (the "coding layer") to serve as high-level embeddings. Assuming these embeddings follow a multivariate Gaussian distribution, the FID is calculated as the Fréchet distance (Wasserstein-2 distance) between the Gaussian fit of the real embeddings (μ_r, Σ_r) and the generated embeddings (μ_g, Σ_g):

$$\text{FID} = \|\mu_r - \mu_g\|_2^2 + \text{Tr}\left(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}\right) \tag{5.2}$$

where μ represents the mean feature vectors, Σ represents the covariance matrices, and Tr denotes the trace. A lower FID score indicates closer statistical proximity between real and synthetic distributions.

Functional Utility Assessment

This strategy ascertains the synthetic data's practical utility as a viable surrogate for real-world data. The most recurrent strategy is *Train on Synthetic, Test on Real* (TSTR). In this paradigm, a supervised downstream model—such as a classifier, anomaly detector, or forecaster—is trained on the synthetic (or augmented) dataset and subsequently evaluated against a real-world test set. An improvement in performance, or performance matching that of real data, indicates that the generative model has successfully captured the essential features required for the task.

Literature examples include Lotte et al. [101], who utilized Linear Discriminant Analysis (LDA) and Common Spatial Patterns (CSP) to evaluate augmented Motor Imagery classification. Similarly, Rommel et al. [102] assessed the performance of Convolutional Neural Networks (CNNs) for sleep stage

classification using synthetic augmentations, while Krell et al. [103] employed Support Vector Machines (SVM) for P300 oddball paradigms. While TSTR provides a measure of utility, it serves as an indirect metric and does not necessarily guarantee statistical fidelity.

Qualitative and Heuristic Inspection

Finally, human perceptual faculties remain essential for identifying failures often overlooked by quantitative metrics. Simple visual inspection, achieved by plotting synthetic samples alongside real ones, can reveal evident issues such as a lack of high-frequency detail, recurring artifacts, or distributional mode collapse—where the model repeatedly generates identical or nearly identical samples.

A more refined qualitative strategy relies on dimensionality reduction algorithms, most notably t-distributed Stochastic Neighbor Embedding (t-SNE) or Principal Component Analysis (PCA). These techniques project high-dimensional datasets into a two- or three-dimensional visualization space. In a successful generation, the clusters formed by the synthetic data should demonstrate substantial overlap with the clusters of the real data [105, 109]. Failures are visually apparent: the formation of distinct, isolated synthetic clusters is indicative of mode collapse, while a failure to populate regions occupied by the real data signifies poor distributional diversity. For a detailed discussion on geometry-based data visualization techniques, see Chapter 7.

Chapter 6

Evaluation of Generative Models

6.1 Study Overview

A first significant study was conducted utilizing two distinct EEG datasets: one capturing resting-state conditions and the other recording subjects during an observational task within an industrial setting. The primary objective was to synthesize time-variant EEG frequency band power indices using GANs. Subsequently, we investigated whether augmenting real datasets with this synthetic data could enhance the predictive performance of signal forecasting models.

The validation pipeline consisted of comparing two forecasting models for each dataset: a baseline model trained exclusively on real data and an experimental model trained on an augmented dataset containing both real and synthetic samples. To assess the fidelity of the generated samples, we computed a domain-adapted version of FID [123]. Through these methodologies, this work aims to establish a robust framework for the evaluation of deep learning-based EEG synthesis.

6.2 Experimental Paradigms

6.2.1 Datasets

The publicly available Sustained Attention to Response Task (SPIS) Resting State Dataset was utilized as a benchmark for relaxed cognitive states [124]. It contains recordings from 10 subjects during both eyes-closed and eyes-open sessions. Data were acquired using 64 Ag/AgCl electrodes positioned according to the international 10-10 system, sampled at 2048 Hz (downsampled to 256 Hz for this study). Each acquisition session lasted approximately 2.5 minutes. The dataset is available at `https://github.com/mastaneht/SPIS-Resting-State-Dataset`.

To represent active cognitive monitoring, a vigilance dataset was acquired based on the protocol described in Section 2.2.3. Data collection was performed at the Industrial Computer Engineering (ICE) Laboratory, University of Verona. Ten subjects were recorded while observing a robotic arm cycling through three stations to move 2x4 cm blocks. Each station had a "correct" and a "wrong" position, occurring with probabilities of 80% and 20%, respectively. Subjects were required to assess the correctness of the block position via button press. The task comprised 40 trials, totaling approximately 23 minutes of recording. Signals were acquired at 256 Hz using 21 Ag/AgCl electrodes (10-20 system), with reference and ground electrodes placed anteriorly and posteriorly to Fz.

6.3 Data Preprocessing and Frequency Band Extraction

Both datasets underwent band-pass filtering (4–30 Hz). Artifacts were removed using Artifact Subspace Reconstruction (ASR) and Independent Component Analysis (ICA) algorithms [125].

To capture non-stationary neural dynamics, time-variant power components for θ, α, and β bands were extracted using the Continuous Wavelet Transform (CWT) [72]. The CWT provides an overcomplete time-frequency representation ideal for detecting transient spectral changes. We utilized the *cwt()* MATLAB function with default parameters (Morse wavelet basis).

Feature Selection for SPIS: For the resting-state dataset, analysis was restricted to the 8–12 Hz α band. Squared coefficients were averaged across frequencies for centro-occipital channels (O1, O2, PO3, PO4). These regions are heavily modulated by α reactivity, particularly the "Berger effect," where posterior α synchronization peaks during eyes-closed relaxation and desynchronizes upon visual stimulation [126, 127].

Feature Selection for Vigilance: For the industrial observation task, features were selected based on neurophysiological markers of vigilance. Parietal α power (P3, P4) was extracted as a primary marker, given its sensitivity to vigilance decrement and internal-external attentional shifts [14]. To increase generation complexity, frontal β (12–30 Hz) and θ (4–8 Hz) components were extracted from channels Fp1 and Fp2. Frontal β typically correlates with focused attention, while increased θ is indicative of mental fatigue [12].

6.4 Data Segmentation and Normalization

The extracted time-variant band powers are smoother than raw EEG signals yet preserve task-relevant dynamics, making them suitable targets for GAN synthesis. The continuous streams were segmented into fixed-length windows of 1 second (256 samples) with a stride of 64 samples (75% overlap). This resulted in 11,000 windows for the SPIS dataset and 55,000 for the Vigilance dataset. All data underwent min-max normalization based on global channel amplitudes.

To evaluate the impact of data scarcity, we created subsets containing 25% and 50% of the total available data by increasing the window stride by factors of four and two, respectively.

6.5 Generative Model Architecture

We designed a GAN framework tailored for time-series generation.

- *Generator (G):* The generator employs a Recurrent Neural Network (RNN) architecture, specifically utilizing Gated Recurrent Units (GRU). GRUs are adept at retaining long-term dependencies, making them ideal for sequential data generation [128, 85]. The final architecture consists of a stack of 5 bidirectional GRU layers with a hidden state size of 64. The input noise vector (channel $\times$ 64 Gaussian time series) is processed by an ensemble of noise-interpolating layers (linear layers interleaved with Leaky ReLU) to match the required input dimensionality.

- *Discriminator (D):* We adopted a 1-D Convolutional Neural Network (CNN) for the discriminator. Preliminary comparisons indicated that 1-D CNNs significantly outperformed GRU-based discriminators in classifying real versus synthetic time series [129]. The discriminator comprises 7 stacked 1-D convolutional layers (kernel size 4, stride 2) interleaved with Leaky ReLU and batch normalization. Filter depth scales progressively by a factor of 2, starting from 32.

Training was executed using the ADAM optimizer for 50 epochs with a batch size of 64. Learning rates were set to 1×10^{-5} for the discriminator and 2×10^{-5} for the generator. A leave-one-out approach was used to train subject-specific generators, from which 100,000 synthetic time series were sampled for each dataset split.

6.6 Evaluation Framework

6.6.1 Statistical Consistency

Output quality was initially assessed via the Cross-Correlation Function (CCF) to evaluate temporal consistency and inter-channel dependencies. To quantify the discrepancy between the real and synthetic CCF distributions, we computed the Weighted Root Mean Squared Error (WRMSE):

$$\text{WRMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\frac{\big(\mu_1(t)-\mu_2(t)\big)^2}{\sigma_1^2(t)+\sigma_2^2(t)}} \tag{6.1}$$

where $\mu_1(t), \mu_2(t)$ are the mean CCF values and $\sigma_1^2(t), \sigma_2^2(t)$ are the variances for real and synthetic signals, respectively. This metric penalizes deviations more heavily in regions of low variability.

6.6.2 Functional Utility: Forecasting Augmentation

To validate utility, we employed a standard paradigm using an ED forecasting model based on GRUs (Fig. 6.1). One model was trained on an augmented dataset and was compared to a baseline model trained on real data only. The ED model (2 stacked GRUs, hidden dimension 64) aims to predict the final 32 samples of a 1-second window given the preceding 224 samples.

Models were trained using leave-one-out cross-validation on three data configurations: (1) Real data only, (2) Real + GAN-augmented data, and (3) Real + Interpolation-augmented data. Performance was evaluated using MAE and SMAPE. Statistical significance was determined via the Wilcoxon signed-rank test ($\alpha = 0.05$).

6.6.3 Fréchet Inception Distance (FID)

We implemented a domain-specific FID score to assess distributional similarity. Instead of the standard Inception v3 image network, we utilized the encoder portion of the aforementioned ED-GRU forecasting model (trained solely on real data) as the feature extractor. The FID was computed between the distributions of the hidden states produced by real and synthetic samples, providing a quantitative measure of feature space overlap.

6.6.4 Baseline Comparison

To contextualize the performance of the GAN, we compared it against a naive augmentation baseline: linear interpolation. Synthetic samples were generated by averaging three randomly selected 1-second real windows with random weights summing to one. This provided a benchmark to determine if the complex GAN architecture yielded benefits over simple signal mixing.

6.7 Generation Results

To assess the efficacy of the proposed generative framework, 100,000 synthetic samples were generated for both the SPIS and Vigilance datasets, creating augmented repositories comparable in size to the original recordings. The evaluation focused on three key dimensions: (1) Signal Quality, assessed via L2 distance and Cross-Correlation Function (CCF); (2) Functional Utility, evaluated by comparing the performance of forecasting models trained on real versus augmented data; and (3) Distributional Similarity, quantified using a domain-adapted Fréchet Inception Distance (FID). These metrics were computed across three data availability scenarios: 100%, 50%, and 25% of the total dataset.

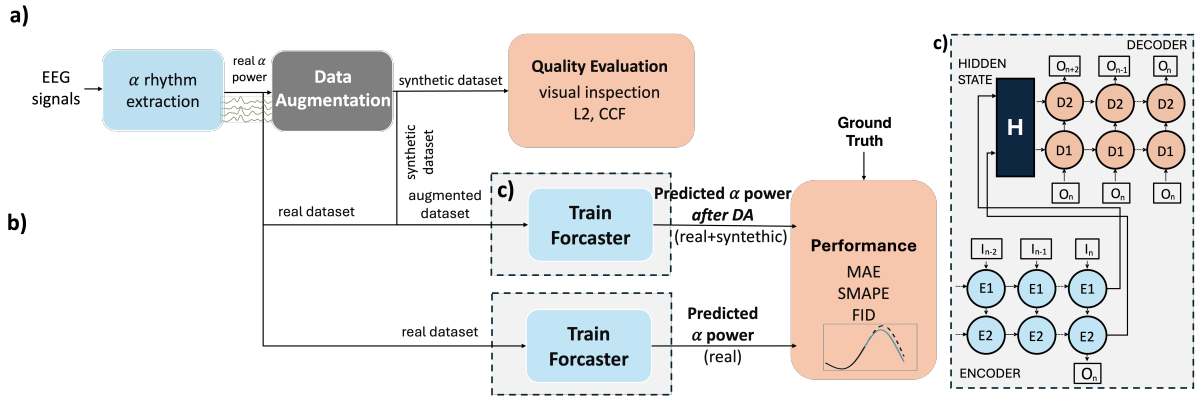


Figure 6.1: Workflow of the proposed method. (a) EEG signals are processed to extract the α rhythm. The real α power is used to train a GAN, which generates synthetic α power data, followed by the computation of L2 and CCF (b). Next, a forecasting model is trained on both original (real) and augmented (real + synthetic) datasets to predict the α power. Finally, the predicted α power, obtained from data augmented with both synthetic and real data, is compared to the α power obtained from real data alone using MAE, SMAPE, and FID.

6.7.1 Signal Quality Assessment

Amplitude Consistency (L2 Distance)

The L2 distance was computed to verify that the generated signals maintained realistic amplitude ranges. For the SPIS dataset, the mean L2 distances between real and synthetic signals were 0.087, 0.091, and 0.147 for the 100%, 50%, and 25% splits, respectively. For the Vigilance dataset, these values were 0.041, 0.055, and 0.074. Considering the average sample amplitude of the real signals (0.066 for SPIS; 0.018 for Vigilance), these results confirm that the generative models successfully learned to synthesize data with realistic energy profiles across all data splits.

Temporal and Spatial Correlations (CCF)

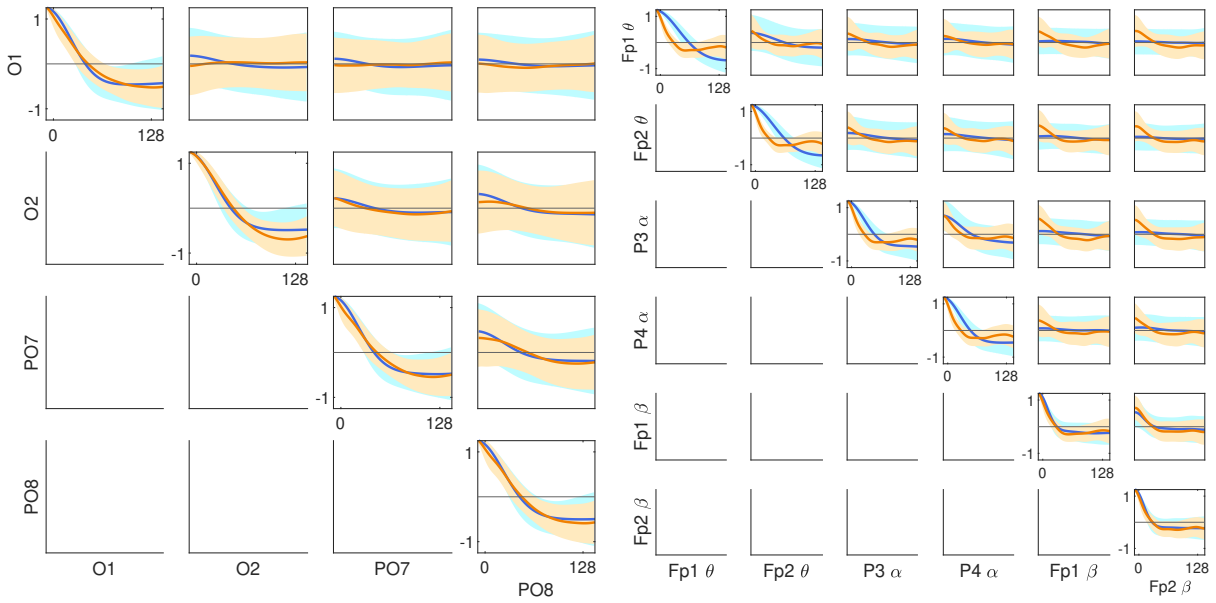
The analysis of the Cross-Correlation Function (CCF) revealed a high degree of affinity between real and synthetic time series when the full dataset was utilized. This alignment was quantified by low Weighted Root Mean Squared Error (WRMSE) values of 0.242 for Vigilance and 0.236 for SPIS. As illustrated in Fig. 6.2 (left), the synthetic data accurately mimics the temporal decay of auto-correlations (diagonal elements) and the variance of inter-channel dependencies (off-diagonal elements), suggesting robust replication of EEG rhythm structures.

However, model fidelity decreased as training data became scarcer. For the 50% split, WRMSE values rose to 0.332 (Vigilance) and 0.346 (SPIS). At the 25% split, degradation was particularly evident in the Vigilance dataset (WRMSE 0.444), where synthetic signals exhibited over-defined waveforms in off-diagonal correlations compared to the high-variance behavior of real signals (Fig. 6.2, right). This indicates a reduction in the model's generalization capabilities under severe data constraints.

6.7.2 Functional Utility: Forecasting Performance

The Encoder-Decoder (ED) model demonstrated the ability to predict the subsequent 32 data points of 1-second windows with reasonable accuracy (qualitative examples in Fig. 6.3). The quantitative impact of data augmentation is detailed in Table 6.1 and Fig. 6.4.

Forecasting models trained on datasets augmented with GAN-generated samples consistently outperformed baselines trained solely on real data. For the SPIS dataset, statistically significant improvements



(a) SPIS (100% Data)

(b) Vigilance (25% Data)

Figure 6.2: Mean CCF (solid lines) and standard deviation (shaded regions) for real (blue) and synthetic (orange) data. Left: Results for the full SPIS dataset show strong overlap. Right: Results for the 25% Vigilance split highlight the divergence in inter-channel correlations due to data scarcity.

in MAE and SMAPE were observed across all data splits. For the Vigilance dataset, improvements were significant for the 100% and 50% splits. However, the 25% split resulted in a slight performance decrement, aligning with the degradation observed in the CCF analysis. Notably, augmentation via linear interpolation resulted in drastic performance deterioration across all cases, validating the necessity of non-linear generative modeling.

		GAN Augmentation			**Interpolation Augmentation**		
Metric	**Dataset**	**100%**	**50%**	**25%**	**100%**	**50%**	**25%**
SMAPE	**SPIS**	26.2%	33.7%	27.3%	-67.4%	-33.3%	-15.0%
	VIGILANCE	19.0%	16.1%	-9.8%	-71.3%	-110.8%	-69.3%
MAE	**SPIS**	29.0%	46.4%	37.4%	-262.0%	-150.7%	-108.0%
	VIGILANCE	15.4%	21.2%	-2.5%	-150.5%	-243.2%	-127.4%

Table 6.1: Percentage improvement in forecasting performance (averaged across subjects) when augmenting real data with GAN-generated vs. Interpolated samples. Positive values indicate error reduction (improvement).

6.7.3 Distributional Similarity (FID)

FID scores were computed to quantify the distance between real and synthetic embeddings (Table 6.2 and Fig. 6.5). GAN-generated data yielded FID scores generally centered around ∼1.0 for SPIS and ∼0.5 for Vigilance, indicating a distribution that is comparable to, yet distinct from, the real data. In contrast, linearly interpolated data produced extremely low FID scores (∼0.01), suggesting near-perfect overlap with the real data distribution manifolds.

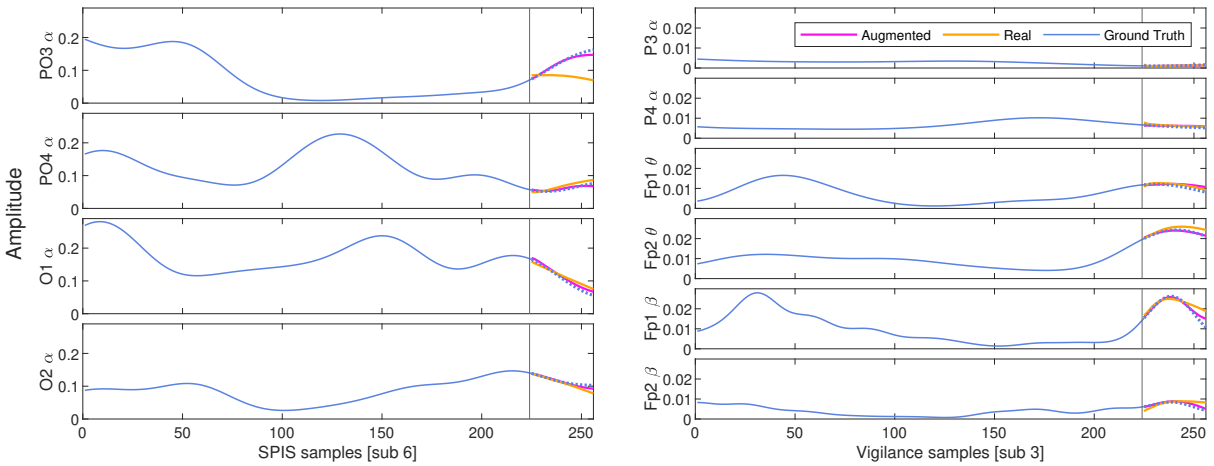
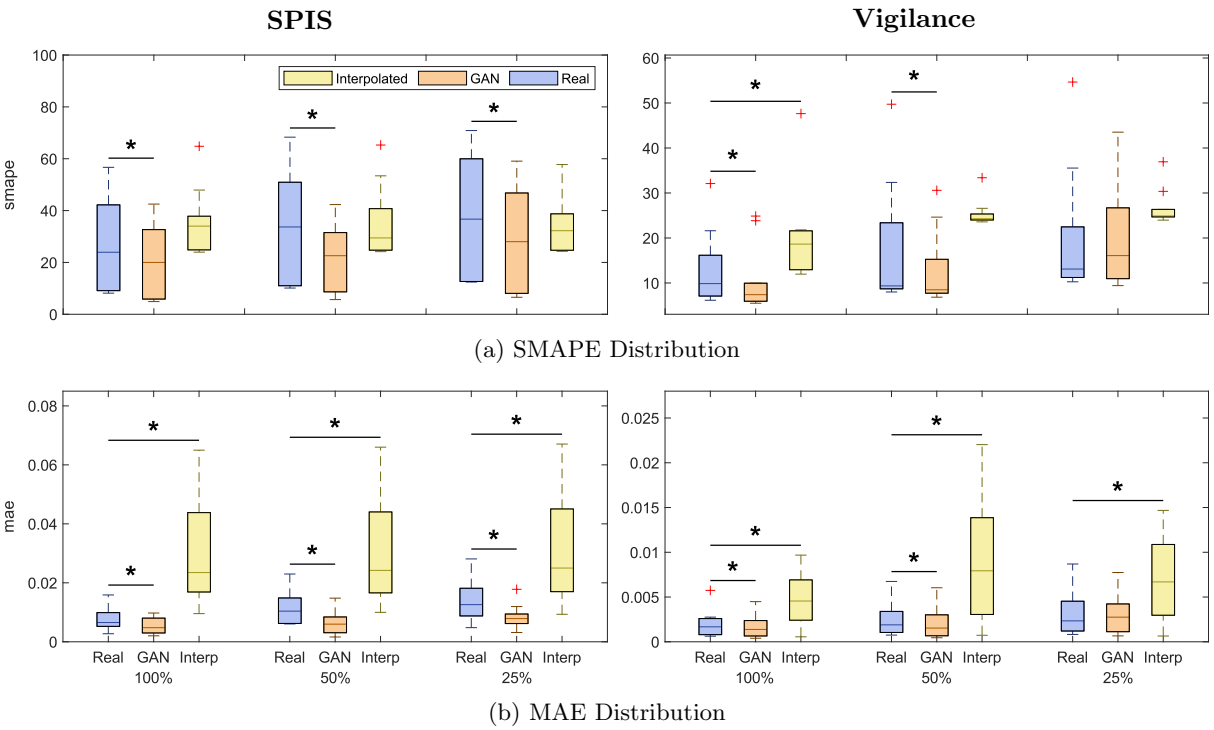


Figure 6.3: Forecasting examples for SPIS (left) and Vigilance (right) datasets using 100% of available data. Dotted lines represent Ground Truth. Predictions from models trained on Real data only (blue) vs. Real + GAN Augmented data (orange) are compared.



(a) SMAPE Distribution

(b) MAE Distribution

Figure 6.4: Error metric distributions across three forecasting models: Real-only, GAN-augmented, and Interpolation-augmented. Asterisks denote statistical significance ($p < 0.05$, Wilcoxon signed-rank test).

In the Vigilance dataset, two significant outliers were identified in the GAN measurements: Subject 8 (100% split) and Subject 5 (25% split), with FID values approaching 3.0. Removing these outliers adjusted the mean FID for the Vigilance GAN case to 0.21 ± 0.09 and 0.59 ± 0.13 for the 100% and 25% splits, respectively. For context, the baseline FID distance between real data and pure white noise was calculated to be $\sim$19.

Dataset	Method	100% Split	50% Split	25% Split
SPIS	**GAN**	1.02 ± 0.53	1.26 ± 0.71	1.43 ± 0.91
	Interp.	0.01 ± 0.00	0.01 ± 0.00	0.01 ± 0.00
Vigilance	**GAN**	0.49 ± 0.88	0.55 ± 0.22	0.90 ± 0.96
	Interp.	0.10 ± 0.04	0.00 ± 0.00	0.00 ± 0.00

Table 6.2: Mean FID values ± SD for GAN and interpolated augmented datasets.

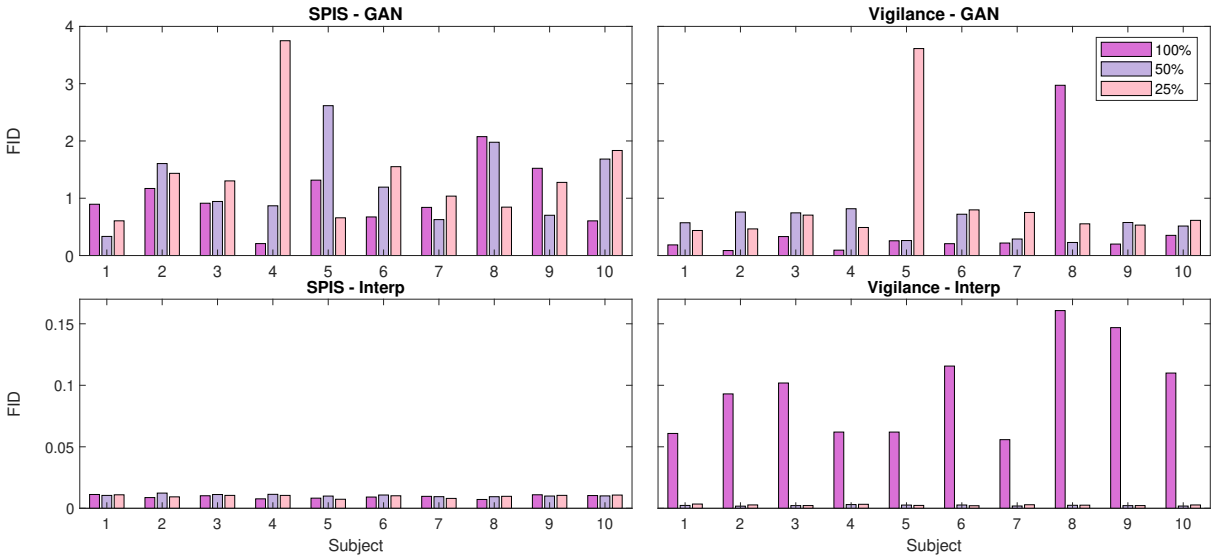


Figure 6.5: Subject-specific FID results for the two datasets, accounting for the forecaster trained on GAN data (above), and the one trained on interpolated data (below). The distance between real and synthetic data is estimated for different percentages of the chosen datasets. Full datasets (100%) are reported in purple, half datasets (50%) are reported in lavender, quarters of datasets are reported in light pink (25%).

6.8 Discussion

This study benchmarked the synthesis of EEG frequency band envelopes to replicate the statistical properties of physiological signals. By targeting the smooth power fluctuations of α, β, and θ bands rather than raw EEG, we filled a specific niche in the literature, addressing applications where time-frequency components are critical for cognitive state assessment.

6.8.1 Interpreting GAN Performance

The generator demonstrated robustness in replicating resting-state signals (SPIS) and vigilance degradation patterns (Vigilance). The FID analysis revealed a crucial insight into the nature of "useful" synthetic data. For the SPIS dataset, FID scores of ∼1.2 corresponded to embedding distributions that were comparable but not identical to real data. This slight distributional shift suggests that the GAN generated novel, informative variations of the signal rather than merely memorizing the training set. This diversity likely contributed to the robust forecasting improvements observed (∼29–37% reduction in MAE).

In the Vigilance dataset, results were more complex due to the multi-channel, multi-band nature of the task. For the 100% and 50% splits, the results mirrored the SPIS success cases (Fig. 6.6a). However,

data scarcity in the 25% split led to degradation. Specifically, Subject 5 presented a massive failure case (FID $>$ 3.0, performance decrement of -136%), characterized by jagged, non-overlapping embeddings (Fig. 6.6b). This failure is likely attributable to discriminator overfitting, where the discriminator distinguished real from fake samples too easily due to the limited training set, preventing the generator from receiving meaningful gradient updates [130].

6.8.2 Implications of Interpolation

A key finding of this work is the contrast between FID scores and functional utility for linearly interpolated data. Interpolation yielded extremely low FID scores ($\sim$0.01) and overlapping embedding distributions (Fig. 6.6c, d). While this indicates high statistical similarity to the real data, it resulted in a drastic deterioration of forecasting performance. We postulate that interpolated data resides too strictly within the convex hull of the existing data, offering no new information about the underlying data manifold. Consequently, the forecasting model treats these samples as redundant noise, failing to learn robust features. This highlights a limitation of FID when used in isolation: a very low score may indicate a lack of generative diversity rather than high quality.

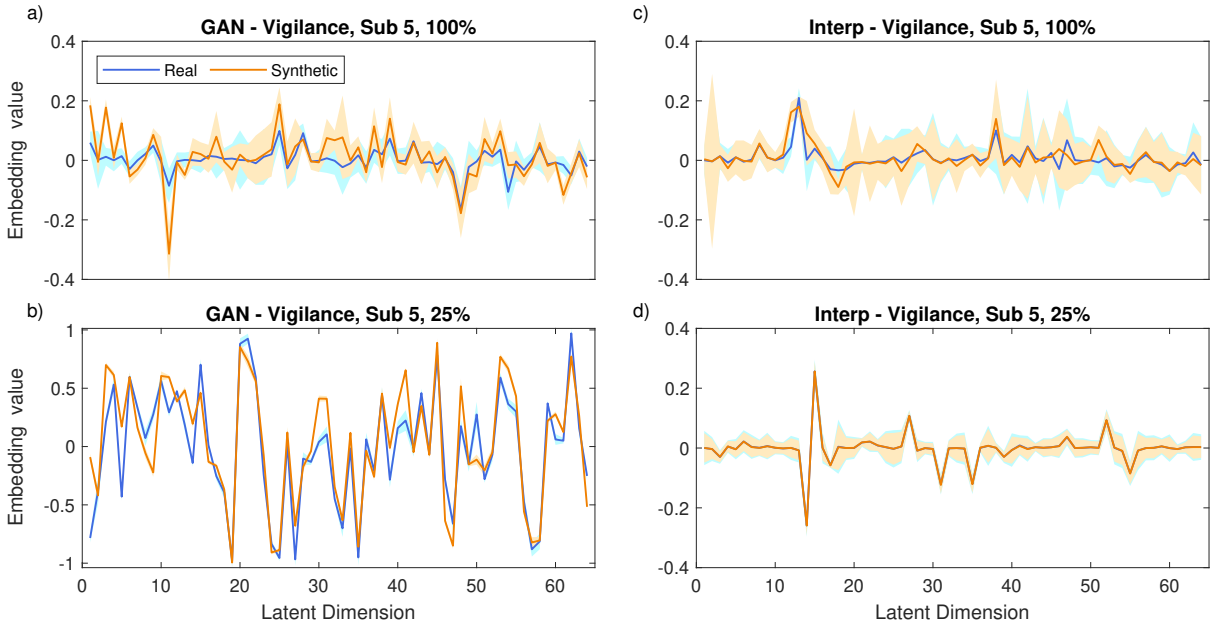


Figure 6.6: Embedding distributions for Subject 5 (Vigilance). Blue: Real, Orange: Synthetic. (a) Successful GAN generation (100% split) showing healthy overlap. (b) Failed GAN generation (25% split) showing mode collapse/divergence. (c-d) Interpolation results showing excessive overlap (redundancy), correlating with poor forecasting performance.

6.8.3 Conclusion

In conclusion, we demonstrated that uniformly low—but not near-zero—FID values are indicative of high-quality synthetic data that improves downstream model performance. Outlier FID values successfully flagged faulty generative processes (e.g., Subject 5), while near-zero values flagged redundant, unhelpful data (interpolation). The proposed GAN framework provides a theoretically grounded solution for simulating EEG dynamics, with significant implications for safety-critical environments like air traffic control [131] or educational monitoring [132], where collecting large-scale pathological or fatigue-related data is resource-intensive.

Chapter 7

Geometry-based Regularization of GANs for EEG Signal Generation

7.1 The EEG Covariance Matrix Manifold

7.1.1 Riemannian Representation of EEG Covariance Matrices

Differential Geometry is the branch of mathematics that employs the tools of differential calculus and multilinear algebra to study geometric problems. Originally formulated for the study of curves and surfaces in three-dimensional Euclidean space ($\mathbb{R}^3$), it evolved to investigate more abstract, multi-dimensional objects known as differentiable manifolds.

An n-dimensional manifold M is a topological space that is locally homeomorphic to the Euclidean space $\mathbb{R}^n$. This means that for any point $p \in M$, there exists an open neighborhood of p that is homeomorphic to an open subset of $\mathbb{R}^n$. This property allows us to apply the familiar concepts of calculus locally on the manifold, by collecting these local Euclidean coordinate systems, or "charts," into an "atlas" for the manifold. If the transition maps between overlapping charts are smooth (C^∞), the manifold is called a **smooth manifold**.

Moreover, for each point $p \in M$, we can associate a vector space called the **tangent space**, denoted T_pM. Intuitively, the tangent space at p consists of all possible velocity vectors of curves passing through p. More formally, a tangent vector can be defined as a derivation on the algebra of smooth real-valued functions defined in a neighborhood of p, providing a linear approximation of the manifold around that point. The disjoint union of all tangent spaces over all points of the manifold forms the tangent bundle TM.

A *Riemannian manifold* (M, g) is a smooth manifold equipped with a Riemannian metric g, defined as a smooth family of positive-definite inner products $g_p : T_pM \times T_pM \to \mathbb{R}$ on the tangent space T_pM at each point $p \in M$. This metric tensor g is necessary to quantify local geometric quantities—such as arc length, angles, and volumes—in a manner invariant under coordinate transformations. This enables the definition of geodesics, which generalize the notion of straight lines as locally distance-minimizing paths.

This chapter places a central emphasis on computing distances within a manifold framework. Because windowed EEG signals are mapped to structured representations that reside on non-Euclidean manifolds, relying on standard Euclidean metrics often leads to distorted comparisons that ignore the underlying geometry. By defining geodesic distances that account for the intrinsic properties of the representation space, a robust Riemannian comparative framework can be established. This approach has demonstrated superior performance across diverse BCI classification domains, including motor imagery, speech processing, and emotion recognition [133, 134, 135]. The main contribution of this work is to extent Riemannian covariance-based analysis to generative modeling, with the goal of developing a method to both estimate and improve synthetic EEG sample quality.

7.1.2 The SPD Covariance Matrix Representation of EEG

The study of Symmetric Positive Definite (SPD) covariance matrices has yielded excellent results in the field of EEG analysis, establishing itself as a powerful representational tool in both Machine Learning (ML) and Deep Learning (DL) methodologies [136]. These matrices serve as robust descriptors of EEG spatial patterns, constraining them to a Riemannian manifold rather than a standard Euclidean vector space.

Covariance Extraction and Preprocessing

Windowed multi-channel EEG data can be represented as a matrix $X \in \mathbb{R}^{N \times T}$, where N denotes the number of channels and T represents the number of time samples. Short temporal windows allow EEG signals to be approximated as pseudo-stationary. Non-invasive EEG data typically range from a couple to 128 channels, depending on the acquisition system and the brain area of interest; the sample frequency is often capped at 256 Hz, so a 2-second window contains 512 samples.

For the Sample Covariance Matrix (SCM) to act as a valid estimator of the population covariance, strict preprocessing steps are required. The standard computation of the SCM assumes the data is mean centered. Therefore, the raw windowed data X must first undergo linear detrending to remove DC offsets and low-frequency drifts that typically contaminate EEG recordings. Letting $\bar{X}$ represent the centered data where the mean of each row is zero, the sample covariance matrix C is computed as:

$$C = \frac{1}{T-1} \bar{X} \bar{X}^T \tag{7.1}$$

This formulation relies on the assumption of weak stationarity: it is implied that the first two moments of the EEG signal, mean and variance, remain constant within the chosen time window. While EEG signals are inherently non-stationary over long durations, they can be approximated as stationary within short epochs, typically between 0.5 and 2 seconds.

Spectral Interpretation via Wiener-Khinchin

The physical meaning of the covariance matrix elements can be understood through the Wiener-Khinchin theorem, which establishes that the Power Spectral Density (PSD) of a wide-sense stationary random process is the Fourier Transform of its autocorrelation function, $R(\tau)$.

The autocorrelation function measures the similarity between a signal $x(t)$ and a delayed version of itself, $x(t-\tau)$. The sample covariance matrix C captures instantaneous correlations, which corresponds to the autocorrelation evaluated at zero-lag ($\tau = 0$). Mathematically, the inverse relationship is given by:

$$R_{ii}(\tau) = \int_{-\infty}^{+\infty} S_{ii}(f) e^{j2\pi f \tau} \, df \tag{7.2}$$

where $S_{ii}(f)$ is the PSD of channel i. By setting $\tau = 0$, the complex exponential term becomes unity ($e^0 = 1$). Consequently, the diagonal elements C_{ii} of the covariance matrix represent the signal variance, and are equivalent to the integral of the PSD over the entire frequency range:

$$C_{ii} = R_{ii}(0) = \int_{-\infty}^{+\infty} S_{ii}(f) \, df \tag{7.3}$$

This implies that the diagonal terms represent the total energy of the signal across all frequencies.

Generalizing this to the off-diagonal elements C_{ij}, these terms represent the cross-covariance at zero lag. In the frequency domain, this corresponds to the integration of the Cross-Power Spectral Density $S_{ij}(f)$:

$$C_{ij} = \int_{-\infty}^{+\infty} S_{ij}(f) \, df \tag{7.4}$$

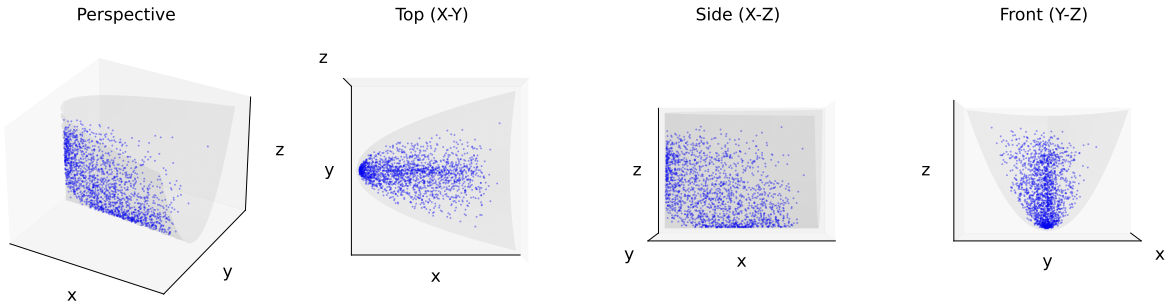


Figure 7.1: Four views of the SPD 3D cone. Each light blue point represents the x, y, z values of the upper diagonal of a 2-by-2 positive-definite matrix. The light grey shape represents the SPD cone surface.

The SPD matrix C physically captures the total synchronized energy shared between channels. Intuitively, it aggregates the functional connectivity between brain regions, specifically filtering for interactions that occur with zero phase lag (or 180-degree anti-phase). This sensitivity to instantaneous synchronization makes SPD matrices effective descriptors for paradigms such as Motor Imagery (MI), which relies heavily on identifying spatial patterns of power synchronization and desynchronization.

7.1.3 The Covariance Matrix Manifold

The set of all SPD matrices forms a differentiable manifold with a specific curved geometry. Given a channel dimension N, the set of all $N \times N$ EEG covariance matrices is a convex cone within the vector space of all $N \times N$ symmetric matrices, denoted by $\mathrm{Sym}_N^+(\mathbb{R})$, and it forms a differentiable manifold of dimension N(N+1)/2. Specifically, positive-definite matrices form the interior of the cone, composing an open manifold, while matrices with zero determinant lie on the cone surface. A visualization of 2-by-2 covariance matrices can be seen in Fig. 7.1.

Affine-Invariant Riemannian Metric (AIRM)

Although the hypercone represents the extrinsic shape of the manifold, we need an intrinsic notion of distance on it. By equipping this space with an appropriate metric, geometric operations that respect the space curvature are made possible; this implies the definition of an inner product on the tangent space at each point $C \in \mathrm{Sym}_N^+(\mathbb{R})$, denoted $T_C \mathrm{Sym}_N^+$. An effective choice for BCI applications is the Affine-Invariant Riemannian Metric (AIRM). For any two tangent vectors $S_1, S_2 \in T_C \mathrm{Sym}_N^+$, the AIRM defines the inner product as:

$$\langle S_1, S_2 \rangle_C = \mathrm{tr}(C^{-1} S_1 C^{-1} S_2)$$

and thus induces a notion of distance on the manifold. The shortest path between two points (matrices) C_1 and C_2 on the manifold is the geodesic, the curve of minimum distance that follows the curvature of the manifold. The unique geodesic $\gamma(t) : [0, 1] \to \mathrm{Sym}_N^+(\mathbb{R})$ connecting C_1 and C_2 is given by:

$$\gamma(t) = C_1^{1/2} \exp\left(t \cdot \log\left(C_1^{-1/2} C_2 C_1^{-1/2}\right)\right) C_1^{1/2}$$

where $\exp(\cdot)$ and $\log(\cdot)$ are the matrix exponential and logarithm, respectively.

The length of this path is the **geodesic distance**, which provides a principled measure of dissimilarity between two covariance matrices. It is calculated as:

$$d_R(C_1, C_2) = \left\| \log\left(C_1^{-1/2} C_2 C_1^{-1/2}\right) \right\|_F \tag{7.5}$$

where $\| \cdot \|_F$ is the Frobenius norm.

A useful reformulation of Eq. 7.5 is the following:

$$d_{\mathrm{R}}(C_1, C_2) = \left(\sum_{i=1}^{n} \ln(\lambda_i)^2 \right)^{1/2} \tag{7.6}$$

with λ_i being the standard eigenvalues of the matrix $C_1^{-1} C_2$.

To better understand the geometrical meaning of this norm, look again at Fig. 7.1, specifically at the Perspective and Side views. It is clear that the point distributions tend to become denser as they approach the cone exterior. In this region, matrices will have at least one eigenvalue λ that tends to zero, since it is approaching the zero-determinant surface. As such, the AIRM distance between a fixed point inside the cone and a matrix near the surface will explode due to $ln(\lambda)^2$. Conversely, the distance between points in the more rarefied central cone region will remain stable.

The AIRM has the particularly desirable property of being affine invariant. If the original EEG data was subjected to an invertible affine transformation A, two generic covariance matrices would transform as $C_1 \rightarrow AC_1A^T$, $C_2 \rightarrow AC_2A^T$; by affine transformation invariance, the distance between two such transformed matrices remains identical to the distance between the original matrices. This provides robustness against perturbations in the EEG recording setup, such as changes in the reference electrode or in the electrode montage [137], and has in fact been proven to be a reliable invariant to covariate shift [138].

7.1.4 Other Metrics on the ACM Manifold

Although the AIRM is widely considered the gold standard due to its theoretical soundness, other metrics emerge from the literature offering distinct computational advantages [139]. These are briefly reviewed in this section for completeness. While they offer improvements in processing speed, the AIRM's inherent geometric rigor makes it the optimal baseline for our evaluations.

Log-Euclidean Metric (LEM)

The Log-Euclidean metric is the most popular alternative to AIRM. It maps SPD matrices to the tangent space at the identity via the matrix logarithm, performs Euclidean operations in that vector space, and maps back via the exponential.

$$d_{\mathrm{LEM}}(A, B) = \| \log(A) - \log(B) \|_F \tag{7.7}$$

It creates a vector space structure for SPD matrices. Unlike AIRM, the geometric mean of a set of matrices can be computed explicitly in a single step ($\exp(\frac{1}{N} \sum \log A_i)$), making it computationally superior for clustering and averaging tasks.

Cholesky Distance

This metric relies on the Cholesky decomposition ($A = L_A L_A^T$), where L is a lower triangular matrix.

$$d_{\mathrm{Chol}}(A, B) = \| L_A - L_B \|_F \tag{7.8}$$

Despite being fast to compute, it is not uniquely defined (due to the sign ambiguity of diagonal elements in the decomposition) and lacks the full affine invariance property, meaning it is dependent on the coordinate system.

Stein Divergence (S-Divergence or Jensen-Bregman LogDet)

Even if it does not satisfy the triangle inequality, S-Divergence acts as an effective proxy for the Riemannian distance. It is derived from the Kullback-Leibler divergence between two multivariate Gaussian

distributions with zero means and covariances A and B.

$$d_{\text{Stein}}(A, B) = \ln \det \left(\frac{A+B}{2} \right) - \frac{1}{2} \ln \det(AB) \tag{7.9}$$

It is cheaper to compute than AIRM, as it avoids matrix diagonalization and logarithms, while retaining desirable properties, namely invariance under inversion and affine transformations.

Wasserstein Distance (Bures Metric)

Originating from optimal transport theory, this metric measures the minimum cost of transporting the probability mass of one Gaussian distribution to another.

$$d_{\text{Wass}}(A, B) = \sqrt{\text{tr}(A) + \text{tr}(B) - 2\text{tr}((A^{1/2}BA^{1/2})^{1/2})} \tag{7.10}$$

Unlike AIRM, this metric is defined on the entire cone, including the boundary. It allows distances to be computed even for singular (rank-deficient) matrices, making it robust in scenarios where $T < N$.

7.1.5 SPD Framework in the Literature

SPD space analysis has been applied in both supervised and unsupervised transfer learning. Authors of [140] use a set of geometrical transformations based on Procrustes Analysis to map inter-session SPD matrices into the same high-dimensional distribution. An alignment strategy is proposed in [141], called Euclidean Alignment, which maps SPD matrices to the Riemannian tangent space and returns a common, zero-mean distributional representation, showing promising results in both motor-imagery and event-related potential classification. Further considerations and extension related to Euclidean Alignment can be found in [142]. A similar approach is also presented for code-modulated Visual Evoked Potentials (c-VEPs) BCIs in [143], where optimal source subjects are selected after orienting data distributions using Log-Euclidean Data Alignment, significantly reducing calibration and improving accuracy compared to subject-specific baselines.

Concurrently, Deep Learning applications that leverage Riemannian representations flourished in the last years. A seminal work, introduced in 2016, was the SPDNet architecture [144], which presents a Riemannian non-linear network architecture for SPD matrix learning, along with layers dedicated to bilinear mapping, eigenvalue rectification, and eigenvalue logarithm. Examples stemming from this research include [145], where the proposed architecture replaces standard filterbank grid searches with end-to-end learnable filterbank layers optimized within the Riemannian manifold, or [146], where EEG-Net, a standard network for BCI, is combined with the SPD cross-covariance representation and bilinear transformations to greatly improve feature quality. A different approach, which leverages Graph Neural Networks, was proposed in [147], adopting manifold-valued graph convolutional techniques to extract spatio-temporal information for classifying 11 MI-EEG datasets.

7.1.6 The Augmented Covariance Matrix Representation

It is possible to extend the covariance matrix by incorporating time-lagged cross-correlations, thereby creating a richer spatio-temporal representation of EEG signals. The resulting structure, arranged as a block Toeplitz matrix, contains sub-matrices of cross-covariances, $C(\tau)$, for a range of time lags τ. The matrix extended in this way remains SPD, provided that an appropriate number of data points is chosen to avoid the matrix becoming singular. Moreover, no strictly linear relationship must be present between channels, a condition often irrelevant due to the noisy nature of the EEG.

For a maximum lag L, the resulting augmented covariance matrix (ACM) would then be expressed as:

$$\mathcal{C} = \begin{pmatrix} C(0) & C(-1) & C(-2) & \cdots & C(-L) \\ C(1) & C(0) & C(-1) & \cdots & C(-L+1) \\ C(2) & C(1) & C(0) & \cdots & C(-L+2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ C(L) & C(L-1) & C(L-2) & \cdots & C(0) \end{pmatrix}$$

The main diagonal of blocks consists of the standard, zero-lag covariance matrix $C(0)$, while off-diagonal blocks capture asynchronous relationships between channels; these lagged components provide a more detailed time-domain representation of the cross-spectrum. If $C(0)$ alone only captures information equivalent to the integral of the cross-spectrum, losing all its phase data, the inclusion of lagged matrices $C(\tau)$ allows for the recovery of the phase relationships from the frequency domain.

More specifically, computing the cross-covariance at a specific time lag τ is mathematically equivalent to performing an Inverse Fourier Transform of the entire cross-spectrum, weighted by a phase rotation term $e^{j2\pi f\tau}$. Consequently, a single lag value does not isolate a specific frequency component but rather extracts the aggregate constructive interference across the entire signal bandwidth, testing for a specific rigid temporal shift between the regions. While the standard covariance matrix (C_{ij}) captures only the instantaneous ($\tau = 0$) synchronization, integrating the real zero-phase component of the spectrum and analyzing non-zero lags allows for the reconstruction of the directionality inherent in neural communication.

ACM is rich in information and has, in fact, caught the interest of the BCI community, with works that formulate implementations in dedicated ML pipelines. The works of I. Carrara stands out: in [148], in which authors study the ACM properties from the optics of chaos theory and obtain top performance over the Mother of all BCI Benchmarks (MOABB), while in [149], authors propose a phase-augmented SPDNet which operates in the SPD space of ACMs, outperforming state-of-the-art DL architecture in MI decoding; a precise geometrical tractation is instead given in [150], where a classification pipeline is built based on the the Siegel disk manifold formed by ACMs.

Downsampling the ACM

The quadratic nature of the ACM formulation, although theoretically correct, suffers from implementation issues: the obtained size, $(N(L+1)) \times (N(L+1))$, quickly becomes prohibitive for extended lag numbers. A tractable representation can be obtained by reducing the total L, and by downsampling the lag intervals. This comes at a cost: the incomplete and potentially distorted estimation of the cross-spectrum. A first cause lies in the potential aliasing introduced by the interval between the selected lags, $\Delta\tau$, which restricts the observable frequency range to $[-\frac{1}{2\Delta\tau}, \frac{1}{2\Delta\tau}]$. Existing spectral content in the true cross-spectrum existing outside this range would be aliased, making it impossible to distinguish low-frequency connectivity from aliased high-frequency components. Secondly, truncating the total range of the selected lags L is equivalent to applying a rectangular window to the true cross-correlation function. This windowing in the temporal domain corresponds to a convolution in the frequency domain between the true cross-spectrum and a sinc function, having its main lobe width inversely proportional to L. Consequently, a smaller lag range produces a wider convolving function, causing greater spectral leakage.

7.2 Geometry-guided GAN Training for EEG Generation

7.2.1 Challenges in EEG-GAN Training

The primary challenge in generating novel signals within the GAN training paradigm is avoiding mode collapse. Mode collapse manifests as reduced diversity in the generated samples, producing redundant outputs by the Generator. This is primarily caused by the Generator's lazy behavior or, in more scientific terms, by the Generator reaching a global minimum in its optimization process from which it is not

incentivised to move further. There are multiple reasons behind this training failure; the two most notable ones are due to an imbalance between the Generator's and Discriminator's learning rates, and an overspecialization of the Generator. The first occurs when either the Generator or the Discriminator completely outmatches its opponents, nullifying its ability to properly optimize its parameters due to the cross-entropy gradient having an insignificant magnitude. At the cost of further muddying the waters between AI and neuroscience, we metaphorically equate this phenomenon to the inhibition in human learning that occurs under excessive cortisol production. This imbalance can be mitigated by adjusting the models' learning parameters properly. The second may occur even if both GAN networks behave properly, as it implies that the Generator has learned that, by replicating a small specific zone of the total data manifold, it can always confuse its adversary. This circumstance is trickier to account for, since it is not caused by any specific architectural failure - it is just the Generator finding a comfortable local minimum. In the upcoming sections, we propose a methodology to enhance the robustness of GAN training for time series by leveraging the geometric properties of the covariance matrix.

7.2.2 Training Stabilization Strategies

Classifying smooth, band-limited temporal signals (as in EEG) is a much simpler task than building a function that transforms pure noise into realistic, credible waveforms. Moreover, the literature is rich in architectures widely acknowledged for their pattern recognition capabilities, and even a simple CNN can distinguish between noisy or perturbed data and clean, realistic data [151, 152]. This is not a problem per se - we want our discriminator to be as capable as possible. However, we noticed that even after a few training epochs, the discriminator had totally overshadowed its opponent, leaving it to scramble through noisy batches in the hope of overcoming an insurmountable wall. Through a coarse yet effective dynamical training procedure, we restored a proper balance between the two parties. Specifically, we reduced the Discriminator learning rate whenever its final output on a batch of fake data fell below a user-set threshold, set to 0.15 in all subsequent experiments. Notice that the Discriminator output was compared before taking the cross-entropy loss, so the 0.15 directly corresponds to a more than 85% accurate classification. The Discriminator learning rate was reduced by an order of magnitude and was restored to its original value whenever its performance remained above threshold. On a final note, the Discriminator updates its weights while considering both real and synthetic data to avoid gradient oscillations between the two steps. Slowing the update procedure, rather than simply skipping it, was essential to avoid skipping entire batches of real data. This entire procedure was combined with label smoothing [153] - real label values were set to 0.95 instead of 1. Label replaces binary targets with softer values, such as 0.1 and 0.9. In the context of GANs, this prevents the Discriminator from becoming overconfident in its classifications, a state that often leads to vanishing gradients and the cessation of Generator learning, by forcing the Discriminator to produce more conservative probability estimates. This smooths the loss landscape and mitigates the risk of training instabilities, such as oscillations or premature mode collapse.

7.2.3 Assessing Generative Variety in the SPD Manifold

Even if the Generator could push through and deliver properly looking samples, no specific, golden-standard method exists to check their distribution, leaving an underlying uncertainty about the model's generalization capabilities. By exploiting the properties of the augmented covariance matrix (ACM), it is possible to both visualize and numerically quantify the spread of synthetic data.

Visualization Through UMAP

To gain an intuitive understanding of the data distribution, we propose encoding EEG data in an N-dimensional covariance manifold and subsequently mapping it to 2D or 3D using an embedding approach, such as Uniform Manifold Approximation and Projection (UMAP) [154]. UMAP works by attempting to preserve the distance between points, which we have thanks to the AIRM, even in low-dimensional representations. More specifically, the UMAP algorithm relies on three core assumptions, all satisfied in

the SPD use-case: the data is uniformly distributed on a Riemannian manifold, the Riemannian metric is locally constant (or can be approximated as such), and the manifold is locally connected.

In the high-dimensional ambient space, UMAP constructs a weighted k-nearest neighbor graph to approximate the manifold's topology. It defines a fuzzy simplicial set where the edge weight $p_{ij|i}$ between a point x_i and its neighbor x_j decays exponentially according to the metric $d(x_i, x_j)$, adjusted by a local distance threshold ρ_i (the distance to the nearest neighbor) and a scaling parameter σ_i [155]. Mathematically, this local fuzzy connectivity is expressed as:

$$p_{ij|i} = \exp\left(-\frac{d(x_i, x_j) - \rho_i}{\sigma_i}\right) \tag{7.11}$$

representing the probability that the 1-simplex (the edge between a pair of nodes i and j) actually exists in the underlying topological manifold. For each pair, the resulting fuzzy union is computed as

$$p_{ij} = p_{ij|i} + p_{jij|j} - p_{ij|i}p_{ji|j} \tag{7.12}$$

These conditional probabilities are symmetrized to form a joint probability distribution P representing the global topological structure. In the target low-dimensional space, the algorithm models a similar graph structure Q, where the probabilities of edges q_{ij} are typically derived using a heavy-tailed distribution (such as the Student t-distribution). The final embedding coordinates are optimized by minimizing the fuzzy set cross-entropy between the high-dimensional representation P and the low-dimensional representation Q:

$$C(P, Q) = \sum_{i \neq j} \underbrace{p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right)}_{Attraction} + \underbrace{(1 - p_{ij}) \log\left(\frac{1 - p_{ij}}{1 - q_{ij}}\right)}_{Repulsion} \tag{7.13}$$

The attraction term minimizes distance between neighboring points where p_{ij} is high, while the repulsion term does the opposite. By minimizing this entire objective via stochastic gradient descent, UMAP balances the preservation of local neighborhood structures with the retention of the global data structure.

In this way, it becomes possible to visualize whether the Generator covers the real data space or not. Practically speaking, we selected only a subset of real waveforms (e.g., 2000 samples), generating an equal number of synthetic ones, and then projecting them onto two dimensions. Note that this implies computing the distance between the SPD representations of both real and synthetic data. A visual overlap between the two distributions implies that the Generator provides a proper coverage of the real data space; otherwise, having small clusters of synthetic data denotes an excessive overspecialization of the Generator.

Coarse Cluster Estimation

To complement the visual inspection of data distributions, we established a numerical indicator to quantify the Generator's total mode coverage.

First, the real data points are partitioned using K-means clustering on the SPD manifold, implemented via *pyriemann* [156]. While K-means is computationally efficient, scaling linearly with the amount of data, clustering is better performed on a representative subset of the real data, due to the cost of the pyriemann functions. The algorithm also exhibits an implicit bias towards generating clusters of equal size, an effect of minimizing the sum of the variances of the clusters.

Subsequently, synthetic samples are classified by assigning them to the nearest real centroid. Using the resulting cluster membership counts for both real and synthetic datasets, the final step involves computing the Jensen-Shannon (JS) divergence, which provides a measure of distributional similarity between 0 and 1, and is apt when handling potentially empty bins.

7.2.4 Geometry-based Regularization of the Generator

As stated before, a regular GAN training procedure can improve sample quality by tweaking hyperparameters, but has no way to enforce sample variety. From hereafter, we delineate the key strategy of this thesis: augmenting the Generator loss function with a geometric regularizer to nudge the generator toward spreading out its outputs. This method is rooted in our knowledge of the SPD space, and is verified in Sec. 7.3.1 with the methods proposed throughout this thesis.

Covariance Matrix Based Generator Loss Function

The regular Generator loss function becomes:

$$\mathcal{L}_G = \underbrace{-\mathbb{E}_{z\sim p_z(z)}\left[\log D(G(z))\right]}_{\text{Standard Cross-Entropy}} - \underbrace{\text{AIRM}_{\text{min}}}_{\text{Spread Term}} \tag{7.14}$$

where

$$\text{AIRM}_{\text{min}} = \min\left\{\text{AIRM}(z_i, z_j) : i, j \in \{1, \ldots, N\}, i \neq j\right\} \tag{7.15}$$

In which N is the total number of elements in a batch, z_i, z_j are elements belonging to that batch, and AIRM is the Affine-Invariant Riemannian Metric. In words, this formula computes all the pairwise distances between SPD data representations inside a batch of synthetic samples, takes the minimum distance, and subtracts it from the regular loss function. In this way, the Generator is incentivized to produce samples that are more spread out, meaning they are farther apart in the SPD space. The minimum is chosen over the mean value to avoid lazy behaviors - if the mean were used, the Generator could simply create a few extremely noisy waveforms to greatly improve the spread term value. An abstract representation of loss function computation is show in Fig. 7.2.

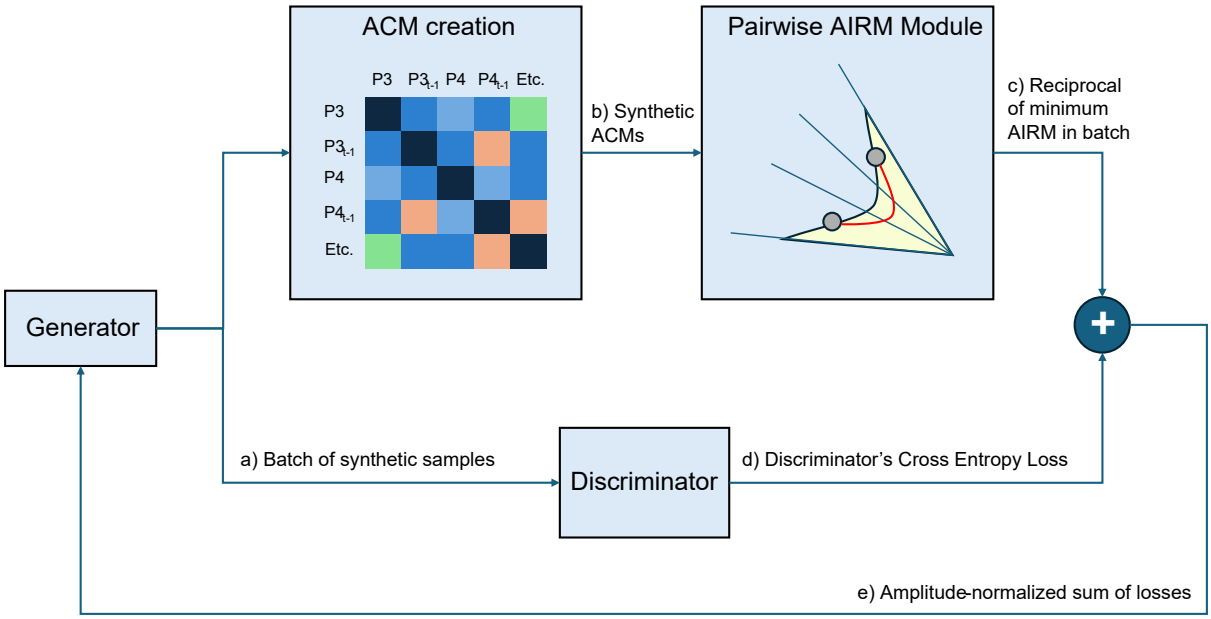


Figure 7.2: High-level explanation on how the geometrically regularized loss function is derived. a) A batch of synthetic samples is generated and fed to both the Discriminator and the ACM extractor module. b) ACM for each element in the batch is generated, pairwise distances can be computed. c) The reciprocal of the smallest AIRM distance in the batch of matrices is taken as the regularizing metric. d) Concurrently, the standard GAN loss function is computed. e) Addition between the two amplitude-normalized term is used as the final Generator's loss function.

Implementively speaking, the AIRM takes the following form:

```
1 import torch as tc
2
3 def torch_riemann_dist_wiki(A, B):
4
5     inv_product = tc.linalg.solve(B, A)
6     eigvals = tc.abs(tc.linalg.eigvals(inv_product))
7     return tc.sqrt(tc.sum(tc.square(tc.log(eigvals))))
```

The function is fully differentiable and efficiently implemented via SVD and eigenvalue decomposition, thereby avoiding explicit matrix inversion. Line 5 computes the matrix product $B^{-1}A$; it avoids explicit matrix inversion, a computationally expensive and numerically unstable operation. `tc.linalg.solve()` instead relies on the robust approach of LU or Cholesky decomposition. Line 6 finds the eigenvalues of $B^{-1}A$; since A and B are SPD, their generalized eigenvalues are theoretically guaranteed to be real and positive. The general solver `tc.linalg.eigvals()` can sometimes produce negligible imaginary components, an issue easily circumvented by applying `tc.abs`. Line 7 fully implements Eq. 7.6. Squaring the logs, summing them, and taking the square root is the definition of the Frobenius Norm of a diagonal matrix. Since the Riemannian distance is the Frobenius norm of the matrix logarithm in the tangent space, this step completes the calculation. The function leverages the fact that the eigenvalues of $B^{-1}A$ are identical to the eigenvalues of $A^{-1/2}BA^{-1/2}$. The latter is the standard definition of the AIRM distance, but the former is much faster to compute because it avoids the matrix square root and inversion operations.

Merging Loss Functions

In multi-task learning, combining multiple loss functions, $\mathcal{L}_{total} = \sum \omega_i \mathcal{L}_i$, poses significant challenges, primarily due to gradient interference and imbalances in task magnitudes. To address these, we adapted gradient surgery and normalization techniques to the AIRM-augmented GAN, in order to fuse the two terms in the Generator loss function (see Eq. 7.14).

Projected Conflicting Gradients (PCGrad) mitigates destructive interference by identifying pairs of gradients $\mathbf{g}_i$ and $\mathbf{g}_j$ that exhibit a negative cosine similarity ($\mathbf{g}_i \cdot \mathbf{g}_j < 0$). When such a conflict occurs, PCGrad projects each gradient onto the normal plane of the other, ensuring that the updated direction does not adversely affect the progress of competing objectives. Mathematically, this can be written as $\mathbf{g}_i = \mathbf{g}_i - \frac{\mathbf{g}_i \cdot \mathbf{g}_j}{\|\mathbf{g}_j\|^2}\mathbf{g}_j$. Complementarily, GradNorm provides a dynamic weighting mechanism to balance the training rates of different tasks. It adjusts the scalar weights $\omega_i(t)$ at each training step by minimizing a secondary loss that encourages the weighted gradient norms $\|\omega_i(t)\mathbf{g}_i\|_2$ at some significant layer to converge toward a common scale, normalized by the relative inverse rate of convergence for each task. Together, these methods transform the optimization from a simple weighted sum into a robust equilibrium search, preventing dominant signals from overwhelming the shared representation while resolving geometric conflicts in the parameter space. However, GradNorm incurs a high computational cost; an easier approach to avoid size-related issues is to simply normalize each loss to 1 by dividing it by its detached magnitude.

7.3 Experimental Evaluation of Geometry-guided GAN Generation

7.3.1 Augmenting Long-Term α

The proposed method was applied to augment the small α-band dataset introduced in section 4.1. The dataset consists of approximately 50,000 overlapping time windows; however, when reduced to non-overlapping windows, it contains only 200 unique samples. This is an excellent case study to assess whether data generation can succeed when starting from such a small cohort of examples and whether

synthetic data actually improves model performance. Hereafter, results detailing data generation, quality inspection, and performance contribution are reported.

EEG Data Generation Architecture

The GAN architecture consists of a GRU-based Generator and a CNN-based Discriminator. The Generator is composed of (i) a noise-interpolating module, alternating linear and Leaky ReLU layers with a dropout rate of 0.2, (ii) a stack of 6 bidirectional GRUs with a latent dimension of size 32, and (iii) an output module consisting of two linear layers interleaved with ReLU activations. The noise-interpolating module takes as input a white-noise sequence of length 64 time points and outputs a $256*C$ vector, where $C = 4$ is the number of channels. The GRUs have a 64-dimensional memory vector and receive properly rearranged interpolated noise as input (256 x C). Before outputting the result, a parameter specifies whether to normalize the output between [0, 1].

The Discriminator consists of a stack of 6 Conv1D layers with a stride of 2, interleaved with Leaky ReLU having dropout set to 0.2. A final convolutional layer with a stride of 4 precedes the sigmoid output layer. The convolutions have an increasing number of filters, ranging from 128 to 2048. Both models were trained on a Nvidia 3070 Ti laptop with 8 GB of VRAM.

Training configuration

Normalized EEG data were generated using 30 training epochs. The Generator and Discriminator learning rates were set to 1e-4, and the Discriminator had a learning rate of 1e-5, respectively. The Discriminator's weight update routine slowed to a tenth whenever it correctly classified at least 85% of a synthetic data batch. In addition, label smoothing was used, with the true label values set to 0.95. The training was repeated for each subject, each time using the remaining nine as the reference distribution. Batches consisted of 32 datapoints. Finally, the ACM representation used to compute the AIRM involved taking 3 lags at 11-sample intervals. This choiche followed the examination of autocorrelation plots, which showed that the first minimum was reached at the 22nd sample.

Two configurations were assessed: a baseline using the standard GAN loss function, and the regularized loss function proposed in section 7.2.4.

Data Quality Inspection

As a first quality assessment, the temporal smoothness, rhythmicity, and structural regularity typical of α-frequency band EEG activity can be evaluated by direct visual inspection of representative time-domain signals.

The aim of the analysis is twofold: first, to understand if the Generator produces plausible-looking signals in comparison to ground truth ones (real EEG data); second, to see if the Generator exhibits laziness, creating repeatedly the same temporal α patterns.

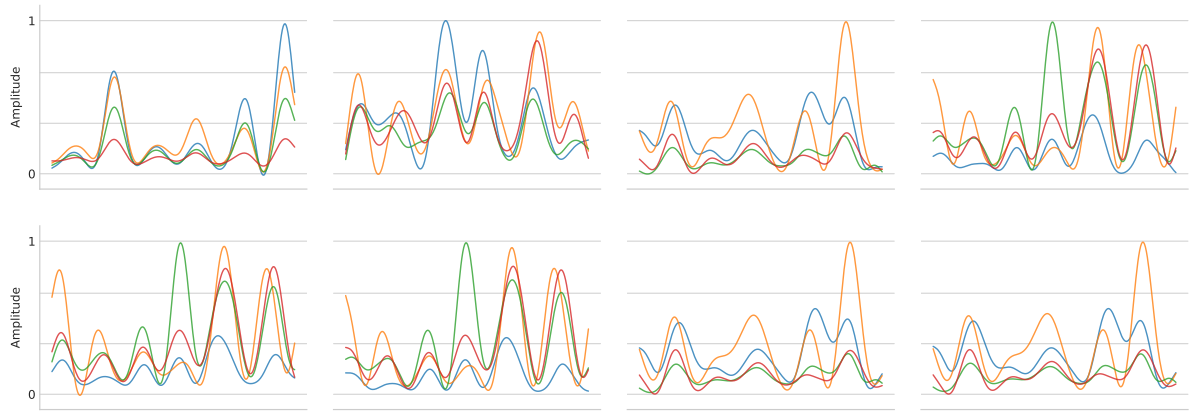
Representative real EEG windows are shown in Fig. 7.3a; they are characterized by a dominant oscillatory component consistent with α behavior and by complex relationships among channels' phases. In the real data, the relationship between channels changes from window to window - they can be either in phase or shifted. See Fig. 7.3a for an example. Such variabilty is expected in EEG recordings and should be reflected in the generated data as well.

Fig 7.3b shows examples produced by standard GAN. Although the oscillatory pattern seems to be generally preserved, several windows look identical to each other. It seems that the generator tends to re-use a limited set of solutions instead of exploring the full real data distribution.

Examples generated with the geometry-regularized GAN are reported in Fig 7.3c. Here, the oscillatory structure is present, along with much broader signal variety. Differences in phase, amplitude and temporal evolution can be observed.

Even if temporal smoothness and regularity can be confirmed for both the standard and the geometrically guided GAN, the former shows multiple repeated signals, as seen in Fig. 7.3b, a symptom of mode collapse not found in the latter - see Fig. 7.3c and notice the wider range of temporal behaviors.

(a) Example of eight real EEG data windows from subject 1. Each color represents one of the four channels (C3, P3, C4, P4). Data amplitudes are normalized to [0, 1]. Data is smooth and regular, and exhibits a dominant oscillation: six peaks are observed in each window. No clear relationship between channels can be found: they may or may not be in phase with each other, showing diverse behavior in every window.



(b) Example of eight synthetic EEG data windows generated for subject 1, using the standard GAN routine. Although a single signal may respect the expected smoothness and regularity, duplicates can still be found even in such a small sample, indicating some form of collapse.

(c) Example of eight synthetic EEG data windows generated for subject 1, using the regularized GAN routine. Data show diverse behavior among samples, denoting both phase-aligned peaks and trailing channels, while still preserving the expected temporal and spectral properties.

Figure 7.3

Latent Space Interpolation

To better compare the two GAN models, we study how the generated EEG signals change along with gradual shifts in the input noise. Each generated signal depends on a noise vector given as an input to the Generator. By fixing two noise vectors and interpolating them, it is possible to generate a sequence of noise samples that gradually move from one point to the other in the noise space. The interpolated noise vectors are then fed to the Generator and the corresponding EEG signals are observed. If the Generator has learned a good representation of the data, then the generated signals are expected to have a gradual change in waveform shape, phase, or amplitude. Differently, if the Generator maps large regions of the noise space to the same output, the generated signals will be unchanged for many consecutive steps of interpolation. This behavior manifests visually and provides another perspective for intuitively understanding model quality.

Practically, a 64-dimensional long white noise vector can be seen as a point lying on a 64-dimensional hypersphere; it is then possible to perform spherical linear interpolation (SLERP), following the arc connecting two points and sampling new noisy vectors along it. Feeding a properly trained Generator with these intermediate vectors should result in a somewhat smooth transition between the chosen generated points. This expected behavior is found for the geometrically regularized GAN: Fig. 7.4 shows the α-band signals generated by 16 noise interpolations. Notice, for example, how the peak in 3 slightly shifts in time, and how the orange peak in 15 is built progressively along samples. It is still worth noting how there is a jump between samples 2 and 3, a phenomenon potentially linked to the presence of sparse clusters in the generated data space. Sec. 7.3.1, 7.3.1 confirm this theory, and relates it to the original data distribution.

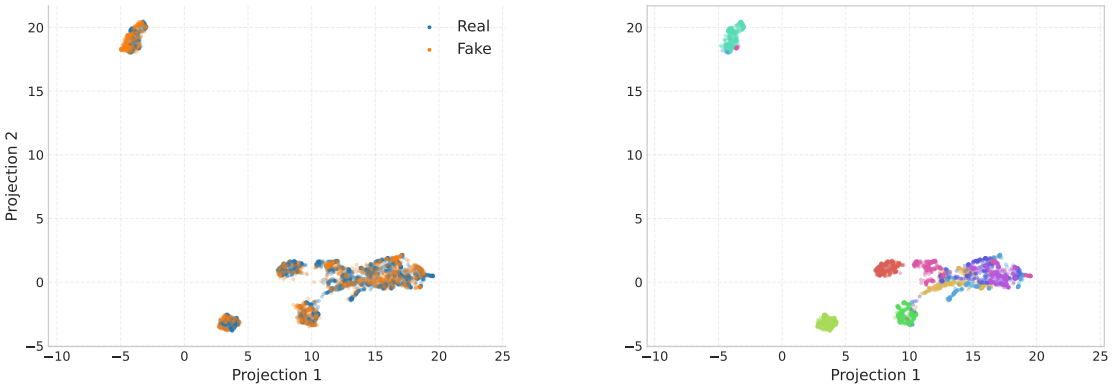
On the contrary, Fig. 7.5 highlights the poor capabilities of the vanilla Generator. Four distinct sections can be found: samples in (0-4), (5-12), (13-14), (15) repeat the same generated series with minimal variation, denoting an overspecialization of the Generator in specific noise space sections.

Figure 7.4: White noise interpolation for the geometrically regularized GAN Generator model. Transitions between the first (0) and last (15) samples are much smoother.

Figure 7.5: Example of white noise interpolation for the vanilla GAN Generator model. Severe jumps between contiguous noise samples may be noticed.

Embedding Visualization in the SPD Manifold

To complement the qualitative inspection, we analyze the distribution of real and synthetic data in the SPD manifold using a low-dimensional embedding. Specifically, a set of 2672 real EEG windows were selected and 1000 synthetic EEG samples were generated. From each window, ACM with 3 lags and a stride of 11 was extracred, and pair-wise distances were evaluated using the AIRM. UMAP was then used to create a joint 2-D representation of their distributions.
Fig. 7.6, 7.7 show the embeddings for the geometry-regularized GAN and the standard GAN, respectively, for a representative subject. In panel (a), real EEG samples are shown in blue and synthetic samples in orange. For the geometry-regularized GAN, the synthetic distribution overlaps the real one, indicating proper coverage and interpolation of the real data space. In contrast, the standard GAN produces sparse, small, and isolated clusters that occupy only parts of the real distribution, indicating zones of overspecialization.
Panel (b) shows the embedding of real data only. Clusters representing the nine left-out subjects are reported in different colors. The presence of distinct clusters indicates subject-specific structure in the SPD manifold and highlights inter-subject variability. At least three subjects (marked with different shades of green) tend to exhibit patterns that are easily distinguishable from those of the others. The two plots show different real data distributions. This can be attributed to the different synthetic inputs fed to UMAP: distinct graphs relating real and synthetic data are built for the two model cases.



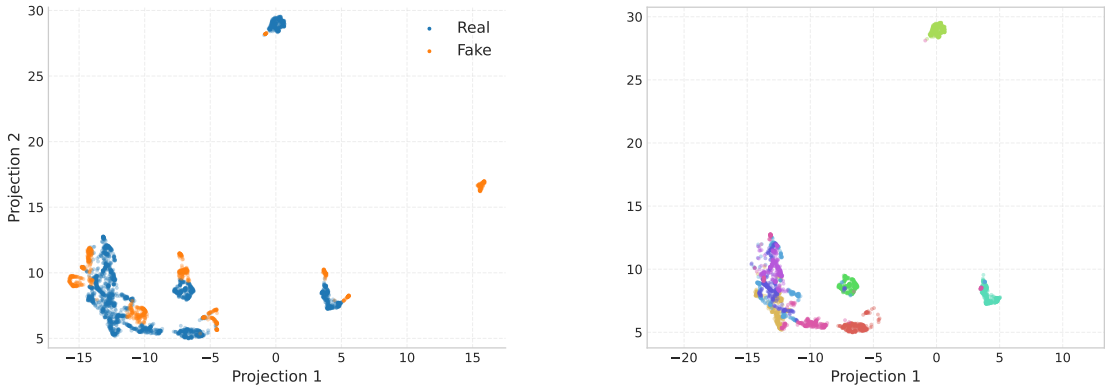
(a) (b)

Figure 7.6: (a) UMAP embeddings of overlapping real (blue) and regularized synthetic (orange) distributions. (b) UMAP embeddings of real data only; different color represent different subjects, 9 in total.

Unsupervised Riemannian Clustering

A more precise indicator of real data space coverage is obtained by leveraging unsupervised clustering in the ACM space. For each subject, augmented covariance matrices were computed from the remaining nine subjects using 3 time lags and a stride of 11. A total of 2672 real ACM samples were used to estimate cluster centroids using Riemannian K-means under the AIRM. Synthetic samples were not included during clustering, but 1000 fake samples were instead assigned to the nearest real-data centroid in the ACM space. The number of clusters was set to 20.

Table 7.1 reports the KL and JS divergence values computed between the real and synthetic cluster distributions with the 10 subject folds, for both the vanilla GAN and the geometry-regularized GAN. For all folds, the geometry-regularized GAN consistently yields a cluster distribution much closer to that



(a) (b)

Figure 7.7: (a) UMAP embeddings of overlapping real (blue) and vanilla synthetic (orange) distributions. (b) UMAP embeddings of real data only; different colors represent different subjects.

Regularized	Subject Folds									
KL	0.091	0.083	0.100	0.105	0.130	0.048	0.056	0.127	0.041	0.066
JS	0.147	0.148	0.155	0.163	0.177	0.111	0.118	0.177	0.100	0.127

Vanilla	Subject Folds									
KL	2.555	1.570	2.178	2.577	2.781	3.596	2.820	3.116	2.822	3.146
JS	0.611	0.517	0.581	0.631	0.649	0.760	0.664	0.701	0.667	0.696

Table 7.1: KL and JS divergence values reported along the distributions taken from the 10 data folds, for both the vanilla GAN method and the geometrically regularized one.

of real data than the vanilla GAN. This can be seen in Tab. 7.1: for all of the 10 subjects, lower values in both smoothed KL divergence and Jean-Shannon divergence can be confirmed. The geometry-regularized GAN yields KL divergence values below 0.14 and JS divergence values below 0.18, whereas the vanilla GAN shows divergences that are several times larger (KL >1.5; JS>0.5) in all folds.

7.3.2 Forecasting Improvements

The final evaluation assesses whether the proposed EEG data augmentation strategy improves the forecasting performance. To this end, two PatchTST models were trained: one using only real data and the other using the augmented dataset generated by the geometry-regularized GAN.

Representative examples of the two different forecasts for a single subject can be seen in Figs. 7.8 and 7.9.

An increase in performance at critical forecast horizons is confirmed quantitatively in Tab. 7.2, where the model trained on augmented data exhibits lower average MAE and reduced variance. Channel-specific metrics at those horizons are shown in Tab. 7.3. A point-by-point comparison, involving all 44 forecast horizons, can be seen in Fig. 7.10. A closer inspection of Tab. 7.3 highlights where the contribution of data augmentation becomes important. At short prediction horizons (11), both models closely track the ground truth, as expected. This is reflected by the R^2 value being close to 1. Small yet consistent improvements are observed at the 22nd horizon, with an average difference of 0.0014 in MAE and 0.0169 in R^2 in favour of the augmented model, and a variance improvement of 0.0006 and 0.0138, respectively,

for MAE and R^2. As the forecast horizon increases, predictions become less and less accurate. In these settings, data augmentation seems to improve stability: the augmented model shows decisively better performance for channel C3, the worst-performing one in the vanilla case, with MAE reductions of 0.0032 and 0.0024 and R^2 improvements of 0.09 and 0.07 at the 33rd and 44th horizons, respectively. Variance behaves in the same way: a reduction of 0.0057, 0.0044 for MAE, and an increment of 0.1158, 0.0909 for R^2, at horizons 33, 44. Other channels show only very small improvements at the 33rd horizon and negligible differences at the 44th, with slightly better MAE but slightly worse R^2 metrics.

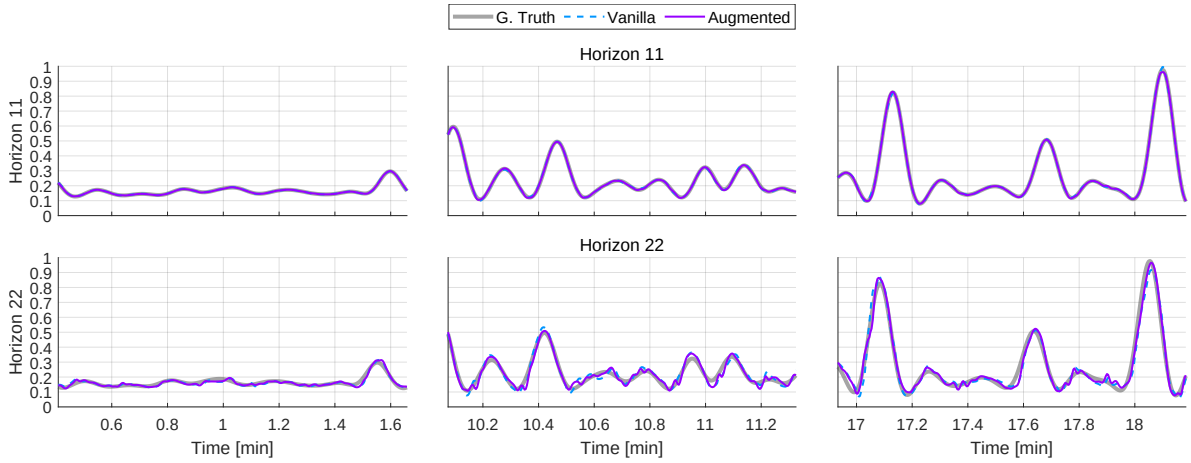


Figure 7.8: Ground truth (in grey) and PatchTST predictions obtained from models trained on real data only (light blue) and augmented data (purple) at the 11^{th} (2.75 s), and 22^{nd} (5.5 s) forecast horizons, for channel P3 of a representative subject (n. 4). Three windows from different time intervals of the task are reported (i.e. 0.4-21, 10-11.3, and 16.9-18.2 minutes).

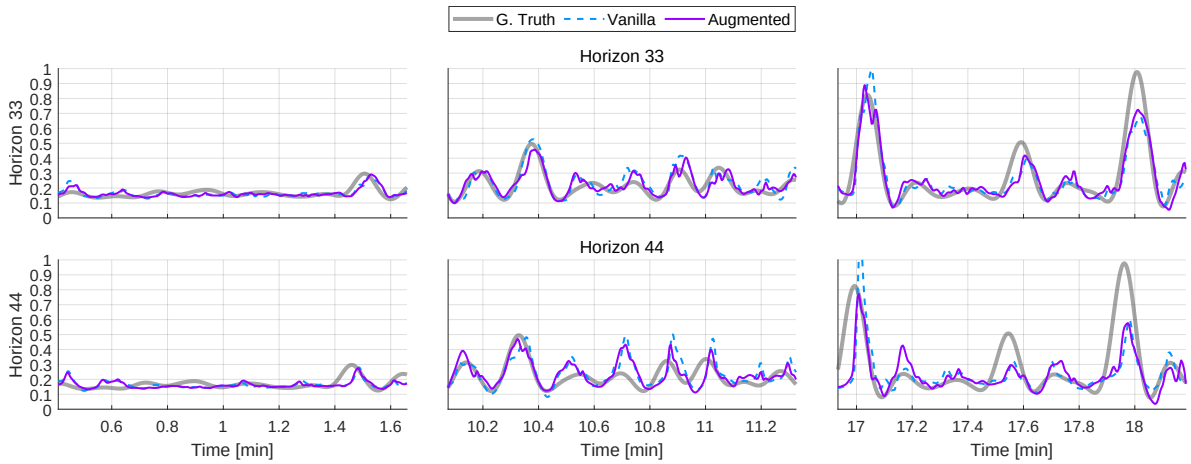


Figure 7.9: Ground truth (in grey) and PatchTST predictions obtained from models trained on real data only (light blue) and augmented data (purple) at the 33^{rd} (7.25 s), and 44^{th} (11 s) forecast horizons, for channel P3 of a representative subject (n. 4).

Horizon		11	22	33	44
MAE	Vanilla	0.0021 ± 0.0007	0.0162 ± 0.0055	0.0366 ± 0.0125	0.0459 ± 0.0138
	Augmented	**0.0016 ± 0.0005**	**0.0148 ± 0.0049**	**0.0350 ± 0.0118**	**0.0450 ± 0.0138**
R^2	Vanilla	0.9981 ± 0.0011	0.9013 ± 0.0430	0.4980 ± 0.2119	0.2097 ± 0.1989
	Augmented	**0.9991 ± 0.0003**	**0.9182 ± 0.0298**	**0.5370 ± 0.1704**	**0.2300 ± 0.1961**

Table 7.2: MAE and R^2 results at four forecast representative forecast horizons, averaged across all subjects and channels, for the two models: the transformer trained on the original data (Vanilla), and the one trained on the augmented dataset (Augmented).

Ch.	**Metric**	**Model**	**Prediction Horizon**			
			11	**22**	**33**	**44**
C3	MAE	Vanilla	0.0027 ± 0.0026	0.0195 ± 0.0165	0.0444 ± 0.0372	0.0547 ± 0.0431
		Augmented	0.0019 ± 0.0016	0.0172 ± 0.0134	0.0412 ± 0.0315	0.0523 ± 0.0387
	R^2	Vanilla	0.9978 ± 0.0021	0.8841 ± 0.0667	0.4004 ± 0.3246	0.1012 ± 0.3486
		Augmented	0.9990 ± 0.0006	0.9115 ± 0.0382	0.4909 ± 0.2078	0.1715 ± 0.2577
C4	MAE	Vanilla	0.0025 ± 0.0015	0.0193 ± 0.0096	0.0438 ± 0.0218	0.0542 ± 0.0264
		Augmented	0.0018 ± 0.0010	0.0176 ± 0.0087	0.0421 ± 0.0213	0.0537 ± 0.0266
	R^2	Vanilla	0.9984 ± 0.0007	0.8959 ± 0.0357	0.4622 ± 0.1814	0.1709 ± 0.1719
		Augmented	0.9991 ± 0.0003	0.9133 ± 0.0277	0.5007 ± 0.1725	0.1790 ± 0.2069
P3	MAE	Vanilla	0.0015 ± 0.0012	0.0120 ± 0.0078	0.0270 ± 0.0177	0.0344 ± 0.0232
		Augmented	0.0012 ± 0.0009	0.0112 ± 0.0068	0.0261 ± 0.0152	0.0340 ± 0.0209
	R^2	Vanilla	0.9988 ± 0.0005	0.9223 ± 0.0405	0.5983 ± 0.2006	0.3097 ± 0.1715
		Augmented	0.9992 ± 0.0003	0.9321 ± 0.0311	0.6078 ± 0.1844	0.3095 ± 0.1960
P4	MAE	Vanilla	0.0019 ± 0.0011	0.0141 ± 0.0066	0.0311 ± 0.0149	0.0402 ± 0.0189
		Augmented	0.0014 ± 0.0007	0.0131 ± 0.0063	0.0304 ± 0.0151	0.0399 ± 0.0188
	R^2	Vanilla	0.9976 ± 0.0027	0.9029 ± 0.0533	0.5312 ± 0.2753	0.2570 ± 0.2287
		Augmented	0.9990 ± 0.0005	0.9159 ± 0.0447	0.5485 ± 0.2604	0.2599 ± 0.2570

Table 7.3: Channel-wise forecasting performance comparison between Vanilla and Augmented transformers across different prediction horizons. Results are reported as Mean ± Standard Deviation.

7.4 Discussion

EEG data processing suffers from a lack of high-quality, extensive, and homogeneous datasets. Differences in acquisition protocols, from recording equipment to the definition of the task itself, make for a fragmented landscape of domain-specific and often incomparable records. The need for generative techniques that can interpolate existing data and produce entirely new, physiologically plausible EEG samples cannot be understated.

Literature certainly does not lack examples of varying degrees of sophistication. From simple signal splicing to yet-to-be-fully-understood genAI approaches, multiple works paved the way since the early 2010s [101]. Inside this context, we provided a way to significantly improve the training stability of supervised genAI methods, specifically by applying and testing it using GANs. Although GANs have been widely used in the BCI community since 2018 [157, 104], few advances have been made in their training protocols, with models or refinements drawn from other fields [158, 158, 159]. A well-known consequence is mode collapse, in which the generator overspecializes on a small subset of the data, ignoring the true underlying distribution and producing realistic but all-too-similar outputs. This failure is intrinsic to the gradient descent approach; why should the generator move out of a local minimum

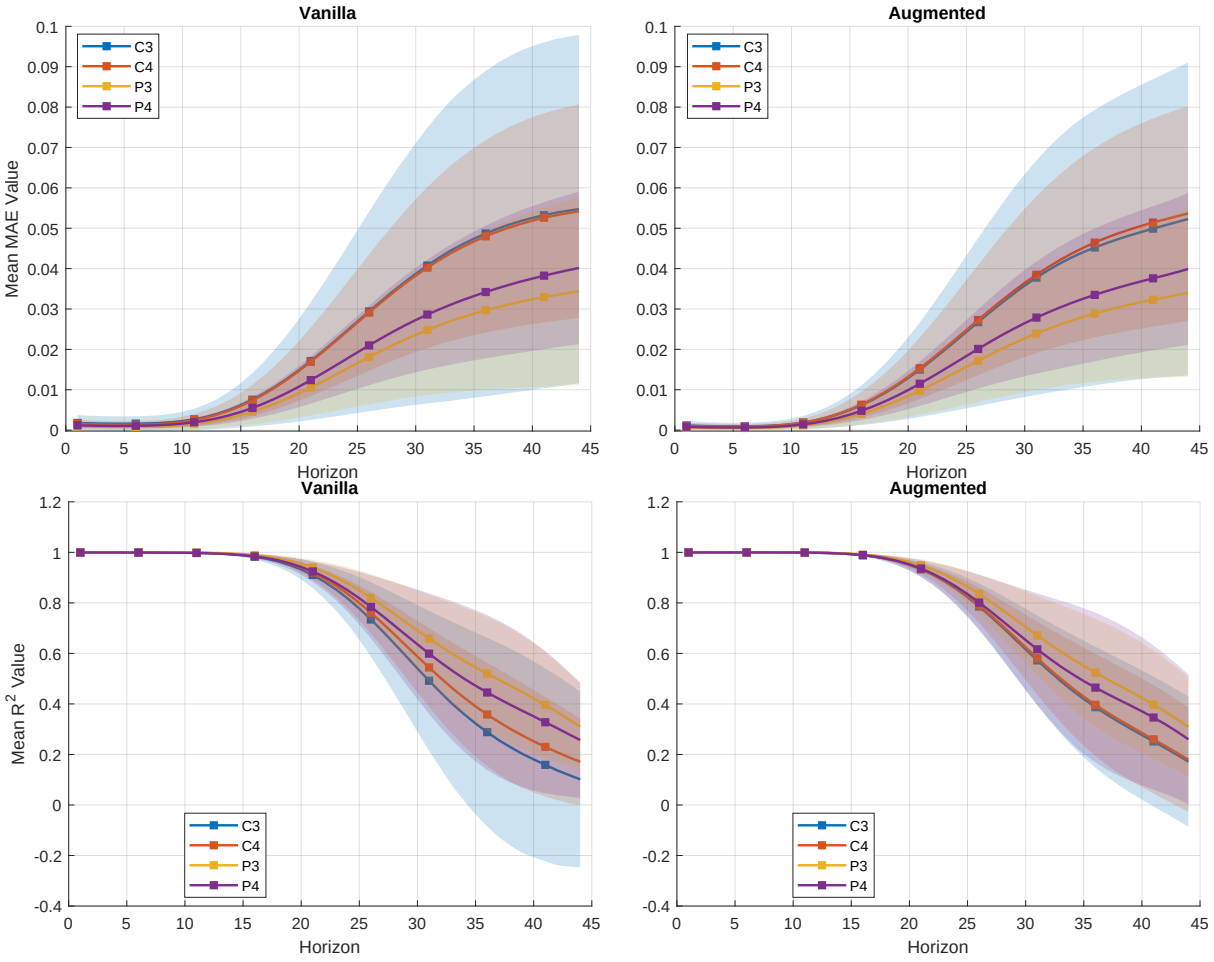


Figure 7.10: Forecast performance of the two PatchTST models for all the prediction horizon from1 to 44. The model trained on real data only (left) is compared with the model trained on the augmented dataset (right). Grand average MAE and R^2 for four channels (C3, C4, P3, and P4) are reported, for each temporal horizon ranging from 1 to 44. The shaded area represents the standard deviation across subjects.

that yields realistic - if low in variety - samples? In this work, we propose a novel method to remove this lazy behavior by introducing a geometry loss term to spread the generator's produced batches. This geometrical loss is especially suited when treating EEG data, since it relies on a representation involving the ACM, a representation widely used in the literature [136, 160, 161].

Riemannian Modeling of EEG Covariance Structure

The proposed generative framework maps windows of decimated α-band envelopes to corresponding ACM matrices, aiming to exploit their geometric properties. ACM matrices are SPD and live in a hyper-dimensional cone; by following the curvature of this space, it is possible to compute a geodesic distance between matrices. In particular, the AIRM provides robustness against linear perturbations, making it widely used in Riemannian EEG practices. This metric can be written to be entirely differentiable and numerically stable; by pairing it with the traditional GAN loss function for the Generator, it becomes possible to force the synthesized samples to spread out. Furthermore, the Riemannian framework is useful for evaluating sample quality: by clustering in the SPD space, it is possible to reconstruct the

distribution of the original data and compare it with the clusters of fake or novel data.

Novelty and Methodological Contribution

Existing EEG data augmentation can currently be expressed by defining two paradigms: geometric transformations such as shifting, noise injection, and splicing, and Euclidean deep generative models. While traditional transformations [101, 102] are computationally inexpensive and have proven effective in specific settings, they fail to introduce genuine distributional variety. Conversely, recent deep learning approaches, such as DeepConvGAN or EEG-GAN, treat neural signals as flat vectors or 2D images, implicitly imposing a Euclidean geometry on the data.

By mapping the signal dynamics into the ACM domain, we instead leverage the intrinsic spatial covariance structure of the EEG. While Riemannian geometry has established itself as an excellent classification tool, involving algorithms such as the seminal Riemannian MDM [162] or Tangent Space Mapping [163], its application to generative processes has been largely unexplored. To the best of our knowledge, this work represents one of the first attempts to directly embed the manifold constraints of SPD matrices into the optimization loop of a neural generator. Unlike standard methods that limit themselves to sample-wise error, our method respects the curvature of the statistical manifold, ensuring that the generated covariance structures remain biologically and mathematically consistent.

The idea is that, by forcing variety in the SPD space with our composed loss function, we can avoid mode collapse, a failure in training rather than a model definition issue.

This approach draws on the literature: multiple GAN modifications have been proposed to nudge the generator towards better generalization. These usually involve editing the discriminator: Gradient Penalty modifies the discriminator's objective to enforce Lipschitz continuity, providing the generator with smoother gradients; Spectral Normalization constrains the discriminator's weights to stabilize training, while methods like Minibatch Discrimination allow the discriminator to examine a batch of samples simultaneously to detect a lack of diversity. However, none of these leverage the data's geometric properties.

The proposed method forces the generator, rather than the discriminator, to push generated samples apart along the geodesics of the SPD cone. This functions as a domain-specific equivalent to Minibatch Discrimination, with the advantage of ensuring that the variance in the generated batch corresponds to meaningful variations in channel synchronization - linear phase and amplitude relationships are captured by the ACM representation - rather than arbitrary addition of noise.

Furthermore, the evaluation framework proposed here — utilizing Riemannian clustering to visualize distributional overlap — offers a new perspective compared to standard metrics, whether based on data features, such as the spectrogram, or neural encodings. The Riemannian approach highlights how the synthetic data is distributed with respect to the real one, immediately denouncing possible failures or regions of overspecialization.

Assessing the Coverage of Synthetic EEG Samples

A first observation that emerges from the results is the importance of a preliminary visual assessment. Looking at the produced samples can immediately capture low-quality or repeated instances, spotting faults in the training procedure. For example, compare the successful case of Fig. 7.3b with the flawed one of Fig. 7.3c. Still, even an apparently satisfactory generator may be moving around a limited region of the data space, warranting further analysis. The dual combination of visualizing ACM data in an embedded 2D space and clustering it in its original form allows for a comprehensive understanding of the synthetic data distribution. Notice how a flawed generation produces small aggregations of synthetic data inside the entire domain, both visually (Fig. 7.7) and in terms of distributional discrepancy (Tab. 7.1, Vanilla). Compare it with the geometrically regularized generator, which shows a proper coverage of the original data domain (Fig. 7.6, Tab. 7.1, Regularized).

Furthermore, we demonstrate that properly generated data can improve the quality of another model. We reintroduce the forecasting PatchTST model proposed in Sec. 4.1 and retrain it on an augmented dataset. Results show that a small yet consistent improvement (roughly 3% in MAE across all subjects)

persists at near-horizon timescales, indicating the effectiveness of the new data. At longer horizons, the augmented model shows improved consistency across channels, significantly limiting the worst-performing channel's error and error variance at the 33rd and 44th horizons.

Why the Geometry-based Regularization Works

The proposed loss function augments the standard generator cross-entropy with an additional term derived from pairwise distances between synthetic samples. This new term is extracted from the computation of all pair-wise distances of synthetic α-band envelope EEG data. Distances are computed by mapping synthetic samples to a common ACM space, in which the AIRM can be defined. The AIRM is a robust, affine-invariant metric that returns the geodesic distance between two SPD matrices. Not only does the AIRM return a precise, consistent distance between samples, but it also has the advantage of being fully differentiable. By taking the closest synthetic pair and maximizing their distance, the new loss term acts as a regularizer that prevents the generator from producing too similar samples.

The regularized term induced by back-propagation could clash with the original cross-entropy term; it is important to ensure synergy between the two, especially during the first phase of training. Adopting gradient surgery techniques such as PCGrad and at least normalizing the weights of the two losses is recommended. In the long run, the second term should not cause any conflict, since proper data coverage actually helps the generator overcome the discriminator.

Scalability and Limitations

The proposed method has two main limitations. One is intrinsic to self-supervised approaches, the other is strictly computational. First of all, the combined loss function does not provide a mechanism for interpolation across subjects, as it does not allow the generator to extend its range beyond the data fed to the discriminator. Despite proper coverage of the existing distribution, samples with realistic ACM properties can only be produced within the existing data clusters. This is evident in Fig. 7.6.

From a computational perspective, the proposed method has a complexity of $\mathcal{O}(B^2)$ with respect to the batch size B, due to the construction of the pairwise cross-distance matrix. Crucially, the cost of a single AIRM computation scales cubically with the dimension of the ACM (N), leading to a total complexity of $\mathcal{O}(B^2N^3)$.

This dependency imposes a necessary compromise between representational accuracy (larger N) and training practicality. Specifically, the geodesic distance relies on `torch.linalg.solve` and `torch.linalg.eigvals`, both of which are $\mathcal{O}(N^3)$ operations. However, eigenvalue decomposition is notably more expensive than solving linear systems; while the latter is a fast cubic operation, the former involves iterative algorithms with a significant constant factor (approximately $10/3\mathrm{N}^3$ for the initial reduction alone).

Integration of the novel method in EEG-BCI Frameworks

The propoesed methodology is appropriate for BCI paradigms where discriminative information is encoded in the covariance structure of ongoing EEG activity. Motor imagery is an example, because class-specific patterns emerge from distributed oscillatory dynamics of sensorimotor networks. At the same time, similar consideration hold for passive BCI, where cognitive states are reflected in large-scale neural activirty patterns captured at the covariance level. A primary advantage of our approach is its compatibility with standard EEG-BCI processing chains. Since the method operates on the covariance manifold, it does not require any modification of the feature extraction or classification stages. The generator acts as a plug-and-play module that enriches the training set upstream, improving the downstream classifier performance without modifying its architecture.

In an operational BCI scenario, the augmentation process is performed during the calibration phase using a pre-trained and fine-tuned model, expanding the data used to train the subsequent classifier. Standard calibration sessions often fail to capture the full range of non-stationary EEG fluctuations, such as changes in attention, fatigue or cognitive engagement, that naturally occur during long-term

use. By forcing the generator to explore the covariance manifold via the proposed regularized loss, the synthetic data can preemptively cover a wider region of the covariance space than covered by the limited calibration recordings. As a result, the classifier will be more robust to natural signal drift, leading to more stable BCI performance over time. Practically speaking, the generator and discriminator can be fine-tuned with few epochs on the small calibration dataset, introducing a negligible delay in the entire pipeline.

This methodology proves particularly impactful in small sample settings, which are endemic to clinical BCI applications. For patient-specific calibration, collecting large datasets from individual patients—especially those with motor impairments—is often arduous and fatiguing. The porposed method can amplify a short calibration session (e.g., a few minutes of motor imagery) into a large, statistically rich dataset, thus enabling the training of high-performance deep learning models that typically require massive data.

Chapter 8

Discussion

8.1 Vigilance and EEG

Throughout this work, a holistic pipeline for the protocol definition, acquisition, analysis, feature extraction, augmentation, and forecasting of EEG data is defined. The focus is on human vigilance recordings and on the involved neural oscillatory bands, especially α.

α band intensity modulates with human attention, being suppressed when concentration is required and bursting during brief neural relief. In Sec. 2.2.3, we define a task aiming to stimulate this kind of brain response. Taking inspiration from the repetitive nature of work in an assembly line, we built an experimental protocol based on observations of a robotic Kuka arm. The task requires sustained attention from participants, who are asked to track the robot's movements and check whether it places a toy block in the correct positions. EEG recordings taken from this protocol manifest recurrent contractions of α and form the basis of all the subsequent research. We observe two types of persistent behavior among participants: their α amplitude is synchronized with the robot's movements, and slowly increases with time-on-task. The first phenomenon indicates suppression and attention during target movement, and release and relaxation during the robot's return to its home position. The second phenomenon denotes steadily increasing boredom and detachment from the task, which should be expected from untrained individuals performing repetitive motions lasting for more than 20 minutes. Although vigilance fluctuates irregularly, with intermittent periods of focus followed by surges of tiredness, an indisputable difference between the amplitudes of α is confirmed for all subjects between the beginning and the end of the task.

The focus on α follows precise considerations: first, it is the main frequency band studied in the human vigilance literature due to its evident synchronous behavior in centro-parietal areas and desynchronization in the frontal area. Second, it is easier to capture under difficult acquisition conditions because it manifests at a higher electric potential, making it less susceptible to electromagnetic noise and jitter in the setup equipment. Our protocol was carried out in an incredibly noisy environment, where external vibrations and radiation from other electronic equipment required us to craft a makeshift EM shield for our cabled equipment. Far from being an ideal controlled environment, such hardships prove that the proposed method is highly robust - mostly due to the aggressive filtering performed on α. After extracting the α envelope using wavelets, we heavily decimate the resulting signal to obtain a smooth, neurologically consistent oscillatory marker of human vigilance.

Implementing this filtering pipeline in a real-time scenario presents specific computational challenges. While efficient, real-time adaptations of the CWT are established in the literature [164], the zero-phase, double-pass filter used in pre-processing is inherently non-causal and cannot be ported directly to an online stream. A feasible approximation involves using a causal second-order IIR Butterworth filter, which minimizes phase distortion but introduces a group delay of approximately 2.14 s. Crucially, the augmented Patch-TST model achieves robust performance at a forecast horizon of 5.5 s. This implies that the model's predictive capability not only accommodates the filter's latency but also effectively overcomes it, functioning as a real-time, zero-phase predictive filter with zero effective latency. The

predictive margin of 3.36 s suggests that α amplitude forecasting is indeed feasible and can be further improved through architectural refinements and increased data availability.

Although this thesis focuses on theoretical validation, these results suggest that the proposed strategy is apt for integration into closed-loop applications. For instance, in safety-critical environments, the system could dynamically modulate task automation levels: detecting a predicted drop in engagement could trigger an increase in automation, thereby preventing failures caused by human distraction.

8.2 Deep Learning Architectures for Vigilance Forecasting

The apparent simplicity of the pre-processed α-band envelope must not be conflated with linearity or stationarity. Despite its smooth temporal profile, the signal preserves non-linear dynamics, expressed through amplitude modulations and inter-channel phase shifts that evolve over time. Consequently, the assumption of weak stationarity—upon which classical methods rely—is fundamentally violated. This explains the baseline VAR model's inability to anticipate upcoming oscillations, as it is by construction constrained to linear dependencies and fails to capture the state-dependent evolution of neural synchronization. To overcome these constraints, we investigated the dominant DL architectures for time-series forecasting and explored their ability to model non-linear temporal dynamics.

Currently, Transformer-based architectures represent the reference standard in time series forecasting. Unlike sequential models, Transformers leverage self-attention to access a global receptive field, enabling the modeling of long-range dependencies without the sequential bottlenecks of recurrent architectures. Prominent instances include the *Informer* [165], which introduces ProbSparse attention to reduce computational complexity; the *Autoformer* [166], which incorporates a decomposition block to explicitly model trend and seasonality; and *PatchTST* [167], which segments time series into sub-sequences (patches) to preserve local semantic information while reducing token length. For smooth signals like the α-envelope, PatchTST is particularly appropriate, as the patching process naturally aligns with the envelope's slow variation, allowing the attention mechanism to learn relationships between macroscopic signal segments rather than point-wise fluctuations. In our experimental setup, the PatchTST model was optimized using the Optuna framework for hyperparameter search. The optimization converged on a patch length of 8, a configuration that underscores the necessity for the model to capture extended temporal strides to effectively model the slow-moving envelope dynamics.

Concurrently, RNNs have evolved significantly. While traditional RNNs suffered from the vanishing gradient problem, gating variants (LSTMs, GRUs) alleviated this issue, making Encoder-Decoder (ED) architectures a standard benchmark for sequence-to-sequence tasks. More recently, the field has seen a resurgence of recurrent-style modeling through Structured State Space Models (SSMs), such as S4 [168] and Mamba [169]. These architectures discretize continuous-time latent state equations, offering the inference speed of RNNs with the training parallelization of Transformers. Theoretically, SSMs are well-suited for smooth, continuous signals such as EEG envelopes, as they excel at modeling continuous dynamics.

Finally, Convolutional Networks specialized in time series (such as the TCNs) offer a third solution. TCN, in particular, uses dilated causal convolutions to expand the receptive field exponentially without the computational overhead of attention mechanisms.

In the subsequent experimental comparison, we benchmark these three dominant paradigms—a Transformer-based model (PatchTST), a recurrent Encoder-Decoder (ED-GRU), and a TCN. Our results demonstrate that while both the ED-GRU and the Transformer can reproduce the dominant signal frequency, the Transformer-based approach achieves superior predictive performance. Specifically, the PatchTST model achieves a remarkable forecasting horizon of 5.5 seconds. Given the experimental protocol, this implies the capability to predict the human vigilance α-marker up to half a trial in advance, providing a substantial window for preemptive intervention.

This forecasting paradigm is not limited to the α rhythm but could be seamlessly extended to other frequency bands, such as θ, commonly associated with fatigue, or β, associated with active concentration. The ability to accurately forecast these biomarkers with seconds' notice holds implications for safety-critical environments. By anticipating drops in vigilance or attention, systems could dynamically adapt

to the operator's mental state, for instance, by increasing automation levels in autonomous driving or industrial robotics, thereby mitigating risky situations before human error occurs.

8.3 Geometrical Generation of α

8.3.1 An overview of EEG data augmentation

EEG data scarcity is a serious research issue that underscores the need to generate high-quality synthetic samples. However, no uniform approach exists in the literature for either performing or evaluating data augmentation, making comparison between existing strategies an arduous task. This limitation is intrinsic to EEG data, in which multiple variables influence the final acquisition, ranging from subject-specific factors to task design. Despite the heterogeneity of methods, two main classes may be distinguished: classical approaches that aim to perturb existing data via linear transformations or spectral analysis, with strategies of varying complexity, and methods relying on neural networks, which draw inspiration from the time series literature.

Despite progress in AI, non-neural approaches have continued to be developed over the years, as demonstrated by multiple recent examples. In [170], the authors propose swapping data between symmetrical electrode pairs, e.g., C3 and C4, with care taken to ensure task-specific label renaming. Notheworhty is the work of [171], where two proposed data augmentation methods leveraging time-frequency decomposition (via Discrete Wavelet Transform and Hilbert–Huang Transform, respectively) address data scarcity, non-stationarity, and inter-subject variability. Both approaches involve decomposing signals into time-frequency components, reassembling sub-signals across different subjects to enhance diversity, and reconstructing them into augmented time-domain trials. Validated across 17 datasets across three BCI paradigms, these methods outperformed nine existing techniques, achieving an average 4% improvement in accuracy. A statistical method is described in [172], where frequency-band correlation analysis is combined with random sampling, selecting only high-correlation samples to ensure realism. Evaluations using distribution analysis, PERMANOVA tests, and a random forest classifier confirm that the synthetic data preserves the statistical and structural properties of real EEG.

Still, the total corpus of related works is dominated by neural models, with a wide variety of architectures emerging across publications. Recent examples include [173], employing a generative pre-trained Transformer (EEGPT) and a contrastive learning strategy (CLTISI) to address EEG non-stationarity and inter-subject variability in the field of emotion recognition. In [174], authors explore using GANs to augment EEG datasets for diagnosing Major Depressive Disorder, aiming to overcome the limitations of small sample sizes. Through a systematic review and empirical testing on two clinical datasets, the authors found that while synthetic data can improve diagnostic accuracy (up to 10% in one case), current generative methods struggle to faithfully reproduce all frequency-domain features and lack standardized evaluation metrics. Chen et al. ([175]) present FusionGen, an encoder-decoder framework that addresses data scarcity and inter-subject variability by leveraging disentangled representation learning and feature fusion, and is validated across multiple public datasets. A weighted co-training framework for EEG emotion recognition, based on a frequency-spatial diffusion transformer, is built in [176] to generate high-fidelity synthetic signals. By employing a pseudo-data weighting module to evaluate and select generated samples for joint training with real data, the method improves classifier generalization.

The wide range of multiple examples stresses the importance of a domain- and model-agnostic methodology. While being applied to a specific use case, namely the improvement of GANs creating windows of smoothed α envelope, our geometrical strategy improves sample variety by adding a regularization term in the model loss function, and facilitates data quality assessment by empirically sampling the data distribution. We believe that, with a few adjustments, it could be easily applied to other domains to further improve existing neural strategies without significant changes.

8.3.2 ACM properties and neural applications

To expand the cohort of available data, and to explore the implications that generative AI could bring in aiding forecasting models, multiple experiments were reported along this thesis. If the results obtained in Ch. 6 confirm that smooth envelope data can be convincingly reproduced by a simple recurrent architecture, the geometrical studies of Ch. 7 push this much further thanks to the properties of the ACM, which make it a desirable representation of EEG data. Not only is it possible to obtain a metric between EEG samples by tracing a geodesic on the SPD cone manifold, but the AIRM metric also computes a distance unperturbed by affine transformations.

In fact, multiple recent works exploit the ACM representation, offering insight into its capabilities. Expanding upon the SPD nature, Carrara et al. demonstrated that these augmented matrices possess an inherent Block-Toeplitz structure, which can be mathematically decomposed into a Cartesian product of an SPD manifold and a Siegel Disk Space [150]. This geometric decomposition significantly enhances BCI decoding efficiency, enabling architectures such as Phase-SPDNet to achieve state-of-the-art motor imagery classification with as few as 3 electrodes. Parallel to this, Sosulski and Tangermann repurposed linear discriminant analysis for event-related potentials by enforcing a block-Toeplitz structure on covariance matrices, reducing calibration times and errors in BCI spellers [177]. The theoretical push into more complex geometries culminated in the development of Siegel Neural Networks [178], which natively embed fully connected and multiclass logistic regression layers within the Siegel domain to process structured matrices without Euclidean approximations. In clinical and applied contexts, Ding et al. advanced epilepsy diagnostics by using quotient-correlation geometries to track pre-ictal seizure trajectories, isolating true connectivity from artifact-driven amplitude spikes [179]. For affective computing, Wosiak et al. introduced geodesic-mean contraction techniques to stabilize trial covariances on the Riemannian manifold, achieving highly robust cross-session emotion recognition [180]. Similarly, Geirnaert et al. leveraged Riemannian Geometry-based Classifiers (RGC) to create unsupervised, self-adapting auditory attention decoders for neuro-steered hearing aids. Bridging deep learning and classical signal processing, Paillard et al. developed GREEN (Gabor Riemann EEGNet), a lightweight architecture that combines Riemannian geometry with learnable wavelet transforms to extract interpretable biomarkers for dementia and pathology detection [181].

Together, these works report a mathematically sound and interpretable paradigm that captures both the spatial and temporal richness of EEG signals. Notably, none of these cited accomplishments adapts the Riemannian framework for data augmentation, although they suggest possible avenues for expanding our current work: considering ACM as belonging to the Siegel Disk Space could significantly reduce the computational cost of pairwise distance extraction.

A couple of works that are strictly related to EEG data generation are [182] and, more recently, [183], which propose a way to interpolate between two EEG samples by moving along the geodesic connecting them in Riemannian space. This approach is completely different from our novel, regularized loss function, but offers another interesting perspective on how to integrate knowledge of the ACM space inside a generative AI methodology.

8.3.3 Reported Results

Using a small sample size of 10 subjects provides an ideal, realistic setting to test the validity of the proposed generative methods. Two main studies were conducted: one testing standard generative procedures and comparing them by evaluating performance and embeddings of an ED-GRU model, and another testing the geometrically regularized GAN. The first batch of generative experiments also covers the SPIS dataset, from which the α envelope was extracted, and proposes an ED-GRU forecaster to study the informational contribution of synthetic samples. In this case, Vigilance samples included both frontal β and parietal θ envelopes. Both the ED-GRU embeddings and the final forecasting performance were considered to assess the quality of the generative procedure; despite showing promising waveforms and an in-depth statistical analysis of the forecasting embeddings, the assessment remains purely empirical. Still, the average MAE and SMAPE improvements of 26.2% and 29% when considering 100% of

the SPIS dataset, and of 19% and 15.4% when considering 100% of the Vigilance dataset, indicate the appropriateness of the chosen GAN training procedure.

The second major study is based on a much less varied dataset: despite the total of roughly 10,000 different windows, they stride by two samples at a time, overlapping 254 out of 256 samples between sequential windows. Moreover, testing the potential improved model's prediction capabilities has more practical implications, as it relates directly to the downsampled Vigilance data already used in our forecasting experiment. The model in question, based on the PatchTST architecture, performs well on its own, making any improvement a benchmark of success for synthetic data. In fact, small but consistent improvements can be seen among all channels and lags (2-3 % for both MAE, R^2), with few exceptions, including slightly worse SD for the final horizons for three of the four channels, and a definite improvement for the otherwise worst-performing C3 channel.

Beyond the demonstrated improvement in forecasting accuracy, the geometric framework provides a rigorous means to analyze the distributional spread of synthetic data. By projecting points from the Riemannian SPD manifold into low-dimensional embeddings via topology-preserving algorithms such as UMAP, it is possible to gain insights into the Generator's behavior. In particular, the synthetic samples in Fig. 7.6 exhibit a diversity that faithfully mirrors the original distribution, while overspecialized regions emerge in Fig.7.7.

8.3.4 AIRM-based Regularization Limitations

The limited data availability confirms, on the one hand, the usefulness of GAN in scarcity settings, while, on the other hand, it requires further experimentation across multiple datasets to properly confirm the proposed method's effectiveness. Still, from our current results, two major blind spots can be observed.

Fig. 7.6 illustrates both the efficacy and the inherent constraints of the proposed method: while the synthetic samples exhibit a diversity that faithfully mirrors the original distribution, they do not extrapolate to unexplored regions of the SPD manifold. This behavior is expected; GANs are fundamentally designed to model the support of the training distribution exposed to the Discriminator, limiting their ability to generate samples outside the convex hull of the observed data.

Another critical consideration for the practical deployment of this framework is the computational burden associated with the training phase. While the inference step (signal generation) is efficient, the training process is heavily constrained by the geometric operations. Specifically, the necessity of computing the pairwise Affine Invariant Riemannian Metric (AIRM) introduces a computational complexity that scales quadratically with the batch size and cubically with the covariance dimension ($\mathcal{O}(B^2N^3)$). Furthermore, integrating multi-objective optimization techniques to balance the adversarial and geometric losses incurs additional overhead.

In our experiments, this bottleneck was particularly acute when employing GradNorm. The method's requirement to compute and normalize gradients for each task weight at every step rendered the training computationally prohibitive; a single model required over 24 hours to complete just 20 epochs, making it unfeasible for iterative research or rapid deployment. Conversely, PCGrad proved to be a far more efficient alternative. By modifying the gradients only when a conflict direction is detected—rather than actively normalizing magnitudes at every step—PCGrad maintained manageable training times (approximately 2–3 hours per model) while delivering satisfactory stabilization and performance. This suggests that, for geometrically constrained GANs, conflict resolution is a more critical optimization factor than magnitude balancing.

8.3.5 Future Works: Improving Geometrical Regularization

A major step forward would be building a model capable of interpolating across-subject samples, rather than only augmenting existing data regions. Drawing inspiration from Riemannian interpolation strategies described in the literature [182, 183], it may be possible to construct geodesic paths between distinct subject clusters. Such a mechanism would facilitate the traversal of sampling-sparse regions of the SPD space, offering a robust solution for covering vast portions of the manifold. This approach holds particular

promise for BCI calibration: by geometrically interpolating around sparse calibration samples, we could significantly expand the representational space available for transfer learning and model fine-tuning.

Secondly, to address the computational bottleneck imposed by AIRM calculation, several optimization strategies warrant investigation. A first improvement involves metric substitution, for example, by replacing the computationally expensive AIRM with the Log-Euclidean Metric (LEM) or the Stein Divergence. The LEM maps matrices to the tangent space, reducing the pairwise comparison to a Euclidean distance calculation ($\mathcal{O}(B^2N^2)$) after a single batch-wise logarithmic map, while the Stein Divergence exploits Cholesky decompositions to approximate the Riemannian geometry without iterative eigenvalue solvers. Alternatively, quadratic batch scaling ($\mathcal{O}(B^2)$) can be mitigated by adopting a Centroid-Based Repulsion strategy, which enforces sample diversity by maximizing each sample's distance from the Riemannian or Log-Euclidean mean of the batch, rather than computing all-to-all distances.

Otherwise, significant computational gains can be realized by rethinking the pairwise loss formulation itself. The diversity loss effectively treats the batch as a fully connected graph, calculating the interaction between every possible pair of samples ($\mathcal{O}(B^2)$). We propose two sampling strategies to reduce this overhead while maintaining the geometric regularization.

The first approach is a Monte Carlo approximation of the diversity loss. Rather than computing the full pairwise distance matrix, one can select a subset of randomly sampled pairs at each training step. By optimizing the AIRM distance over only $K \ll B^2$ pairs, the computational burden scales linearly with the batch size. While this introduces variance into the gradient estimate, such stochasticity can be advantageous in non-convex optimization, potentially helping the generator escape saddle points synonymous with partial mode collapse.

A more sophisticated sampling strategy involves exploiting the known geometry of the input latent space. Since Euclidean distances in the Gaussian noise hypersphere ($\mathcal{Z}$) are computationally negligible, we can pre-calculate the full pairwise distance matrix in $\mathcal{Z}$ to guide the selection of samples in the Riemannian space. Specifically, identifying the pair (z_i, z_j) with the minimum Euclidean distance in the batch allows us to target the generator's local sensitivity. By forcing these two neighboring noise vectors to map to sufficiently distant points on the SPD manifold (maximizing $d_{\text{AIRM}}(G(z_i), G(z_j))$), we explicitly prevent the surjective mapping characteristic of mode collapse. This enforces a lower bound on the local Lipschitz constant of the generator, ensuring that the model remains sensitive to small perturbations in the latent space and preserving the fine-grained diversity of the generated signal envelopes.

Another significant optimization lies in exploiting the specific structure of the ACMs used in this work. As demonstrated by Carrara et al. [150], treating ACMs merely as generic large SPD matrices ignores their inherent Block-Toeplitz structure. This manifold can be mathematically decomposed into a Cartesian product of a standard, smaller SPD manifold, representing the spatial covariance block, and a sequence of Siegel Disk Spaces, representing the lagged temporal blocks. By transforming these temporal blocks into Verblunsky coefficients, which reside natively within the Siegel domain, the high-dimensional ACM analysis can be split into smaller, geometrically appropriate components, processed with specialized tools such as Kähler potentials. Integrating this decomposition into the generative loss function would not only drastically reduce computational demand by breaking down the N^3 cost but also enforce stricter adherence to the specific temporal geometry of the EEG signal, potentially enhancing both training speed and generation fidelity.

After considering all potential optimizations, the final step to implement this methodology for real-time calibration data augmentation is fine-tuning the model. As established in the generative literature, transfer learning via GAN fine-tuning allows models trained on large, diverse source domains to rapidly adapt to data-scarce target environments without overfitting [184]. Some established methodologies include freezing the initial layers of the discriminator [185], or changing only the statistical layers, such as batch normalization [186]. A GAN pre-trained on a multi-subject pool requires only a short and patient-specific calibration session, for example, a few minutes of motor imagery, to adapt its weights to the new user's specific covariance distribution. The proposed regularized loss forces the generator to expand upon these sparse calibration samples, synthesizing a statistically rich dataset that improves the training of downstream deep learning models.

Geometry-Informed Deep Learning

This work can be considered belonging within Geometry-Informed Deep Learning (GIDL). GDL specifically focuses on constraining the network's outputs or internal representations using known geometric rules and boundaries. This broad term covers multiple recent works: authors in [187] create 3D designs by solving structural optimization problems guided purely by user-specified topological constraints, avoiding the usage of training data; authors of [188] propose a geometrical module to unfold 2D representations into a volumetric space, enabling accurate and sparse 3D CT and volumetric MRI reconstruction with sub-second acquisition times; while Hu et al. developed a deep learning framework integrating graph convolutional networks and geometric transfer matrices to accurately reconstruct cardiac transmembrane potentials from standard 12-lead ECGs [189].

Standard GANs implicitly assume that the data support lies on a manifold embedded in Euclidean space and use Euclidean metrics (such as MSE or standard cross-entropy) to optimize this mapping. However, EEG covariance descriptors naturally reside on the Riemannian manifold of SPD matrices, a space with non-positive curvature. By implementing AIRM as a differentiable loss function, we transform the generator into a geometry-informed model. The network is no longer minimizing error in a flat vector space but is optimizing trajectories along the geodesics of the SPD cone. This ensures that the diversity enforced by the loss function reflects the true topological distance between biological states, rather than minimizing an arbitrary sample-wise difference.

8.3.6 Final Words

To conclude, we went through all the major steps that compose an EEG acquisition inside an industrial setting. From data collection, struggling with noise and filtering, to choosing specific frequency envelopes as appropriate time-variant features; experimenting with a wide range of forecasting models, ranging from the classical VAR to modern DL architectures; and finally, proposing a rigorous method to escape the curse of data availability. In particular, the geometrical approach underlying the new GAN loss function opens the door to numerous future works and establishes a sound angle from which to approach the issues of generative AI.

Bibliography

[1] Hamwira Yaacob et al. “Application of Artificial Intelligence Techniques for Brain–Computer Interface in Mental Fatigue Detection: A Systematic Review (2011–2022)”. In: *IEEE Access* 11 (2023), pp. 74736–74758.

[2] Rabie A Ramadan and Ahmed B Altamimi. “Unraveling the potential of brain-computer interface technology in medical diagnostics and rehabilitation: A comprehensive literature review”. In: *Health and Technology* 14.2 (2024), pp. 263–276.

[3] Xiu-Yun Liu et al. “Recent applications of EEG-based brain-computer-interface in the medical field”. In: *Military Medical Research* 12.1 (2025), p. 14.

[4] Jiayan Zhang et al. “Recent progress in wearable brain–computer interface (BCI) devices based on electroencephalogram (EEG) for medical applications: A review”. In: *Health data science* 3 (2023), p. 0096.

[5] Raed Mohammed Hussein, Firas Sabar Miften, and Loay E George. “Driver drowsiness detection methods using EEG signals: a systematic review”. In: *Computer methods in biomechanics and biomedical engineering* 26.11 (2023), pp. 1237–1249.

[6] Thiago Gabriel Monteiro, Charlotte Skourup, and Houxiang Zhang. “Using EEG for mental fatigue assessment: A comprehensive look into the current state of the art”. In: *IEEE Transactions on Human-Machine Systems* 49.6 (2019), pp. 599–610.

[7] Andra Mahu et al. “Validation of Vigilance Decline Capability in a Simulated Test Environment: A Preliminary Step Towards Neuroadaptive Control”. In: *Neuroergonomics and Cognitive Engineering* (2024), p. 45.

[8] Rahul Ajmeria et al. “A critical survey of EEG-based BCI systems for applications in industrial internet of things”. In: *IEEE Communications Surveys & Tutorials* 25.1 (2022), pp. 184–212.

[9] Klara Hemmerich et al. “Understanding vigilance and its decrement: theoretical, contextual, and neural insights”. In: *Frontiers in Cognition* 4 (2025), p. 1617561.

[10] John K Hawley and Anna L Mares. *Developing effective human supervisory control for air and missile defense systems*. Tech. rep. ARMY RESEARCH LAB FORT BLISS TX, 2006.

[11] Ilia Pershin et al. “Vigilance described by the time-on-task effect in EEG activity during a cued Go/NoGo task”. In: *International Journal of Psychophysiology* 183 (2023), pp. 92–102.

[12] Edmund Wascher et al. “Frontal theta activity reflects distinct aspects of mental fatigue”. In: *Biological psychology* 96 (2014), pp. 57–65.

[13] Leonard J Trejo et al. “EEG-based estimation and classification of mental fatigue”. In: *Psychology* 6.05 (2015), p. 572.

[14] Marika Sebastiani et al. “Neurophysiological vigilance characterisation and assessment: Laboratory and realistic validations involving professional air traffic controllers”. In: *Brain sciences* 10.1 (2020), p. 48.

[15] Henriques Zacarias et al. “ECG Forecasting System Based on Long Short-Term Memory”. In: *Bioengineering* 11.1 (2024), p. 89.

[16] Peter G Jacobs et al. “Artificial intelligence and machine learning for improving glycemic control in diabetes: best practices, pitfalls and opportunities”. In: *IEEE Reviews in Biomedical Engineering* 17 (2023), pp. 19–41.

[17] Sarala Ghimire et al. “Deep learning for blood glucose level prediction: How well do models generalize across different data sets?” In: *PLoS One* 19(9) (2024), e0310801.

[18] Gourav Siddhad et al. “Awake at the Wheel: Enhancing Automotive Safety through EEG-Based Fatigue Detection”. In: *arXiv preprint arXiv:2408.13929* (2024).

[19] Fo Hu et al. “EEG-Based Driver Fatigue Detection Using Spatio-Temporal Fusion Network With Brain Region Partitioning Strategy”. In: *IEEE Transactions on Intelligent Transportation Systems* 25.8 (2024), pp. 9618–9630.

[20] Hanna Pankka et al. “Forecasting EEG time series with WaveNet”. In: *bioRxiv* (2024), pp. 2024–01.

[21] Sichao Liu, Lihui Wang, and Robert X Gao. “Cognitive neuroscience and robotics: Advancements and future research directions”. In: *Robotics and Computer-Integrated Manufacturing* 85 (2024), p. 102610.

[22] Nebojsa Bacanin et al. “Multivariate energy forecasting via metaheuristic tuned long-short term memory and gated recurrent unit neural networks”. In: *Information Sciences* 642 (2023), p. 119122.

[23] Abdul Khalique Shaikh et al. “A new approach to seasonal energy consumption forecasting using temporal convolutional networks”. In: *Results in Engineering* 19 (2023), p. 101296.

[24] Liyilei Su et al. “A systematic review for transformer-based long-term series forecasting”. In: *Artificial Intelligence Review* 58.3 (2025), p. 80.

[25] Ajay Bandi, Pydi Venkata Satya Ramesh Adapa, and Yudu Eswar Vinay Pratap Kumar Kuchi. “The power of generative ai: A review of requirements, models, input–output formats, evaluation metrics, and challenges”. In: *Future Internet* 15.8 (2023), p. 260.

[26] Ian J Goodfellow et al. “Generative adversarial nets”. In: *Advances in neural information processing systems* 27 (2014).

[27] Jiaqi Liang et al. “Generative Adversarial Networks for Modeling Bio-Electric Fields in Medicine: A Review of EEG, ECG, EMG, and EOG Applications”. In: *Bioengineering* 13.1 (2026), p. 84.

[28] Ettore Cinquetti, Gloria Menegaz, and Silvia F Storti. “Toward in-silico data assessment for passive BCIs: generating EEG rhythms with GANs”. In: *Journal of Neural Engineering* 22.6 (2025), p. 066007.

[29] Preeti Sharma et al. “Generative adversarial networks (GANs): introduction, taxonomy, variants, limitations, and applications”. In: *Multimedia tools and applications* 83.41 (2024), pp. 88811–88858.

[30] Sanaz Rezvani et al. “A review on the performance of brain-computer interface systems used for patients with locked-in and completely locked-in syndrome”. In: *Cognitive Neurodynamics* (2023), pp. 1–25.

[31] Carmen Vidaurre et al. “Challenges of neural interfaces for stroke motor rehabilitation”. In: *Frontiers in Human Neuroscience* 17 (2023), p. 1070404.

[32] Aayushi Khajuria and Deepak Joshi. “EEG based cortical investigation for the limit of stability analysis in transfemoral amputees: a comparison with able-bodied individuals”. In: *Biocybernetics and Biomedical Engineering* 42.1 (2022), pp. 362–374.

[33] M Sharanreddy and PK Kulkarni. “Automated EEG signal analysis for identification of epilepsy seizures and brain tumour”. In: *Journal of medical engineering & technology* 37.8 (2013), pp. 511–519.

[34] Ingeborg H Hansen et al. "Detection of a sleep disorder predicting Parkinson's disease". In: *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2013, pp. 5793–5796.

[35] CWNFCW Fadzal, W Mansor, and LY Khuan. "Review of brain computer interface application in diagnosing dyslexia". In: *2011 IEEE Control and System Graduate Research Colloquium*. IEEE. 2011, pp. 124–128.

[36] Clerc Maureen, Bougrain Laurent, and Lotte Fabien. *Brain-Computer Interfaces 1: Methods and Perspectives*. Wiley, 2016, pp. 89–90.

[37] Ines Assali et al. "CNN-based classification of epileptic states for seizure prediction using combined temporal and spectral features". In: *Biomedical Signal Processing and Control* 82 (2023), p. 104519.

[38] Amirmasoud Ahmadi, Hanieh Bazregarzadeh, and Kamran Kazemi. "Automated detection of driver fatigue from electroencephalography through wavelet-based connectivity". In: *Biocybernetics and Biomedical Engineering* 41.1 (2021), pp. 316–332.

[39] Wei Wei et al. "A MultiModal Vigilance (MMV) dataset during RSVP and SSVEP brain-computer interface tasks". In: *Scientific data* 11.1 (2024), p. 867.

[40] Soo-Yeon Han et al. "Classification of pilots' mental states using a multimodal deep learning network". In: *Biocybernetics and Biomedical Engineering* 40.1 (2020), pp. 324–336.

[41] Khalida Douibi et al. "Toward EEG-based BCI applications for industry 4.0: challenges and possible applications". In: *Frontiers in Human Neuroscience* 15 (2021), p. 705064.

[42] Bo'an Wei et al. "Construction Site Hazard Identification and Worker Adverse Reaction Monitoring Using Electroencephalograms: A Review". In: *Buildings* 14.1 (2024), p. 180.

[43] Tanghan Jiang et al. "Understanding construction workers' cognitive processes under risky scenarios through electroencephalography". In: *Automation in Construction* 166 (2024), p. 105674.

[44] Wolfgang Klimesch. "Memory processes, brain oscillations and EEG synchronization". In: *International journal of psychophysiology* 24.1-2 (1996), pp. 61–100.

[45] Gert Pfurtscheller, A Stancak Jr, and Ch Neuper. "Event-related synchronization (ERS) in the alpha band—an electrophysiological correlate of cortical idling: a review". In: *International journal of psychophysiology* 24.1-2 (1996), pp. 39–46.

[46] Ashley Craig et al. "Regional brain wave activity changes associated with fatigue". In: *Psychophysiology* 49.4 (2012), pp. 574–582.

[47] European Commission and Directorate-General for Research and Innovation, A Renda, et al. *Industry 5.0, a transformative vision for Europe – Governing systemic transformations towards a sustainable industry*. Publications Office of the European Union, 2021.

[48] European Commission and Directorate-General for Research and Innovation and J Müller. *Enabling Technologies for Industry 5.0 – Results of a workshop with Europe's technology leaders*. Publications Office, 2020.

[49] Yevheniy Dmytriyev, Marco Carnevale, and Hermes Giberti. "Enhancing flexibility and safety: collaborative robotics for material handling in end-of-line industrial operations". In: *Procedia Computer Science* 232 (2024), pp. 2588–2597.

[50] Shanhe Lou et al. "A human-cyber-physical system enabled sequential disassembly planning approach for a human-robot collaboration cell in Industry 5.0". In: *Robotics and Computer-Integrated Manufacturing* 87 (2024), p. 102706.

[51] Muhammad Hamza Zafar, Even Falkenberg Langås, and Filippo Sanfilippo. "Exploring the synergies between collaborative robotics, digital twins, augmentation, and industry 5.0 for smart manufacturing: A state-of-the-art review". In: *Robotics and Computer-Integrated Manufacturing* 89 (2024), p. 102769.

[52] Zainab Saleem et al. “A review of external sensors for human detection in a human robot collaborative environment”. In: *Journal of Intelligent Manufacturing* (2024), pp. 1–25.

[53] Ce Zheng et al. “Deep learning-based human pose estimation: A survey”. In: *ACM Computing Surveys* 56.1 (2023), pp. 1–37.

[54] Giovanni Boschetti et al. “3D collision avoidance strategy and performance evaluation for human–robot collaborative systems”. In: *Computers & Industrial Engineering* 179 (2023), p. 109225.

[55] Maurizio Faccio et al. “Human factors in cobot era: a review of modern production systems features”. In: *Journal of Intelligent Manufacturing* 34.1 (2023), pp. 85–106.

[56] Aitor Toichoa Eyam, Wael M Mohammed, and Jose L Martinez Lastra. “Emotion-driven analysis and control of human-robot interactions in collaborative applications”. In: *Sensors* 21.14 (2021), p. 4626.

[57] Carlo Caiazzo et al. “Development of a Neuroergonomic Assessment for the Evaluation of Mental Workload in an Industrial Human–Robot Interaction Assembly Task: A Comparative Case Study”. In: *Machines* 11.11 (2023), p. 995.

[58] Saranjuu Chulakit et al. “A Review of HRV and EEG Technology Applications in Industry 5.0: Emphasising Manufacturing Efficiency and Worker Well-Being”. In: *International Journal of Integrated Engineering* 16.5 (2024), pp. 324–340.

[59] Zohreh Zakeri et al. “Multimodal assessment of cognitive workload using neural, subjective and behavioural measures in smart factory settings”. In: *Sensors* 23.21 (2023), p. 8926.

[60] Marta Rinaldi et al. “Evaluation of mental stress in human-robot interaction: an explorative study”. In: *Procedia Computer Science* 232 (2024), pp. 726–735.

[61] Simone Borghi et al. “Assessing operator stress in collaborative robotics: A multimodal approach”. In: *Applied Ergonomics* 123 (2025), p. 104418.

[62] Mohd Umair Iqbal, Babji Srinivasan, and Rajagopalan Srinivasan. “Multi-class classification of control room operators’ cognitive workload using the fusion of eye-tracking and electroencephalography”. In: *Computers & Chemical Engineering* 181 (2024), p. 108526.

[63] Xin Fang et al. “Real-Time Monitoring of Mental Fatigue of Construction Workers Using Enhanced Sequential Learning and Timeliness”. In: *Automation in Construction* 159 (2024), p. 105267.

[64] Ethel Pruss et al. “Restoring Engagement in Human-Robot Interaction: A Brain-Computer Interface for Adaptive Learning with Robots”. In: *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE. 2023, pp. 3247–3252.

[65] Bitao Yao et al. “Task reallocation of human-robot collaborative production workshop based on a dynamic human fatigue model”. In: *Computers & Industrial Engineering* 189 (2024), p. 109855.

[66] Sruthi Ramadurai et al. “Physiological indicators of fluency and engagement during sequential and simultaneous modes of human-robot collaboration”. In: *IISE Transactions on Occupational Ergonomics and Human Factors* 12.1-2 (2024), pp. 97–111.

[67] Loıc Couture et al. “Adaptive System to Enhance Operator Engagement during Smart Manufacturing Work”. In: *Sensors & Transducers.* 265.2 (2024), pp. 106–119.

[68] Hardik Chauhan, Youjin Jang, and Inbae Jeong. “Predicting human trust in human-robot collaborations using machine learning and psychophysiological responses”. In: *Advanced Engineering Informatics* 62 (2024), p. 102720.

[69] Gianluca Borghini et al. “Avionic technology testing by using a cognitive neurometric index: a study with professional helicopter pilots”. In: *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2015, pp. 6182–6185.

[70] Miyakoshi Kothe, Christian Delorme, and Arnaud Makoto. *clean_rawdata*. Version 2.7. Software. 2019.

[71] Ettore Cinquetti et al. “Enhancing Safety in Industry 4.0: The Use of Passive Brain Computer Interfaces for Vigilance Monitoring”. In: *Eighth National Congress of Bioengineering – Proceedings 2023*. GNB atti. Pàtron editore, 2023, pp. 774–777.

[72] Mehmet Akin. “Comparison of wavelet transform and FFT methods in the analysis of EEG signals”. In: *J. Med. Syst.* 26.3 (2002), pp. 241–7.

[73] Tilmann Gneiting. “Making and evaluating point forecasts”. In: *Journal of the American Statistical Association* 106.494 (2011), pp. 746–762.

[74] Erfan Nozari et al. “Macroscopic resting-state brain dynamics are best described by linear models”. In: *Nature biomedical engineering* 8.1 (2024), pp. 68–84.

[75] Parinthorn Manomaisaowapak, Anawat Nartkulpat, and Jitkomut Songsiri. “Granger causality inference in EEG source connectivity analysis: a state-space approach”. In: *IEEE Transactions on Neural Networks and Learning Systems* 33.7 (2021), pp. 3146–3156.

[76] Danish M Khan, Norashikin Yahya, and Nidal Kamel. “Optimum order selection criterion for autoregressive models of bandlimited EEG signals”. In: *2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)*. IEEE. 2021, pp. 389–394.

[77] Andrew A Neath and Joseph E Cavanaugh. “The Bayesian information criterion: background, derivation, and applications”. In: *Wiley Interdisciplinary Reviews: Computational Statistics* 4.2 (2012), pp. 199–203.

[78] Laith Alzubaidi et al. “Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions”. In: *Journal of big Data* 8 (2021), pp. 1–74.

[79] Yao Wang et al. “Identifying mental fatigue of construction workers using EEG and deep learning”. In: *Automation in Construction* 151 (2023), p. 104887.

[80] Mobarak Abumohsen et al. “Hybrid machine learning model combining of CNN-LSTM-RF for time series forecasting of Solar Power Generation”. In: *e-Prime-Advances in Electrical Engineering, Electronics and Energy* 9 (2024), p. 100636.

[81] Xiao-Qian Lu et al. “CNN-LSTM based incremental attention mechanism enabled phase-space reconstruction for chaotic time series prediction”. In: *Journal of Electronic Science and Technology* 22.2 (2024), pp. 100256–100256.

[82] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. “An empirical evaluation of generic convolutional and recurrent networks for sequence modeling”. In: *arXiv preprint arXiv:1803.01271* (2018).

[83] Pradeep Hewage et al. “Temporal convolutional neural (TCN) network for an effective weather forecasting using time-series data from the local weather station”. In: *Soft Computing* 24 (2020), pp. 16453–16482.

[84] Pouya Bashivan et al. “Learning representations from EEG with deep recurrent-convolutional neural networks”. In: *arXiv preprint arXiv:1511.06448* (2015).

[85] Yong Yu et al. “A review of recurrent neural networks: LSTM cells and network architectures”. In: *Neural computation* 31.7 (2019), pp. 1235–1270.

[86] Junyoung Chung et al. “Empirical evaluation of gated recurrent neural networks on sequence modeling”. In: *arXiv preprint arXiv:1412.3555* (2014).

[87] Ralf C Staudemeyer and Eric Rothstein Morris. “Understanding LSTM–a tutorial into long short-term memory recurrent neural networks”. In: *arXiv preprint arXiv:1909.09586* (2019). arXiv: `1909.09586 [cs.NE]`.

[88] Kyunghyun Cho et al. “Learning phrase representations using RNN encoder-decoder for statistical machine translation”. In: *arXiv preprint arXiv:1406.1078* (2014). arXiv: `1406.1078 [cs.CL]`.

[89] Sadaqat Ali Rammy et al. “Sequence-to-sequence deep neural network with spatio-spectro and temporal features for motor imagery classification”. In: *Biocybernetics and Biomedical Engineering* 41.1 (2021), pp. 97–110.

[90] Nicole Gruber and Alfred Jockisch. "Are GRU cells more specific and LSTM cells more sensitive in motive classification of text?" In: *Frontiers in artificial intelligence* 3 (2020), p. 40.

[91] Min Xia et al. "A Stacked GRU-RNN-Based Approach for Predicting Renewable Energy and Electricity Load for Smart Grid Operation". In: *IEEE Transactions on Industrial Informatics* 17.10 (2021), pp. 7050–7059.

[92] Ettore Cinquetti et al. "A glimpse ahead: Forecasting 5.5-s human vigilance for enhanced safety in Industry 5.0". In: *Biocybernetics and Biomedical Engineering* 45.2 (2025), pp. 258–268.

[93] Oscar Almanza-Conejo et al. "Emotion recognition in EEG signals using the continuous wavelet transform and CNNs". In: *Neural Computing and Applications* 35.2 (2023), pp. 1409–1422.

[94] Jonathan M Lilly and Sofia C Olhede. "Generalized Morse wavelets as a superfamily of analytic wavelets". In: *IEEE Transactions on Signal Processing* 60.11 (2012), pp. 6036–6041.

[95] Yvonne Tran et al. "The influence of mental fatigue on brain activity: Evidence from a systematic review with meta-analyses". In: *Psychophysiology* 57.5 (2020), e13554.

[96] Skipper Seabold and Josef Perktold. "statsmodels: Econometric and statistical modeling with python". In: *SciPy* 7.1 (2010).

[97] Wen Guo et al. "Back to mlp: A simple baseline for human motion prediction". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2023, pp. 4809–4819.

[98] Simanto Saha and Mathias Baumert. "Intra-and inter-subject variability in EEG-based sensorimotor brain computer interface: a review". In: *Frontiers in computational neuroscience* 13 (2020), p. 87.

[99] Mamunur Rashid et al. "Current Status, Challenges, and Possible Solutions of EEG-Based Brain-Computer Interface: A Comprehensive Review". In: *Frontiers in Neurorobotics* 14 (2020).

[100] Andreas Schwarz et al. "Analyzing and decoding natural reach-and-grasp actions using gel, water and dry EEG systems". In: *Front. Neurosci.* 14 (2020), p. 567658.

[101] Fabien Lotte. "Generating artificial EEG signals to reduce BCI calibration time". In: *5th International Brain-Computer Interface Workshop*. 2011, pp. 176–179.

[102] Cédric Rommel et al. "Data augmentation for learning predictive models on EEG: a systematic comparison". In: *Journal of Neural Engineering* 19.6 (2022), p. 066020.

[103] Mario Michael Krell and Su Kyoung Kim. "Rotational data augmentation for electroencephalographic data". In: *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2017, pp. 471–474.

[104] Kay Gregor Hartmann, Robin Tibor Schirrmeister, and Tonio Ball. "EEG-GAN: Generative adversarial networks for electroencephalograhic (EEG) brain signals". In: *arXiv preprint arXiv:1806.01875* (2018).

[105] Yun Luo and Bao-Liang Lu. "EEG data augmentation for emotion recognition using a conditional Wasserstein GAN". In: *2018 40th annual international conference of the IEEE EMBC*. IEEE. 2018, pp. 2535–8.

[106] Chenxi Tian et al. "Dual-encoder VAE-GAN with spatiotemporal features for emotional EEG data augmentation". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 31 (2023), pp. 2018–2027.

[107] Yingying Jiao et al. "Driver sleepiness detection from EEG and EOG signals using GAN and LSTM networks". In: *Neurocomputing* 408 (2020), pp. 100–1. ISSN: 0925-2312. DOI: `https://doi.org/10.1016/j.neucom.2019.05.108`.

[108] Seif Eldawlatly. "On the role of generative artificial intelligence in the development of brain-computer interfaces". In: *BMC Biomedical Engineering* 6.1 (2024), p. 4.

[109] Yudie Hu et al. "Human-robot collaborative disassembly enabled by brainwaves and improved generative adversarial network". In: *Advanced Engineering Informatics* 60 (2024), p. 102434.

[110] Ettore Cinquetti et al. “EEG-driven GAN for Alpha Rhythm Generation in Passive BCI”. In: *2024 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE)*. 2024, pp. 389–394.

[111] Ahmed G Habashi et al. “Generative adversarial networks in EEG analysis: an overview”. In: *Journal of neuroengineering and rehabilitation* 20.1 (2023), p. 40.

[112] Cristóbal Esteban, Stephanie L Hyland, and Gunnar Rätsch. “Real-valued (medical) time series generation with recurrent conditional gans”. In: *arXiv preprint arXiv:1706.02633* (2017).

[113] Jinsung Yoon, Daniel Jarrett, and Mihaela Van der Schaar. “Time-series generative adversarial networks”. In: *Advances in neural information processing systems* 32 (2019).

[114] Farhad Mortezapour Shiri et al. “A comprehensive overview and comparative analysis on deep learning models: CNN, RNN, LSTM, GRU”. In: *arXiv preprint arXiv:2305.17473* (2023).

[115] Atiye Sadat Hashemi et al. “Time-series anonymization of tabular health data using generative adversarial network”. In: *2023 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2023, pp. 1–8.

[116] Cemal Öztürk. “Enhancing financial time-series analysis with timegan: A novel approach”. In: *2024 9th International Conference on Computer Science and Engineering (UBMK)*. IEEE. 2024, pp. 447–450.

[117] Peihao Tang et al. “Time series data augmentation for energy consumption data based on improved timegan”. In: *Sensors* 25.2 (2025), p. 493.

[118] Rama Cont et al. “Tail-gan: Learning to simulate tail risk scenarios”. In: *Management Science* (2025).

[119] Ling Chen. “Risk Management with Feature-Enriched Generative Adversarial Networks (FE-GAN)”. In: *arXiv preprint arXiv:2411.15519* (2024).

[120] J Qian et al. “TimeLDM: Latent Diffusion Model for Unconditional Time Series Generation. arXiv 2024”. In: *arXiv preprint arXiv:2407.04211* (2024).

[121] Ettore Cinquetti et al. “EEG-driven GAN for Alpha Rhythm Generation in Passive BCI”. In: *2024 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE)*. IEEE. 2024, pp. 389–394.

[122] Lin Long et al. “On llms-driven synthetic data generation, curation, and evaluation: A survey”. In: *arXiv preprint arXiv:2406.15126* (2024).

[123] Martin Heusel et al. “GANs trained by a two time-scale update rule converge to a local nash equilibrium”. In: *Advances in neural information processing systems* 30 (2017).

[124] Mastaneh Torkamani-Azar et al. “Prediction of Reaction Time and Vigilance Variability From Spatio-Spectral Features of Resting-State EEG in a Long Sustained Attention Task”. In: *IEEE Journal of Biomedical and Health Informatics* 24.9 (2020), pp. 2550–2558.

[125] Tim R Mullen et al. “Real-time neuroimaging and cognitive monitoring using wearable dry EEG”. In: *IEEE transactions on biomedical engineering* 62.11 (2015), pp. 2553–2567.

[126] Hans Berger. “Über das elektrenkephalogramm des menschen”. In: *DMW-Deutsche Medizinische Wochenschrift* 60.51 (1934), pp. 1947–9.

[127] Wiremu Hohaia et al. “Occipital alpha-band brain waves when the eyes are closed are shaped by ongoing visual processes”. In: *Scientific reports* 12.1 (2022), p. 1194.

[128] Soroush Mehri et al. “SampleRNN: An unconditional end-to-end neural audio generation model”. In: *arXiv preprint arXiv:1612.07837* (2016).

[129] Serkan Kiranyaz et al. “1D convolutional neural networks and applications: A survey”. In: *Mechanical systems and signal processing* 151 (2021), p. 107398.

[130] Tiantian Fang, Ruoyu Sun, and Alex Schwing. "DigGAN: Discriminator gradient gap regularization for GAN training with limited data". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 31782–31795.

[131] Frédéric Dehais et al. "Monitoring pilot's mental workload using ERPs and spectral power with a six-dry-electrode EEG system in real flight conditions". In: *Sensors* 19.6 (2019), p. 1324.

[132] Jiahui Xu and Baichang Zhong. "Review on portable EEG technology in educational research". In: *Computers in Human Behavior* 81 (2018), pp. 340–9.

[133] Minchao Wu et al. "A study on the combination of functional connection features and Riemannian manifold in EEG emotion recognition". In: *Frontiers in Neuroscience* 17 (2024), p. 1345770.

[134] Chuong H Nguyen, George K Karavas, and Panagiotis Artemiadis. "Inferring imagined speech using EEG signals: a new approach using Riemannian manifold features". In: *Journal of neural engineering* 15.1 (2017), p. 016002.

[135] Nicolas Ivanov and Tom Chau. "Riemannian geometry metric-based visual feedback for BCI user training: towards exploratory learning of motor imagery". In: *Brain-Apparatus Communication: A Journal of Bacomics* 3.1 (2024), p. 2366193. DOI: 10.1080/27706710.2024.2366193.

[136] Imad Eddine Tibermacine et al. "Riemannian geometry-based eeg approaches: A literature review". In: *arXiv preprint arXiv:2407.20250* (2024).

[137] Florian Yger, Maxime Berar, and Fabien Lotte. "Riemannian approaches in brain-computer interfaces: a review". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25.10 (2016), pp. 1753–1762.

[138] Paolo Zanini et al. "Transfer learning: A Riemannian geometry framework with applications to brain–computer interfaces". In: *IEEE Transactions on Biomedical Engineering* 65.5 (2017), pp. 1107–1116.

[139] Sylvain Chevallier et al. "Review of Riemannian distances and divergences, applied to SSVEP-based BCI". In: *Neuroinformatics* 19.1 (2021), pp. 93–106.

[140] Pedro Luiz Coelho Rodrigues, Christian Jutten, and Marco Congedo. "Riemannian procrustes analysis: transfer learning for brain–computer interfaces". In: *IEEE Transactions on Biomedical Engineering* 66.8 (2018), pp. 2390–2401.

[141] He He and Dongrui Wu. "Transfer learning for brain–computer interfaces: A Euclidean space data alignment approach". In: *IEEE Transactions on Biomedical Engineering* 67.2 (2019), pp. 399–410.

[142] Dongrui Wu. "Revisiting Euclidean alignment for transfer learning in EEG-based brain-computer interfaces". In: *Journal of Neural Engineering* (2025).

[143] Jiahui Ying, Qingguo Wei, and Xichen Zhou. "Riemannian geometry-based transfer learning for reducing training time in c-VEP BCIs". In: *Scientific reports* 12.1 (2022), p. 9818.

[144] Zhiwu Huang and Luc Van Gool. "A riemannian network for spd matrix learning". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 31. 1. 2017.

[145] Daniel Wilson et al. "Deep riemannian networks for EEG decoding". In: *arXiv preprint arXiv:2212.10426* (2022).

[146] Georgios Rousis et al. "Combining EEGNet with SPDNet towards an end-to-end architecture for imagined speech decoding". In: *2024 32nd European Signal Processing Conference (EUSIPCO)*. IEEE. 2024, pp. 1531–1535.

[147] Ce Ju and Cuntai Guan. "Graph neural networks on spd manifolds for motor imagery classification: A perspective from the time–frequency analysis". In: *IEEE Transactions on Neural Networks and Learning Systems* (2023).

[148] Igor Carrara and Théodore Papadopoulo. "Classification of BCI-EEG based on the augmented covariance matrix". In: *IEEE Transactions on Biomedical Engineering* 71.9 (2024), pp. 2651–2662.

[149] Igor Carrara et al. “Geometric neural network based on phase space for BCI-EEG decoding”. In: *Journal of Neural Engineering* 22.1 (2025), p. 016049.

[150] Igor Carrara and Theodore Papadopoulo. “Enhancing Computational Efficiency of Motor Imagery BCI Classification with Block-Toeplitz Augmented Covariance Matrices and Siegel Metric”. In: *arXiv preprint arXiv:2406.16909* (2024).

[151] Alexander Craik, Yilei He, and Jose L Contreras-Vidal. “Deep learning for electroencephalogram (EEG) classification tasks: a review”. In: *Journal of Neural Engineering* 16.3 (2019), p. 031001.

[152] Robin Tibor Schirrmeister et al. “Deep learning with convolutional neural networks for EEG decoding and visualization”. In: *Human brain mapping* 38.10 (2017), pp. 5391–5420.

[153] Tim Salimans et al. “Improved techniques for training gans”. In: *Advances in neural information processing systems* 29 (2016).

[154] Benyamin Ghojogh et al. “Uniform manifold approximation and projection (UMAP) and its variants: tutorial and survey”. In: *arXiv preprint arXiv:2109.02508* (2021).

[155] Leland McInnes, John Healy, and James Melville. “Umap: Uniform manifold approximation and projection for dimension reduction”. In: *arXiv preprint arXiv:1802.03426* (2018).

[156] Alexandre Barachant et al. *pyRiemann*. Version v0.9. July 2025. DOI: 10.5281/zenodo.593816. URL: https://doi.org/10.5281/zenodo.593816.

[157] Isaac A Corley and Yufei Huang. “Deep EEG super-resolution: Upsampling EEG spatial resolution with generative adversarial networks”. In: *2018 IEEE EMBS international conference on biomedical & health informatics (BHI)*. IEEE. 2018, pp. 100–103.

[158] Yeganeh Farahzadi, Morteza Ansarinia, and Zoltan Kekecs. “YARE-GAN: Yet Another Resting State EEG-GAN”. In: *arXiv preprint arXiv:2503.02636* (2025).

[159] Zhishui You et al. “Virtual Electroencephalogram Acquisition: A Review on Electroencephalogram Generative Methods”. In: *Sensors (Basel, Switzerland)* 25.10 (2025), p. 3178.

[160] Kun Chen et al. “EEG-based mental fatigue detection using linear prediction cepstral coefficients and Riemann spatial covariance matrix”. In: *Journal of Neural Engineering* 19.6 (2022), p. 066021.

[161] Byung Hyung Kim et al. “A discriminative SPD feature learning approach on Riemannian manifolds for EEG classification”. In: *Pattern Recognition* 143 (2023), p. 109751.

[162] Alexandre Barachant et al. “Classification of covariance matrices using a Riemannian-based kernel for BCI applications”. In: *Neurocomputing* 112 (2013), pp. 172–178.

[163] Sara Omari, Adil Omari, and Mohamed Abderrahim. “Multiple tangent space projection for motor imagery EEG classification”. In: *Applied Intelligence* 53.18 (2023), pp. 21192–21200.

[164] Lukas PA Arts and Egon L Van den Broek. “The fast continuous wavelet transformation (fCWT) for real-time, high-quality, noise-resistant time–frequency analysis”. In: *Nature Computational Science* 2.1 (2022), pp. 47–58.

[165] Haoyi Zhou et al. “Informer: Beyond efficient transformer for long sequence time-series forecasting”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 35. 12. 2021, pp. 11106–11115.

[166] Haixu Wu et al. “Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting”. In: *Advances in neural information processing systems* 34 (2021), pp. 22419–22430.

[167] Yuqi Nie et al. “A time series is worth 64 words: Long-term forecasting with transformers”. In: *arXiv preprint arXiv:2211.14730* (2022).

[168] Shriyank Somvanshi et al. “A survey on structured state space sequence (s4) models”. In: *arXiv e-prints* (2025), arXiv–2503.

[169] Albert Gu and Tri Dao. "Mamba: Linear-time sequence modeling with selective state spaces". In: *First conference on language modeling*. 2024.

[170] Ziwei Wang et al. "Channel reflection: Knowledge-driven data augmentation for EEG-based brain–computer interfaces". In: *Neural Networks* 176 (2024), p. 106351.

[171] Ziwei Wang et al. "Time–frequency transform based EEG data augmentation for brain–computer interfaces". In: *Knowledge-Based Systems* 311 (2025), p. 113074.

[172] Gideon Vos et al. "A Statistical Approach for Synthetic EEG Data Generation". In: *arXiv preprint arXiv:2504.16143* (2025).

[173] Zhijiang Wan et al. "Data generation for enhancing EEG-based emotion recognition: Extracting time-invariant and subject-invariant components with contrastive learning". In: *IEEE Transactions on Consumer Electronics* (2024).

[174] FP Carrle, Y Hollenbenders, and A Reichenbach. *Generation of synthetic EEG data for training algorithms supporting the diagnosis of major depressive disorder*. 2023.

[175] Yuheng Chen et al. "FusionGen: Feature fusion-based few-shot EEG data generation". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2025, pp. 351–360.

[176] Yufan Yi et al. "A weighted co-training framework for emotion recognition based on EEG data generation using frequency-spatial diffusion transformer". In: *IEEE Transactions on Affective Computing* 15.4 (2024), pp. 2055–2069.

[177] Jan Sosulski and Michael Tangermann. "Introducing block-Toeplitz covariance matrices to remaster linear discriminant analysis for event-related potential brain–computer interfaces". In: *Journal of neural engineering* 19.6 (2022), p. 066001.

[178] Xuan Son Nguyen, Aymeric Histace, and Nistor Grozavu. "Siegel Neural Networks". In: *arXiv preprint arXiv:2511.09577* (2025).

[179] Tao Ding, Tom MW Nye, and Yujiang Wang. "Manifold-valued models for analysis of EEG time series data". In: *Computational Statistics & Data Analysis* 209 (2025), p. 108168.

[180] Agnieszka Wosiak, Anita Tereszczuk, and Katarzyna Żykwińska. "Determining Levels of Affective States with Riemannian Geometry Applied to EEG Signals". In: *Applied Sciences* 15.19 (2025), p. 10370.

[181] Joseph Paillard, Jörg F Hipp, and Denis A Engemann. "GREEN: A lightweight architecture using learnable wavelets and Riemannian geometry for biomarker exploration with EEG signals". In: *Patterns* 6.3 (2025).

[182] Emmanuel Kalunga, Sylvain Chevallier, and Quentin Barthélemy. "Data augmentation in Riemannian space for brain-computer interfaces". In: *STAMLINS*. 2015.

[183] Xuan Kan et al. "R-mixup: Riemannian mixup for biological networks". In: *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2023, pp. 1073–1085.

[184] Yaxing Wang et al. "Transferring gans: generating images from limited data". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 218–234.

[185] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. "Freeze the discriminator: a simple baseline for fine-tuning gans". In: *CVPR AI for Content Creation Workshop*. 2020.

[186] Atsuhiro Noguchi and Tatsuya Harada. "Image generation from small datasets via batch statistics adaptation". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 2750–2758.

[187] Arturs Berzins et al. "Geometry-informed neural networks". In: *arXiv preprint arXiv:2402.14009* (2024).

[188] Liyue Shen et al. "A geometry-informed deep learning framework for ultra-sparse 3D tomographic image reconstruction". In: *Computers in Biology and Medicine* 148 (2022), p. 105710.

[189] Shujin Hu, Huafeng Liu, and Huihui Ye. “Physiology and geometry-informed deep learning framework for 12-lead ecg cardiac transmembrane potential recovery”. In: *Proceedings of the 2024 International Conference on Innovation in Artificial Intelligence*. 2024, pp. 124–130.