International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Unsupervised Generative Modeling of Sleep EEG Feature Distributions using Wasserstein GANs

Mandakini Keshavrao Jadhwar
Dept of EDT, National Institute of Electronics & Information Technology Aurangabad, Maharashtra, India.

Dr. Anirban Jyoti Hati
(Scientist 'C')
Dept of EDT, National Institute of Electronics & Information Technology Aurangabad, Maharashtra, India.

Saurabh Bansod
(Scientist 'C')
Dept of EDT, National Institute of Electronics & Information Technology Aurangabad, Maharashtra, India.

Shashank Singh
(Scientist 'B')
Dept of EDT, National Institute of Electronics & Information Technology Aurangabad, Maharashtra, India.

Abstract -- **The deep learning architectures for automatic sleep staging have approached expert-level performance, they remain contingent on large annotated corpora and perpetuate the class imbalance inherent in sleep stage prevalence distributions constraints that motivate a label-free approach to distributional modelling. This thesis investigates whether a Wasserstein generative adversarial network with gradient penalty (WGAN-GP) can learn to approximate the population-level feature distribution of sleep EEG in a fully unsupervised manner, without access to stage annotations at any point in training .The model is trained on spectral and time-domain features extracted from the Sleep-EDF Expanded database 197 whole-night polysomnographic recordings spanning a heterogeneous population aged 25 to 101 and evaluated on six held-out subjects drawn from both sub-studies. Two feature configurations are compared: a seven-dimensional set (v1) of spectral band powers, entropy, and amplitude descriptors, and a ten-dimensional extension (v2) that additionally incorporates Hjorth and line-length features. The results demonstrate that the WGAN-GP learns a generative model whose synthetic feature distributions bear statistically meaningful similarity to those of unseen test subjects, with v1 outperforming v2 on both primary metrics: v1 achieves lower MMD^2 in four of six test recordings and lower maximum KS statistic in all six a unanimous result corresponding to an approximate 32% reduction in worst-case distributional divergence relative to v2. UMAP visualizations corroborate these findings, showing stronger manifold interleaving between real and synthetic samples under v1 in both per-recording and aggregate analyses.**

Keywords - EEG, Sleep Analysis, Unsupervised Learning, GANs, WGAN-GP, UMAP, MMD

I. INTRODUCTION

Sleep is a fundamental biological process whose disruption carries measurable consequences for physical health, cognitive function, and psychological wellbeing. Chronic insomnia, the most prevalent sleep disorder, affects an estimated ten to fifteen percent of the general adult population worldwide, with well-documented associations with cardiovascular disease, metabolic syndrome, depression, and all-cause mortality — and an annual economic cost, expressed in lost productivity and increased healthcare utilization, running to tens of billions of dollars in high-income economies alone [3]. Clinical assessment of sleep architecture relies on polysomnography (PSG), a multimodal overnight recording in which a trained specialist manually assigns each thirty-second EEG epoch to a sleep stage. The taxonomy codified by Rechtschaffen and Kales [1] and subsequently revised by the American Academy of Sleep Medicine [2] defines the five-class W–N1–N2–N3–REM scheme that remains universally adopted in both clinical practice and research.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Data augmentation has been proposed as one mitigation, but conventional strategies — noise injection, time-shifting, and feature interpolation — offer a limited ceiling [14]. Generative synthesis, which learns the data distribution rather than perturbing existing samples, represents the

The central objective is to investigate whether a Wasserstein GAN with gradient penalty (WGAN-GP), trained in a fully unsupervised manner on the Sleep-EDF Expanded database [21,22] — 197 whole-night PSG recordings spanning a heterogeneous population aged 25 to 101, recorded under both ambulatory and hospital conditions — can learn a generative model of sleep EEG feature distributions that achieves meaningful statistical similarity to held-out subjects never seen during training. Two feature configurations are compared: a seven-dimensional set (v1) comprising relative spectral band powers in the delta, theta, alpha, beta, and gamma bands, spectral entropy [25], and RMS amplitude; and a ten-dimensional extension (v2) that adds Hjorth mobility, complexity [23], and signal line length. Distributional similarity is quantified using the maximum mean discrepancy with RBF kernel [19], per-feature Kolmogorov-Smirnov tests with Benjamini-Hochberg correction, and UMAP [20] visualization of embedding overlap across both individual recordings and the aggregate test set.

The principal contributions of this work are fourfold. First, an end-to-end reproducible pipeline for unsupervised sleep EEG distributional modelling is implemented, spanning MNE-Python preprocessing, spectral and time-domain feature extraction, Standard Scaler normalization, WGAN-GP training, and multi-metric distributional evaluation. Second, a controlled ablation between the two feature set dimensionalities reveals that the compact v1 set achieves lower distributional divergence than v2 on the majority of held-out recordings — a finding whose mechanistic basis is developed in Chapter 5. Third, the systematic over-dispersion of generated feature distributions relative to real distributions is characterized and explained as a structural consequence of unconditional generation learning population-level variance rather than within-subject variance. Fourth, UMAP visualization confirms the distributional findings qualitatively and reveals that the degree of manifold interleaving between real and synthetic samples mirrors the quantitative metric ranking between v1 and v2.

II. LITERATURE REVIEW

Before the rise of deep learning, automated sleep staging relied on hand-crafted features — relative spectral band powers, Hjorth parameters, spectral entropy — fed to classical classifiers such as support vector machines and linear discriminant analysis. The transition to deep learning resolved the feature engineering bottleneck by allowing models to learn discriminative representations directly from labelled data, and the resulting performance gains have been substantial and consistent across independent benchmark evaluations.

DeepSleepNet [4] was among the first architectures to demonstrate that a single-channel EEG recording is sufficient for near-expert automatic staging when paired with a sufficiently expressive model. Its two-branch convolutional extractor learned features at both fine and coarse temporal scales before passing the resulting representations to a bidirectional LSTM that captured inter-epoch transition dynamics. This

cohorts could generalize to unseen recording sites and electrode configurations, a result that highlighted both the power of large multi-site training sets and the residual difficulty of cross-dataset generalization. XSleepNet [5] took a different route, showing that jointly learning from raw EEG waveforms and time-frequency spectrograms — and adaptively weighting the two views based on their informativeness for a given epoch — produced state-of-the-art figures on the full Sleep-EDF-153 benchmark.

Sleep EEGNet [7] addressed this directly through a sequence-to-sequence framework with focal loss and targeted augmentation, reporting substantial F1 improvements on minority classes, but acknowledged that the ceiling on such improvements is ultimately set by the scarcity of real minority-class examples in the training set. These limitations collectively motivate the investigation of generative approaches that can model the sleep EEG distribution without requiring any labels at all.

The generative adversarial network, introduced by Goodfellow et al. [8], established a new paradigm for learning generative models from data. Rather than maximising a likelihood objective explicitly, the GAN framework trains two networks simultaneously: a generator G that maps samples from a simple prior distribution to the data space, and a discriminator D that attempts to distinguish real samples from generated ones. The generator is rewarded for fooling the discriminator, and the discriminator is penalised for being fooled, driving both toward an equilibrium in which G's output distribution matches the true data distribution.. The result is mode collapse, where the generator learns to produce

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

a small subset of the target distribution rather than its full diversity.

Arjovsky et al. [9] addressed this fundamental instability by replacing the Jensen-Shannon divergence with the Wasserstein-1 distance, also known as the Earth Mover's Distance. The Wasserstein distance measures the minimum cost of transporting mass from one distribution to another and remains finite and differentiable even when the two distributions have disjoint support — properties that the Jensen-Shannon divergence does not share. The resulting WGAN critic, which estimates the Wasserstein distance rather than a classification probability, produces loss curves that correlate meaningfully with sample quality throughout training, making training progress interpretable in a way that was not possible with the original GAN objective. The Wasserstein distance requires the critic function to satisfy a Lipschitz continuity constraint, which the original WGAN enforced through weight clipping — a mechanism that proved effective but introduced pathological gradient behavior, limiting the critic's representational capacity and slowing convergence. Gulrajani et al. [10] replaced weight clipping with a gradient penalty computed at points interpolated between real and generated samples, directly regularising the critic's gradient norm where it matters most..

Alternative generative frameworks exist and merit brief consideration. The variational autoencoder (VAE) [17] learns a probabilistic encoder-decoder in which the latent space is regularised toward a Gaussian prior, enabling both generation and structured latent space interpolation. VAEs are more stable to train than GANs and provide an explicit likelihood bound, but are known to produce over-smoothed samples because the reconstruction objective penalises all deviations from the training data equally, regardless of perceptual or distributional importance. Denoising diffusion probabilistic models [18] have more recently achieved state-of-the-art generative quality in image and audio synthesis by learning to reverse a gradual Gaussian noise process, but their sequential sampling procedure is computationally expensive and their advantage over well-tuned adversarial models on low-dimensional feature distributions — as opposed to high-dimensional raw signals — has not been established. For the compact, low-dimensional feature-space setting of this thesis, the WGAN-GP offers the best combination of distributional fidelity, training stability, and computational tractability.

In the domain of sleep EEG specifically, Zhang et al. [13] proposed Sleep EGAN, a conditional GAN trained to generate synthetic thirty-second EEG epochs conditioned on sleep stage labels. The primary motivation was class imbalance correction: by generating additional N1 and N3 epochs and augmenting training batches before classification, Sleep EGAN achieved substantial F1-score improvements on minority classes across several staging backbones evaluated on the Sleep-EDF benchmark. Sleep EGAN is the closest direct precursor to this thesis in terms of dataset and application domain. Its key limitation, from the perspective of the present work, is the conditioning on stage labels: the model cannot be deployed in settings where labels are unavailable, and its generative quality is evaluated only instrumentally — through downstream classifier performance — rather than in terms of distributional fidelity as an end in itself. The broader augmentation landscape for EEG deep learning, surveyed by Lashgari et al. [14], similarly shows that virtually all generative augmentation strategies are task-conditioned and evaluated through their effect on classification accuracy, leaving the question of unsupervised distributional modelling largely unaddressed.

Panwar et al. [15] provided the most methodologically relevant precedent for the present work by applying a Wasserstein GAN to model EEG feature distributions — rather than raw waveforms — in a rapid serial visual presentation (RSVP) event detection context.

III. METHODS AND MATERIALS

The overall processing pipeline is summarized in Figure 3.1

Figure 3.1: End-to-end system pipeline. EEG recordings from the Sleep-EDF Expanded Database are preprocessed and transformed into compact feature vectors, which are used to train a WGAN-GP generative model. The trained generator is evaluated against real EEG features from unseen subjects using MMD, KS tests, and UMAP

All EEG recordings used in this study are drawn from the Sleep-EDF Expanded Database (version 1.0.0), hosted on PhysioNet. The database contains 197 whole-night polysomnographic (PSG) recordings collected across two independent studies. In both studies, EEG signals were

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

acquired at 100 Hz from two frontal-to-central electrode pairs, Fpz–Cz and Pz–Oz, alongside horizontal EOG and submental EMG. Corresponding hypnogram files encode expert-annotated sleep stages (W, R, 1, 2, 3, 4, M, ?) scored according to the 1968 Rechtschaffen and Kales manual. Critically, no hypnogram annotations are used at any point in this study; the framework is entirely unsupervised, and stage labels are mentioned here only for contextual reference. Raw EEG signals contain frequency components well outside the range of sleep-relevant neural activity, and expressing 30-second waveforms as feature vectors is necessary to make the generative modelling tractable and interpretable.. The Power Spectral Density (PSD) of each epoch is estimated using Welch's method, with a segment length of 4 seconds and density scaling. Welch's method is preferred over the direct periodogram because averaging across overlapping sub-segments substantially reduces the variance of the spectral estimate — an important consideration when extracting stable features from short epochs. Only the sub-band [0.5, 45] Hz is retained for feature computation, consistent with the filter passband. Five relative spectral band powers are computed, The extracted feature matrices span heterogeneous numerical ranges: relative band powers are bounded in [0, 1], while line length and RMS are expressed in physical units that vary across subjects and nights. To ensure that no single feature dimension dominates the training loss by virtue of its scale, each feature dimension j is standardized to zero mean and unit variance using the statistics of the training data:

$$\tilde{x}_{ij} = \frac{x_{ij} - u_j}{\sigma_j} \qquad (3.3)$$

The mean μ_j and standard deviation σ_j are computed from the training set only and stored alongside the trained model weights. The identical scaler is applied to test recordings at inference time without any re-fitting, which prevents data leakage and ensures that the test distribution is projected into the same normalised space in which the generator was trained.corresponding to the canonical EEG frequency bands associated with different stages of sleep and wakefulness: delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), and gamma (30–45 Hz). The absolute power in each band B is obtained by trapezoidal integration of the PSD over the band's frequency range, and each value is then normalised by the total power across the full analysis band to yield a relative measure:

$$\hat{p}(B) = \frac{\int_B S(f)\,df}{\int_{0.5}^{45} S(f)\,df + \varepsilon} \qquad (3.1)$$

where S(f) is the Welch PSD estimate and $\varepsilon = 10^{-12}$ is a small constant introduced to prevent division by zero on silent epochs.

Spectral entropy is included as a measure of the frequency-domain complexity of the EEG signal. The normalized PSD is treated as a probability mass function over frequency bins, and Shannon entropy is applied to that distribution:

$$LL(x) = \sum_{i=1}^{N-1} |X_{i+1} - X_i| \qquad (3.2)$$

Line length increases with higher frequency content or larger amplitude, making it sensitive to the presence of sleep spindles and other high-frequency transient activity. Together, the Hjorth parameters and line length capture waveform morphology information that spectral features summaries only indirectly. Two feature configurations are defined for the ablation study. Feature set v1 is a 7-dimensional vector composed of the five relative band powers, spectral entropy, and RMS. Feature set v2 extends v1 to 10 dimensions by appending Hjorth mobility, Hjorth complexity, and line length. Both configurations are summarized in Table 3.1.

Table 3.1: Feature definitions for v1 (7-dimensional) and v2 (10-dimensional) configuration

#	Feature	Description	Included in
1	rel_delta	Relative delta power [0.5–4 Hz]	v1 and v2
2	rel_theta	Relative theta power [4–8 Hz]	v1 and v2
3	rel_alpha	Relative alpha power [8–13 Hz]	v1 and v2
4	rel_beta	Relative beta power [13–30 Hz]	v1 and v2
5	rel_gamma	Relative gamma power [30–45 Hz]	v1 and v2
6	entropy	Spectral entropy of the normalised PSD	v1 and v2
7	rms	Root Mean Square signal amplitude	v1 and v2
8	hjorth_mobility	Hjorth Mobility — proxy for mean signal frequency	v2 only
9	hjorth_complexity	Hjorth Complexity — waveform shape irregularity	v2 only
10	line_length	Line Length — cumulative amplitude variation	v2 only

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

2.3. Generative Framework: WGAN-GP

The generative model at the heart of this study is a Wasserstein Generative Adversarial Network with Gradient Penalty (WGAN-GP). Standard GANs optimise the Jensen–Shannon divergence between the real data distribution P_r and the generator distribution P_g. When these two distributions have disjoint supports — which is common early in training, before the generator has learned anything useful — the Jensen–Shannon divergence saturates to a constant, providing no gradient signal to the generator. The result is training instability and mode collapse, where the generator maps all latent codes onto a small subset of the target distribution. The WGAN replaces the Jensen–Shannon divergence with the Wasserstein-1 distance (also known as the Earth Mover's Distance), which remains informative and smooth even when the two distributions are disjoint. By the Kantorovich–Rubinstein duality, it can be expressed as:

$$W(P_r, P_g) = \sup_{\|f\|_L \leq 1} [\mathbb{E}_{x\sim} p_r[f(x)] - \mathbb{E}_{x\sim} p_g[f(x)]] \quad (3.4)$$

where the supremum is taken over all functions f satisfying the 1-Lipschitz condition. In practice, this function is parameterised by the critic network (a terminology preferred over "discriminator" to reflect the absence of a sigmoid output). The original WGAN enforced the Lipschitz constraint by clipping critic weights, but this approach is known to produce degenerate gradients and constrain the critic to pathologically simple functions. The WGAN-GP replaces weight clipping with a gradient penalty that directly regularises the critic's gradient norm at interpolated points between real and generated samples, yielding more stable training and a better-conditioned critic throughout the learning process.

The model consists of two multilayer perceptrons: a generator G and a critic C. The architecture is illustrated in Figure 3.2.

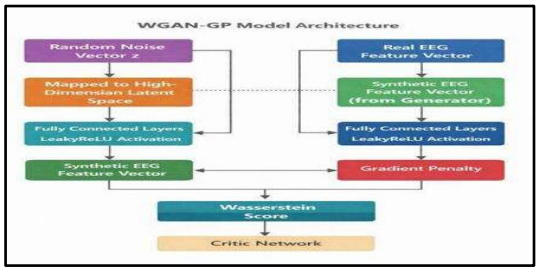

Figure 3.2: WGAN-GP model architecture. A random latent vector z is mapped through the generator's fully connected layers to produce a synthetic EEG feature vector. Both the synthetic and real EEG feature vectors are independently passed through the critic network, which outputs an unbounded Wasserstein score. A gradient penalty term, computed at interpolated points between real and fake samples, enforces the 1-Lipschitz constraint during critic training.

The generator maps a latent noise vector $z \in \mathbb{R}^{32}$ — drawn from an isotropic Gaussian prior $z \sim \tau(0, I)$ — to a synthetic feature vector $G(z) \in \mathbb{R}^{D}$, where $D \in \{7, 10\}$ depending on the feature configuration. It consists of two hidden layers of width 128 with LeakyReLU activations (slope 0.2), followed by a linear output layer with no activation. The latent dimension of 32 was chosen to provide sufficient expressiveness for a low-dimensional target space while keeping the generative mapping compact; the absence of an output activation allows the generator to produce unbounded values in the standardised feature space without any range restriction. The critic maps any feature vector — real or synthetic — to a scalar Wasserstein score. Its architecture mirrors the generator's hidden structure, using two 128-unit hidden layers with LayerNorm followed by LeakyReLU activations, and a single linear output neuron. LayerNorm is used in place of BatchNorm because the gradient penalty requires per-sample gradient computations: BatchNorm introduces inter-sample dependencies within a mini-batch that would corrupt these gradients, whereas LayerNorm normalises each sample independently and preserves the required independence structure.For each mini-batch of real samples {x_i} and generated samples {G(z_i)}, the gradient penalty is computed at random linear interpolations between real and fake points:

$$\hat{x}_j = \varepsilon_i \cdot x_i + (1 - \varepsilon_i).\, G(z_i), \quad \varepsilon_i \sim Uniform(0,1) \quad (3.5)$$

The gradient penalty term is then:

$$GP = \mathbb{E}_{\hat{x}}[(\|\nabla_{\hat{x}} C(\hat{x})\|_2 - 1)^2] \quad (3.6)$$

where $\nabla_{\{\hat{x}\}} C(\hat{x})$ is the gradient of the critic with respect to its input, computed by automatic differentiation. This penalty encourages the critic's gradient norm to remain close to unity along the interpolation path, which is the region of feature space most relevant for enforcing the Lipschitz constraint.

The full critic loss combines the Wasserstein distance estimate with the gradient penalty:

$$\mathcal{L}_c = \mathbb{E}\big[C\big(G(z)\big)\big] - \mathbb{E}[C(x)] + \lambda\, .\, GP \quad (3.7)$$

and the generator loss is simply the negated critic score on generated samples:

$$\mathcal{L}_c = -\mathbb{E}\big[C\big(G(z)\big)\big] \quad (3.8)$$

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

The penalty coefficient $\lambda = 10$ follows the recommendation of Gulrajani et al. and has been widely validated across applications. Both networks are optimized with Adam at a learning rate of 1×10^{-4} and momentum parameters $\beta_1 = 0.0$, $\beta_2 = 0.9$. Setting $\beta_1 = 0$ (rather than the conventional 0.9) is recommended for WGAN training to reduce the momentum-induced oscillations that can destabilize the Wasserstein loss landscape. The critic is updated five times per generator step, a standard asymmetric schedule that ensures the critic provides a near-optimal Wasserstein score estimate before each generator update. Mini-batches smaller than 8 samples are skipped to avoid numerical instability in the gradient penalty. The training data are reshuffled at the start of each epoch by random permutation. All training hyperparameters are summarised in Table 3.2.

Table 3.2: WGAN-GP training hyperparameters.

Hyperparameter	Value	Justification
Latent dimension z	32	Compact but expressive for 7–10-dim. target
Hidden width (G and C)	128	Matched capacity in both networks
Critic updates per G step	5	Ensures near-optimal critic before each G update
Batch size	256	Stable gradient estimates; large relative to D
Training epochs	30	Convergence observed within this range
Learning rate	1×10^{-4}	Recommended for WGAN-GP
Adam β_1	0.0	Avoids momentum instability in Wasserstein training
Adam β_2	0.9	Standard second-moment decay
Gradient penalty λ	10	Recommended value from Gulrajani et al.
Critic normalisation	Layer Norm	Compatible with per-sample GP computation
Activation (G and C)	Leaky ReLU (0.2)	Avoids dying neurons in both networks
Random seed	42	Fixed across NumPy and PyTorch for reproducibility

Two independent WGAN-GP models are trained: one on the v1 feature matrix and one on v2. Each trained model is persisted as three artefacts: the generator state dictionary (generator.pth), the fitted normalisation scaler (scaler.pkl), and a configuration file recording all relevant hyperparameters and feature names (config.json).

3.4 Evaluation Framework

of unseen subjects? Three complementary methods are employed — a global kernel-based distance, per-feature marginal tests, and a manifold visualisation —

At test time, real feature vectors are extracted from each held-out recording using the identical pipeline described in Section 3.2, scaled with the training-fitted scaler, and capped at 2,000 samples per recording. An equal number of synthetic samples is then drawn from the trained generator:

$$X_{fake} = G(Z), \quad Z \sim N(0, I_{32}), \quad Z \in \mathbb{R}^{n*32} \tag{3.9}$$

where n is the number of real epochs from the test recording (up to 2,000). These two sample sets, X_{real} and X_{fake}, are then compared using the metrics described below.

3.4.1 Maximum Mean Discrepancy

Maximum Mean Discrepancy (MMD) is a kernel-based statistic that measures the distance between two distributions by comparing their mean embeddings in a Reproducing Kernel Hilbert Space. An unbiased estimate of MMD^2 between samples X = {x_i}*{i=1}^n and Y = {y_j}*{j=1}^m is given by:

$$MMD^2(X,Y) = \frac{\sum_{i\neq i'} k(x_i,x_{i'})}{n(n-1)} + \frac{\sum_{j\neq j'} k(y_j,y_{j'})}{m(m-1)} - \frac{2\Sigma_{i,j} k(x_i,y_j)}{nm} \tag{3.10}$$

where the first two sums exclude the diagonal (unbiased estimation) and the third sum covers all cross-pairs. The RBF kernel is used throughout:

$$k(x,y) = \exp\left(-\frac{\|k-y\|_2^2}{2\sigma 2}\right) \tag{3.11}$$

The bandwidth σ is set using the median heuristic: it is taken as the median pairwise Euclidean distance among a random subsample of up to 2,000 points drawn from the pooled real and fake samples. This data-adaptive choice avoids manual bandwidth tuning and scales naturally with the local

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

geometry of the feature space. A lower MMD² indicates closer distributional alignment, with MMD² = 0 if and only if the two distributions are identical (for a characteristic kernel).

3.4.2 Kolmogorov–Smirnov Tests with FDR Correction

MMD provides a global distributional distance but does not reveal which individual features account for any observed divergence. Two-sample Kolmogorov–Smirnov (KS) tests are therefore applied separately to each feature dimension, comparing the empirical CDFs of the real and generated samples:

$$KS_j = sup\left|F_{real.j}(t) - F_{fake,j}(t)\right| \qquad (3.12)$$

where F_{real,j} and F_{fake,j} are the empirical cumulative distribution functions for feature dimension j. A KS statistic of zero indicates perfect marginal agreement, while a value of one indicates complete separation. Because D simultaneous tests are performed (D ∈ {7, 10}), raw p-values are corrected using the Benjamini–Hochberg (BH) procedure at significance level α = 0.05. The BH correction controls the false discovery rate by rejecting the null hypotheses for ranks j = 1, …, k where:

$$K = \max\{j: p_{(j)} \leq \frac{j}{D}.\alpha\} \qquad (3.13)$$

and p_{(j)} denotes the j-th smallest p-value. Three statistics are reported per test recording: the maximum KS statistic across all features (KS_max), the mean KS statistic (KS_mean), and the count of features with statistically significant divergence after correction (#sig).

3.4.3 UMAP Manifold Visualisation

Uniform Manifold Approximation and Projection (UMAP) is used to project the feature vectors into two dimensions, providing a qualitative view of how well the generated distribution covers the real EEG feature manifold. Real and synthetic samples are concatenated and jointly embedded, colour-coded by origin, so that spatial overlap in the projection reflects distributional alignment. Prior to embedding, PCA is applied to reduce the joint matrix to min(6, D) components. This pre-reduction step is standard practice before UMAP on small-dimensional data and helps to suppress noise while speeding up the k-nearest-neighbour graph construction. The UMAP embedding uses n_neighbors = 30 (balancing local and global structure preservation) and min_dist = 0.05 (producing tightly clustered embeddings that make cluster separation visible). Two types of UMAP are generated: single-recording plots for the best- and worst-performing test subjects (as ranked by MMD²), and an aggregate plot constructed by pooling up to 600 real and 600 synthetic samples from each test recording. The aggregate projection reveals whether the generator captures the global sleep EEG feature manifold across the entire test population, rather than simply fitting individual recordings.

3.4.4 Ablation Study Design

The two trained models — one on v1 features, one on v2 — are evaluated on identical test recordings using the identical evaluation pipeline described above. The comparison is intentionally controlled: the only difference between the two experimental conditions is the feature configuration; the model architecture, hyperparameters, training procedure, and evaluation protocol are all identical. Per-recording scatter plots of v1 versus v2 metrics (MMD² and KS_max) and a summary bar chart of means across all test files are used to characterise consistent performance differences. The central hypothesis motivating the ablation is that spectral features, being derived from the power spectrum and thus reflecting broad rhythmic patterns rather than individual waveform shapes, should generalise more uniformly across subjects than morphological features such as the Hjorth parameters, which are known to exhibit stronger inter-individual variability.

IV. RESULTS

Figures 4.1 and 4.2 show the epoch-wise generator and critic loss curves for v1 and v2 respectively, recorded over 30 training epochs. In WGAN-GP, the critic loss approximates the negative Wasserstein-1 distance between the real and generated distributions, and the generator loss is the negated expected critic score on synthetic samples. A rising generator loss therefore does not indicate training failure; it indicates that the critic is progressively assigning lower scores to synthetic samples, which in turn provides stronger gradient signal to drive the generator toward the real distribution. Convergence of the critic loss toward zero reflects a well-calibrated approximate Wasserstein estimator, with the gradient penalty term enforcing the Lipschitz constraint throughout.

For v1 (Figure 4.1), the generator loss begins at approximately 3.8 and rises steeply over the first eight epochs, reaching a plateau in the range 8.0 to 8.6 by around epoch ten. It then remains broadly stable through epoch 25, followed by a modest downward drift to approximately 7.5 at epoch 29. The critic loss starts near −1.5, recovers rapidly during the first five epochs, and stabilises in the narrow range [−0.3, 0.0] for the remainder of training. Neither curve

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

exhibits the sudden collapse nor the unbounded divergence that would indicate mode collapse or training instability. The trajectory for v1 is smooth and consistent throughout all 30 epochs.

The v2 training curve (Figure 4.2) is qualitatively similar but quantitatively distinct. The generator loss begins at approximately 4.1 and rises more steeply and steadily than v1, reaching a higher plateau of approximately 11.5 to 12.0 near epoch 16, after which it remains essentially flat. The critic loss follows the same convergence pattern as v1, starting near −1.5 and settling close to zero by epochs 5 to 8. The higher 4.2 Quantitative Evaluation on Unseen Subjects

The primary quantitative evaluation compares distributional similarity between real and synthetic features across six held-out test recordings. Two complementary metrics are reported throughout. The squared maximum mean discrepancy (MMD²), computed using an RBF kernel with per-recording median heuristic bandwidth, captures joint distributional similarity across the full feature space simultaneously. The per-feature Kolmogorov–Smirnov statistic (KS), reported as both a per-feature maximum (KS max) and a mean across all features (KS mean) after Benjamini–Hochberg false discovery rate correction at $\alpha = 0.05$, captures marginal distributional divergence one feature at a time. These two metrics are complementary: MMD² is sensitive to joint structure and correlations between features, while KS max identifies the single most divergent marginal dimension and is therefore a useful indicator of worst-case per-feature fidelity. Lower values are better for all reported metrics. Figure 4.3 presents the per-recording MMD² for v1 and v2 as a scatter plot, with the y = x diagonal serving as the reference line. Points above the diagonal indicate that v2 incurs higher MMD² than v1 for that recording; points below indicate the reverse. Of the six test recordings, v1 achieves lower MMD² in four. The two recordings where v2 performs better on MMD² both fall in the intermediate difficulty range, with MMD² values of approximately 0.11 to 0.17 for v1. Importantly, v1 dominates at both extremes of the difficulty range: the recordings with the lowest MMD² (approximately 0.05) and highest MMD² (approximately 0.20) both favour v1 by clear margins. The winner tally of v1: 4, v2: 2 on MMD² therefore reflects a consistent but not unconditional advantage for the spectral-only model.

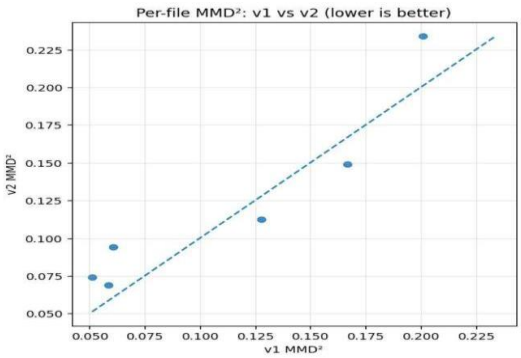

Figure 4.3. Per-recording MMD² for v1 (x-axis) versus v2 (y-axis) across six test subjects. The dashed line is the y = x reference; points above this line indicate v1 achieves lower MMD². v1 outperforms v2 in four of six recordings, with the two v2 wins concentrated in the intermediate MMD² range. Lower is better.

Figure 4.4 presents the corresponding per-recording KS max values. In marked contrast to the MMD² results, v1 achieves lower KS max across all six test recordings without exception, placing every data point above the y = x diagonal. The advantage is frequently substantial in magnitude: for two of the six recordings, the v2 KS max value is approximately 2.6 times and 1.5 times the corresponding v1 value respectively (approximately 0.28 versus 0.74, and approximately 0.54 versus 0.83). Even in the case where the margin is smallest — approximately 0.61 versus 0.62 — v1 remains the lower of the two. The unanimous 6/6 advantage on KS max, combined with the magnitude of several of these margins, is the most decisive quantitative finding of the ablation study, indicating that the addition of morphological features consistently worsens the worst-case marginal distributional agreement on unseen subjects.

generator plateau in v2 — approximately 12 versus approximately 8 in v1 — is consistent with the increased modelling complexity introduced by the ten-dimensional feature space: the addition of morphological features presents a harder distributional target, and the wider Wasserstein distance at convergence reflects this. Both models nevertheless converge to stable solutions without any signs of pathological training behaviour.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

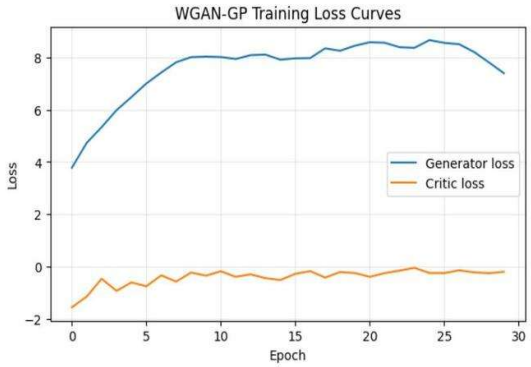

Figure 4.1. WGAN-GP training loss curves for v1 (7-dimensional spectral feature set). Blue: generator loss; Orange: critic loss. The generator loss rises to a stable plateau of approximately 8.0–8.6, while the critic loss converges near zero, indicating a well-calibrated Wasserstein estimator throughout training.

Figure 4.2. WGAN-GP training loss curves for v2 (10-dimensional spectral and morphological feature set). The generator loss reaches a higher plateau of approximately 11.5–12.0, reflecting the greater complexity of the augmented feature space. The critic loss follows the same convergence pattern as v1.

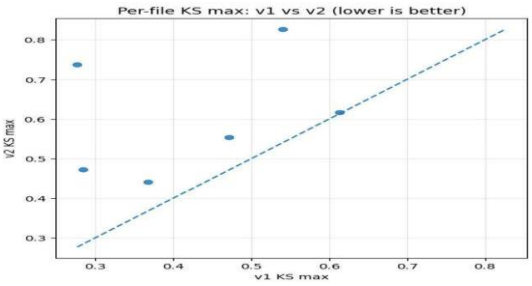

Figure 4.4. Per-recording KS max for v1 (x-axis) versus v2 (y-axis) across six test subjects. All six points lie above the y = x diagonal, indicating that v1 achieves lower KS max in every recording without exception. For two recordings the v2 KS max is approximately 1.5× and 2.6× the corresponding v1 value. Lower is better.

Figure 4.5 summarises the mean values of four evaluation metrics averaged across all six test recordings, with v1 shown in blue and v2 in orange. All four metrics favour v1, though the magnitude of the advantage varies. Mean MMD² is approximately 0.10 for v1 and approximately 0.12 for v2 — a modest but consistent difference. The KS max metric reveals a more substantial gap: v1 averages approximately 0.42 compared to approximately 0.62 for v2, representing a reduction of roughly 32 percent. Mean KS shows a smaller difference (approximately 0.20 for v1 versus approximately 0.22 for v2), suggesting that the primary source of divergence in v2 is concentrated in its worst-performing marginal dimensions rather than distributed uniformly across the feature set.

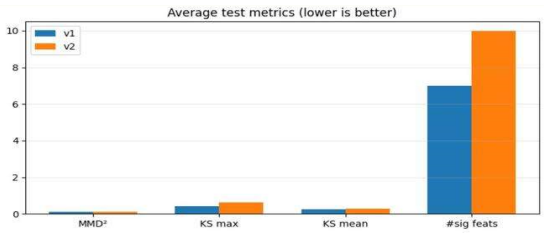

Figure 4.5. Mean evaluation metrics averaged across six test recordings for v1 (blue) and v2 (orange). Lower values are better for all four metrics. v1 is superior on every metric, with the largest absolute advantage on KS max (~0.42 vs ~0.62, a ~32% reduction) and on the number of significantly divergent features (7 vs 10, corresponding to 100% of each model's feature set).

4.3 Manifold Visualisation Using UMAP

Scalar distributional metrics, while informative, do not reveal the geometric relationship between real and synthetic feature distributions. UMAP projections provide a complementary view by embedding the high-dimensional feature space into two dimensions while approximately preserving local neighbourhood structure. For each projection, features were first reduced to min(6, D) principal components before UMAP embedding (n_neighbors = 30, min_dist = 0.05, random_state = 42). Real test epochs are shown in blue and synthetic epochs generated by the trained model in orange. Strong manifold alignment is indicated by thorough interleaving of the two point clouds; systematic spatial separation indicates that the generator places probability mass in regions of feature space that the real distribution does not occupy, or vice versa.

4.3.1 Per-Recording UMAP: Best-Performing Test Subject

Figures 4.6 and 4.7 show the UMAP projections for v1 and v2 respectively on the best-performing test recording by MMD², ST7242J0-PSG.edf. This recording belongs to the Sleep Telemetry sub-study of the Sleep-EDF Expanded dataset, in which participants were monitored in a clinical setting during a pharmacological trial. It is important to note that these projections represent the most favourable case encountered in the test set; the per-recording UMAP for a harder test subject would be expected to show greater divergence between the two clouds.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

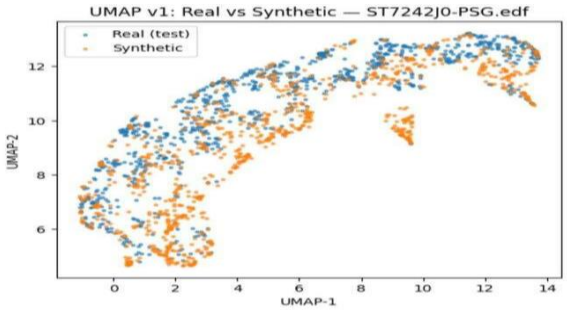

Figure 4.6. UMAP projection for v1 on the best-performing test recording (ST7242J0-PSG.edf, Sleep Telemetry sub-study). Real (blue) and synthetic (orange) samples share a continuous arc-shaped manifold with strong interleaving in the high-density central region. A small isolated real-only cluster is visible near coordinates (9, 9.5). This represents the most favourable alignment case in the test set.

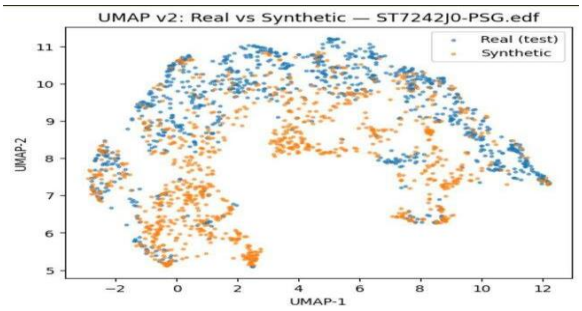

Figure 4.7. UMAP projection for v2 on the same recording (ST7242J0-PSG.edf). The global arc topology is preserved, but local alignment is visibly poorer. Two synthetic-only clusters appear at the base of the arc near coordinates (2, 5.2), and the right arm shows real samples with insufficient synthetic coverage, indicating that the morphological features introduce out-of-distribution mass for this subject.

4.3.2 Aggregate UMAP: All Test Subjects

Figures 4.8 and 4.9 show aggregate UMAP projections computed by pooling 600 samples per recording from all six test files, yielding a population-level view of manifold alignment that captures the diversity of sleep EEG feature distributions across subjects of different ages, recording conditions, and pharmacological states.

For v1 (Figure 4.8), the resulting projection produces a large, complex curved manifold reflecting the inter-subject variability in the test set. Real and synthetic samples are broadly co-distributed across the entire structure, including the long central body, the lower tail, and the upper-right arm. No systematic directional separation of the two point clouds is apparent at the population level; pockets of locally elevated density for one class appear scattered without consistent pattern, consistent with subject-level variability rather than a systematic population-level generator bias.

The v2 aggregate projection (Figure 4.9) shows a similarly curved global manifold, confirming that v2 also reproduces the broad topological structure of the sleep EEG feature space across the test population. However, regions of greater spatial separation between the two clouds are more prominent than in v1. The left edge and lower regions of the manifold in particular contain synthetic clusters with reduced real-sample coverage, suggesting that the v2 generator places probability mass in areas of the feature space that are systematically underrepresented in the real recordings from these test subjects. This population-level observation is consistent with the per-recording KS max values reported in Section 4.2.1 and corroborates the hypothesis that morphological features introduce subject-specific variability that the generator, trained on the full population distribution, cannot adapt to at inference time.

4.4 Feature-Level Distributional Analysis

To examine which features contribute most to the observed aggregate divergence, Figures 4.10 through 4.15 present marginal distribution histograms for each of the six v1 features on ST7242J0-PSG.edf, the best-performing test recording. All values are Z-scored using the training scaler, so the x-axis represents standardised units. Both histograms are density-normalised, making peak heights directly comparable within each panel. This analysis is conducted for the v1 model only; a complete analogous analysis for v2 — which would isolate the specific contribution of the three morphological features to the observed distributional divergence — is identified as a direction for further investigation. As these histograms represent the best-performing subject in the test set, the levels of agreement observed here should be considered an upper bound on per-subject fidelity; recordings with higher MMD² values would be expected to show larger discrepancies.

4.4.1 Well-Matched Features: rel_theta and entropy

The relative theta power distribution (Figure 4.10) shows the closest agreement between real and synthetic samples of all six features. Both distributions are right-skewed and unimodal, with modal values aligning closely at approximately −1.0 on the standardised axis. The peak densities are within approximately 0.05 of one another (real ~0.51, synthetic ~0.46). The synthetic distribution is modestly over-dispersed in the positive tail, extending to approximately 5 compared to the real tail which effectively decays near 3, but the core distributional structure — direction of skew, modal location, and approximate bulk width — is faithfully reproduced. This agreement reflects the relative stability of theta band oscillations across subjects and sleep stages: theta power is associated with both light non-REM sleep and quiet wakefulness, and its moderate ubiquitous presence across the training population gives the generator sufficient regularity to generalise to unseen subjects.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Spectral entropy (Figure 4.11) also shows good qualitative agreement. Both real and synthetic distributions approximate a symmetric bell shape centred near zero, with the real distribution peaked at approximately −0.05 to 0.0 (density ~0.73) and the synthetic distribution shifted marginally rightward to approximately +0.3 (density ~0.60). The synthetic distribution is somewhat broader, extending further in both tails, but the overall shape and central tendency are reproduced reasonably well. The relatively good match for spectral entropy is expected on methodological grounds: entropy integrates power across the entire 0.5–45 Hz spectrum, making it an aggregate summary measure that is less sensitive to subject-specific concentration in any particular frequency band, and hence less variable across subjects than any individual narrowband power feature.

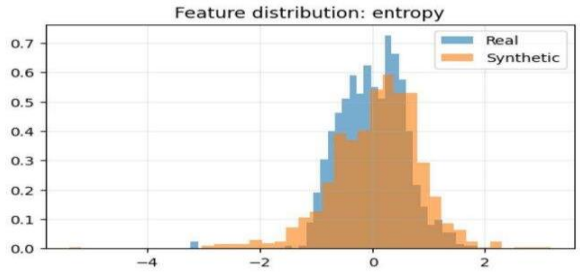

Figure 4.11. Marginal distribution histograms for spectral entropy (v1, best test recording). Both distributions are approximately symmetric and bell-shaped, with closely matching central tendencies. The synthetic distribution is slightly broader and shifted rightward by approximately 0.3 standard deviations.

4.4.2 Moderately Matched Feature: rel_beta

The relative beta power distribution (Figure 4.12) shows an intermediate level of agreement. The real distribution is tightly concentrated in the standardised range −1.5 to 0.0, with a sharp peak of approximately 1.4 at around −0.8. The synthetic distribution shares the same leftward modal location but is substantially flatter (peak density ~0.80) and develops a pronounced right tail extending to approximately 8 on the standardised axis — a spread of roughly 9.5 standard deviations compared to approximately 2.0 for the real distribution. Beta power is primarily elevated during active wakefulness and, to a lesser degree, during REM sleep. The spurious right tail in the synthetic distribution therefore corresponds to wakefulness-associated and arousal-associated epochs drawn from training subjects that are not well-represented in this particular test recording. The modal region is captured correctly, but the generator assigns probability mass to high-beta regions in proportion to their frequency in the training population rather than in this individual subject's recording.

4.3 Poorly Matched Features: rel_alpha, rel_delta, and rel_gamma

The remaining three features show the greatest divergence from the real distribution, each exhibiting a distinct pattern of mismatch with a physiologically interpretable cause.

The relative alpha power distribution (Figure 4.13) displays a sharp, narrow real-data concentration at approximately −0.5 to −0.3 (standardised), with a peak density of approximately 1.13 — the highest peak density of any single feature in this analysis. Alpha activity is strongly suppressed during non-REM sleep; since this test subject spent the substantial majority of the recording in sleep, their alpha power is consistently low and tightly concentrated. The synthetic distribution shares the same modal location but is considerably broader, generating a heavy right tail extending to approximately 7 on the standardised axis. This tail corresponds to wakefulness intervals from training subjects whose recordings included prolonged periods of pre-sleep or post-awakening wakefulness. The tail mass is non-trivial and directly elevates the KS statistic for this feature.

The relative delta power distribution (Figure 4.14) presents a qualitatively different pattern of mismatch. The real distribution is a narrow, concentrated spike in the standardised range 0 to +1.2, with a peak density of approximately 0.97 and near-zero mass below 0. This reflects a recording in which delta band power is consistently above the training mean — a hallmark of recordings with a high proportion of slow-wave sleep. The synthetic distribution, by contrast, is broadly spread from approximately −4.5 to +1.5 with a gradual ramp-up shape, peaking around 0 to +0.5 at a density of approximately 0.55 to 0.60. Crucially, the mismatch here is not confined to the distribution tails: the shapes of the two distributions are fundamentally different — a narrow high-density spike versus a broad, low-density ramp. The generator has learned the population-level delta distribution, which spans subjects with a wide range of sleep depths and slow-wave sleep proportions, but this particular test subject's distribution is concentrated at the upper end of that population range in a manner the generator does not reproduce.

The relative gamma power distribution (Figure 4.15) represents the most extreme case of marginal divergence across all six features. The real distribution is effectively confined to a single narrow bin centred at approximately −0.75 to −0.5 (standardised), with a

deviations around this mode. The synthetic distribution shares the same approximate modal location (peak density ~1.8) but spreads mass across a range from −1 to approximately +9 on the standardised axis, spanning roughly ten standard deviations. Gamma band power in frontal sleep EEG is typically very low and highly uniform within any

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

single recording session; inter-session variability, however, is driven by electrode impedance, amplifier gain, and subject-specific scalp resistance — factors that differ between training and test subjects in ways the generator cannot adjust for at inference time. The extreme concentration of the real gamma distribution and the very broad synthetic counterpart produce the largest marginal KS statistic of any feature in this analysis, and are the most probable primary driver of the KS max value reported for this recording.

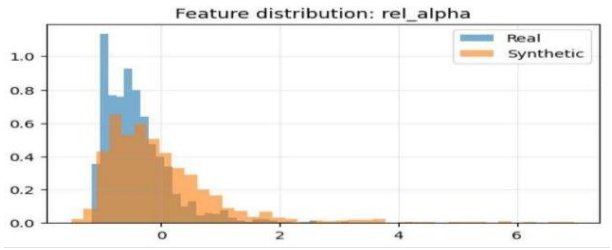

Figure 4.13. Marginal distribution histograms for rel_alpha (v1, best test recording). The real distribution is a narrow spike at approximately −0.5 standardised (alpha suppression during sleep), while the synthetic distribution generates a heavy right tail extending to approximately 7 standard deviations, corresponding to wakefulness-associated alpha activity absent in this test subject's recording.

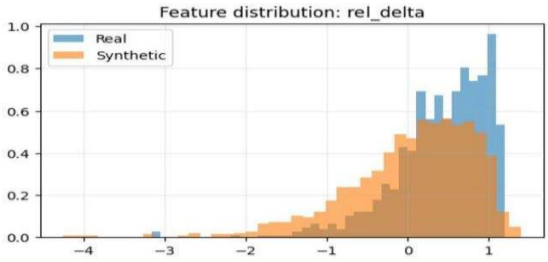

Figure 4.14. Marginal distribution histograms for rel_delta (v1, best test recording). The real distribution is a narrow spike in the range 0 to +1.2 standardised, reflecting consistently elevated slow-wave power, while the synthetic distribution is a broad ramp spanning −4.5 to +1.5. The mismatch is not confined to the tails but extends to the overall distributional shape.

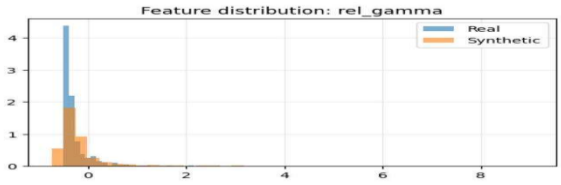

Figure 4.15. Marginal distribution histograms for rel_gamma (v1, best test recording). The most extreme marginal mismatch of any feature: the real distribution is an exceptionally narrow spike (peak density ~4.4) while the synthetic distribution is broadly spread across approximately ten standard deviations. This feature is the most probable primary driver of the observed KS max statistic for this recording.

Table 4.1. Summary of marginal distributional agreement for v1 on the best-performing test recording (ST7242J0-PSG.edf). Features are listed in order of presentation (Figures 4.10–4.15). "Over-dispersion" refers to whether the synthetic distribution is visibly wider than the real distribution in the histogram; this was observed for all six features.

Figure	Feature	Agreement	Primary Failure Mode	Over-Dispersion
4.10	rel_theta	Good	Slight tail over-dispersion in positive tail	Mild
4.11	entropy	Good	Slight rightward shift of ~0.3 SD; modest broadening	Mild
4.12	rel_beta	Moderate	Spurious right tail to ~8 SD (wakefulness episodes)	Moderate
4.13	rel_alpha	Poor	Heavy right tail to ~7 SD (wakefulness alpha)	Severe
4.14	rel_delta	Poor	Shape mismatch — narrow spike vs broad ramp across 6 SD	Severe
4.15	rel_gamma	Very Poor	Near-constant real vs ~10 SD synthetic spread	Extreme

V. DISCUSSION AND CONCLUSION

The training loss curves for both v1 and v2 (Figures 4.1 and 4.2) are consistent with the theoretical expectations for a well-behaved WGAN-GP optimisation. In the original Wasserstein GAN formulation, Arjovsky et al. [1] demonstrated that the critic loss provides a meaningful and continuous surrogate for the Wasserstein-1 distance between the real and generated distributions, in contrast to the saturating or oscillating behaviour frequently observed in the original GAN objective. The gradient penalty introduced by Gulrajani et al. [2] further stabilises training by enforcing the Lipschitz constraint through soft penalisation of the gradient norm at interpolated points, rather than the hard weight clipping used in the original WGAN. The convergence of the critic loss to near-zero in both models — reached by approximately epoch five in each case — is consistent with the critic having learned a well-calibrated linear approximation of the Wasserstein distance in the relevant region of feature space, after which further training primarily refines the generator.

The plateau reached by v1 at approximately 8.0–8.6 and by v2 at approximately 11.5–12.0 represents an approximate equilibrium at which the critic's discriminative capacity and the generator's ability to fool it are balanced. The higher equilibrium value for v2 is consistent with a harder optimisation landscape: the ten-dimensional feature space containing morphological features presents a more complex target distribution, and the Wasserstein distance at equilibrium reflects the residual gap between the generator

and real distributions that the fixed-capacity architecture was unable to fully close within 30 epochs.

One aspect of the v1 training curve that merits attention is the modest downward drift in the generator loss over the final five epochs, from approximately 8.6 at epoch 25 to approximately 7.5 at epoch 29. This drift does not coincide with any deterioration in the held-out evaluation metrics — the MMD² and KS values reported in Chapter 4 reflect performance at the end of training, not at the peak of the generator loss curve — which indicates that the late-stage loss movement is a training-time fluctuation rather than an onset of instability. The evaluation results themselves provide the most reliable evidence that 30 epochs was sufficient for the v1 model to reach a stable, well-generalising solution

This project successfully developed an unsupervised generative framework to model sleep EEG feature distributions using a Wasserstein Generative Adversarial Network (WGAN-GP). Unlike traditional sleep analysis approaches that rely on labeled sleep stages, the proposed system learns population-level EEG patterns without supervision, enabling a deeper understanding of intrinsic sleep dynamics.The system demonstrated the ability to generate realistic synthetic EEG feature samples and generalize to previously unseen subjects, as validated using distribution-based evaluation metrics such as Maximum Mean Discrepancy (MMD), Kolmogorov–Smirnov (KS) tests, and UMAP visualization. Experimental results further revealed that spectral EEG features generalize more reliably across subjects than morphology-sensitive features, highlighting an important trade-off between physiological detail and cross-subject robustness.Overall, the project confirms the feasibility and effectiveness of generative artificial intelligence for sleep EEG modeling, providing a foundation for future research in unsupervised sleep analysis, synthetic EEG generation, and population-level brain signal modeling.

REFERENCES

[1] Rechtschaffen, A., & Kales, A. (Eds.). (1968). *A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects.* US Department of Health, Education, and Welfare.

[2] Berry, R. B., Brooks, R., Gamaldo, C. E., Harding, S. M., Lloyd, R. M., Marcus, C. L., & Vaughn, B. V. (2012). *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications, Version 2.0.* American Academy of Sleep Medicine.

[3] Morin, C. M., & Benca, R. (2012). Chronic insomnia. *The Lancet*, *379*(9821), 1129–1141. https://doi.org/10.1016/S0140-6736(11)60750-2

[4] Supratak, A., Dong, H., Wu, C., & Guo, Y. (2017). DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *25*(11), 1998–2008. https://doi.org/10.1109/TNSRE.2017.2721116

[5] Phan, H., Chén, O. Y., Tran, M. C., Koch, P., Mertins, A., & De Vos, M. (2022). XSleepNet: Multi-view sequential model for automatic sleep staging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(9), 5903–5915. https://doi.org/10.1109/TPAMI.2021.3070057

[6] Perslev, M., Darkner, S., Kempfner, L., Nikolic, M., Jennum, P. J., & Igel, C. (2021). U-Sleep: Resilient high-frequency sleep staging. *npj Digital Medicine*, *4*, 72. https://doi.org/10.1038/s41746-021-00440-5

[7] Mousavi, S., Afghah, F., & Acharya, U. R. (2019). SleepEEGNet: Automated sleep stage scoring with sequence to sequence deep learning approach. *PLOS ONE*, *14*(5), e0216456. https://doi.org/10.1371/journal.pone.0216456

[8] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems, 27.* https://proceedings.neurips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html

[9] Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein GAN. *Proceedings of the 34th International Conference on Machine Learning (ICML), PMLR 70*, 214–223. https://arxiv.org/abs/1701.07875

[10] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. C. (2017). Improved training of Wasserstein GANs. *Advances in Neural Information Processing Systems, 30.* https://proceedings.neurips.cc/paper/2017/hash/892c3b1c6dccd52936e27cbd0ff683d6-Abstract.html

[11] Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv:1511.06434.* https://arxiv.org/abs/1511.06434

[12] Hartmann, K. G., Schirrmeister, R. T., & Ball, T. (2018). EEG-GAN: Generative adversarial networks for electroencephalographic (EEG) brain signals. *arXiv:1806.01875.* https://arxiv.org/abs/1806.01875

[13] Zhang, J., Yao, R., & Ge, W. (2021). SleepEGAN: A generative adversarial network for data augmentation of minority classes in sleep EEG class imbalance problems. *Biomedical Signal Processing and Control*, *67*, 102516. https://doi.org/10.1016/j.bspc.2021.102516

[14] Lashgari, E., Liang, D., & Maoz, U. (2020). Data augmentation for deep-learning-based electroencephalography. *Journal of Neuroscience Methods*, *346*, 108885. https://doi.org/10.1016/j.jneumeth.2020.108885

[15] Panwar, S., Rad, P., Nam, C. S., & Bhatt, P. (2020). Modeling EEG data distribution with a Wasserstein generative adversarial network to predict RSVP events. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *28*(2), 474–483. https://doi.org/10.1109/TNSRE.2019.2958787

[16] Pascual, S., Bonafonte, A., & Serrà, J. (2017). SEGAN: Speech enhancement generative adversarial network. *Proceedings of Interspeech 2017*, 3642–3646. https://doi.org/10.21437/Interspeech.2017-1428

[17] Kingma, D. P., & Welling, M. (2013). Auto-encoding variational Bayes. *arXiv:1312.6114.* https://arxiv.org/abs/1312.6114

[18] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, *33*, 6840–6851. https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html

[19] Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, *13*, 723–773. http://www.jmlr.org/papers/v13/gretton12a.html

[20] McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv:1802.03426.* https://arxiv.org/abs/1802.03426

[21] Kemp, B., Värri, A., Rosa, A. C., Nielsen, K. D., & Claasen, J. (1992). A simple format for exchange of digitized polygraphic recordings. *Electroencephalography and Clinical Neurophysiology*, *82*(5), 391–393. https://doi.org/10.1016/0013-4694(92)90009-7

[22] Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C. K., & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, *101*(23), e215–e220. https://doi.org/10.1161/01.CIR.101.23.e215

[23] Hjorth, B. (1970). EEG analysis based on time domain properties. *Electroencephalography and Clinical Neurophysiology*, *29*(3), 306–310. https://doi.org/10.1016/0013-4694(70)90143-4

[24] Welch, P. D. (1967). The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics*, *15*(2), 70–73. https://doi.org/10.1109/TAU.1967.1161901

[25] https://doi.org/10.1016/0013-4694(91)90138-T

[26] Shorten, C., & Khoshgoftaar, T. M. (2019). A surveon image data augmentation for deep learning. *Journal of Big Data*, *6*, 60. https://doi.org/10.1186/s40537-019-0197-0