Brain Informatics

https://doi.org/10.1186/s40708-026-00327-9

Article in Press

Integrating convolutional variational autoencoders and the Gaussian mixture model for efficient manifold learning and clustering of spatially preserved EEG topographic maps

Received: 13 December 2025

Accepted: 3 August 2026

Published online: 24 August 2026

Cite this article as: Faremi S. & Longo L. Integrating convolutional variational autoencoders and the Gaussian mixture model for efficient manifold learning and clustering of spatially preserved EEG topographic maps. *Brain Inf.* (2026). https://doi.org/10.1186/s40708-026-00327-9

Saheed Faremi & Luca Longo

Integrating Convolutional Variational Autoencoders and the Gaussian Mixture Model for efficient manifold learning and clustering of spatially preserved EEG topographic maps

Saheed Faremi[1,2] and Luca Longo[1,2*]

[1]School of Computer Science and Information Technology, University College Cork, Western Road, Cork, T12 XF62, Munster, Ireland.
[2*]Artificial Intelligence and Cognitive Load Research Lab, School of Computer Science and Information Technology, University College Cork, Western Road, Cork, T12 XF62, Munster, Ireland.

*Corresponding author(s). E-mail(s): luca.longo@ucc.ie;
Contributing authors: 126100912@umail.ucc.ie;

Abstract

EEG microstate analysis segments continuous brain activity into brief, quasi-stable topographic configurations whose quantity, duration, and transition dynamics index cognitive state and neuropsychiatric conditions. The standard algorithm, modified $\boldsymbol{k}$-means (ModKMeans), operates in electrode space. It performs hard partitioning, assigning each data point to exactly one cluster, and provides no learned representation, no generative decoder, and no probabilistic soft assignment. In this research, a Convolutional Variational Deep Embedding model (Conv-VaDE) is designed, a topographic-map adaptation of Variational Deep Embedding (VaDE) that places a Gaussian Mixture Model (GMM) prior on the latent space of a convolutional variational autoencoder and jointly trains the encoder, decoder, and GMM. Microstate polarity invariance, the property that a topography $\mathbf{x}$ and its sign-inverted counterpart $-\mathbf{x}$ index the same underlying configuration, is preserved through three coordinated mechanisms: sign-flip augmentation of the training set, a polarity-invariant reconstruction loss, and an encoder regulariser on the latent embedding. The model is evaluated on a sample of healthy participants ($\boldsymbol{N = 203}$ healthy adults, eyes-closed) from the LEMON resting-state EEG dataset using a multi-quadrant evaluation framework, with each method assessed in its native frame and cross-method comparisons performed via rank-based Wilcoxon signed-rank tests. Conv-VaDE

instead offers a focused set of capabilities the standard ModKMeans pipeline does not provide: a learned generative latent manifold (silhouette 0.232 at $\boldsymbol{K = 3}$, with very narrow cross-subject interquartile-range width $\boldsymbol{\approx 0.02}$ across the 203 participants, indicating highly reproducible latent geometry), decoded GMM cluster centres inspectable as scalp topographies, probabilistic soft assignment for transitional states, and a generative framework supporting downstream representation analysis. A rank-based meta-criterion over silhouette, Calinski–Harabasz, and Davies–Bouldin yields modal $\boldsymbol{K = 4}$ for Conv-VaDE (39.4%) and $\boldsymbol{K = 3}$ for ModKMeans (29.6%); $\boldsymbol{K = 4}$ is adopted to align with the canonical Lehmann A/B/C/D literature. Within the classical cluster-validity register, ModKMeans attains a higher silhouette and 4–5 percentage points higher Global Explained Variance (GEV) at every $\boldsymbol{K}$ (Wilcoxon, BH-corrected, $\boldsymbol{p < 10^{-30}}$). The two methods occupy distinct positions in the methodological design space, with the cross-method comparison characterising the trade-off between direct electrode-space optimisation and learned generative representation rather than adjudicating method superiority.

Keywords: Deep Clustering, CNN-VAE, Gaussian Mixture Models, Topographic Maps, Unsupervised Learning, Neuroimaging, EEG Signal Processing, Microstate theory;

1 Introduction

Electroencephalography (EEG) records scalp-level electrical activity at millisecond resolution, making it the dominant non-invasive modality for studying large-scale neural dynamics in humans [1]. Despite advances in neural signal processing, supervised methods that require manual labelling remain prevalent in EEG analysis, limiting the discovery of naturally occurring unlabelled activity patterns [2, 3]. A central challenge in EEG analysis is the automatic identification of recurrent neural patterns within and across subjects. Microstate theory [4] addresses this challenge by decomposing multi-channel EEG into sequences of transient, quasi-stable scalp topographies, microstates, hypothesised to reflect activity in large-scale neuronal networks. Microstate metrics have been associated with cognitive, affective, and clinical states across psychiatric and neurological conditions, supporting their use as candidate biomarkers [5]. Computationally, microstates correspond to clusters of EEG topographies derived from multichannel recordings, but identifying a stable, reproducible set of clusters across subjects remains methodologically unresolved. K-means and its microstate-specific variants, polarity-invariant (ModKMeans), spherical, and weighted formulations, operate on raw topographies or flattened electrode vectors, relying on manually defined similarity measures that assume fixed geometric relationships rather than learned representations. EEG topographies are believed to lie on nonlinear, high-dimensional manifolds [4], where Euclidean distance between flattened electrode vectors poorly reflects topographic similarity. Furthermore, principled statistical validation of cluster solutions in EEG microstate analysis, particularly cross-method, cross-representation comparisons, remains underdeveloped [6–8]. In contrast, deep

clustering methods, including autoencoder-based and contrastive embeddings, jointly learn nonlinear latent representations and cluster assignments, capturing topographic similarity structure that flat Euclidean distance cannot resolve. Autoencoder-based deep clustering also denoises implicitly through reconstruction, and regularisers such as dropout and adversarial training further constrain the latent geometry.

This research study evaluates a Convolutional Variational Deep Embedding model (Conv-VaDE) for unsupervised clustering of EEG topographic maps. Conv-VaDE is a topographic-map adaptation of Variational Deep Embedding (VaDE; [9]), which places a Gaussian Mixture Model (GMM) prior over the latent space of a convolutional variational autoencoder and jointly trains the encoder, decoder, and GMM via a mixture-augmented evidence lower bound. A microstate-specific feature of this adaptation is the explicit treatment of polarity invariance, the canonical microstate property that a topography $\mathbf{x}$ and its sign-inverted counterpart $-\mathbf{x}$ index the same underlying configuration of cortical generators [4, 10], which the standard ModKMeans pipeline absorbs through an absolute-value spatial-correlation similarity [11]. Conv-VaDE preserves this invariance through sign-flip augmentation, a polarity-invariant reconstruction loss, and a latent regulariser. The architectural configuration (latent dimension 32, network depth 4, channel width 32) was identified as cross-K-robust in earlier pilot work [12]. The study acknowledges that pilot-derived hyperparameters carry transfer assumptions when scaled to larger datasets. The present study scales this configuration from a 10-participant pilot to the full LEMON resting-state subjects ($N = 203$ healthy adults, eyes-closed) and undertakes a methodologically matched comparison against the field-standard ModKMeans baseline, implemented via pycrostates [11].

Two methodological contributions are introduced. First, a Multi-quadrant evaluation framework reports each cluster validity index across four representation $\times$ distance conditions. Each method is evaluated in its native frame, Conv-VaDE in latent space with sklearn Euclidean distance, ModKMeans in its native pycrostates pipeline, and cross-method comparisons use rank-based Wilcoxon signed-rank tests. A decoded-topomap evaluation, in which Conv-VaDE is assessed under the ModKMeans/pycrostates pipeline, is reported separately as a methodological robustness check. Second, a rank-based meta-criterion is applied across the canonical microstate cluster-validity indices, silhouette, Calinski-Harabasz, and Davies-Bouldin, within each method's native frame to identify the participant-level optimal cluster count. Bootstrap confidence intervals on the rank-stability statistic and pairwise Wilcoxon signed-rank tests with Benjamini-Hochberg correction quantify uncertainty across K. The central research question is to characterise the design trade-off between Conv-VaDE, evaluated in its native frame at participant scale, and the field-standard ModKMeans pipeline: where the two methods sit on classical cluster-validity, and what complementary capabilities Conv-VaDE provides that the standard pipeline does not.

The remainder of this manuscript is structured as follows. Section 2 reviews the existing literature on EEG microstate clustering and deep generative methods. Section 3 details the LEMON dataset, the Conv-VaDE architecture, the Multi-Quadrant Evaluation framework, and the statistical pipeline. Section 4 presents the matched-conditions results, the meta-criterion for K selection, and the per-subject distributions. Section 5 discusses the findings and presents limitations and technical extensions. Section 6 concludes the study.

2 Related Work

Unsupervised representation learning for EEG is often based on: deep generative modelling, microstate clustering, and joint embedding-clustering frameworks. Deep learning has reduced reliance on hand-crafted EEG features, with convolutional networks (CNNs) dominating supervised analysis, appearing in roughly 67 to 75% of EEG deep-learning studies surveyed up to 2019 [2, 3, 13, 14] and reporting mean accuracy gains of approximately 5.4 percentage points over hand-crafted baselines. Rakhmatulin et al. [15] grouped CNN architectures for EEG into standard, recurrent-convolutional hybrid, decoder-based, and combined categories, with effectiveness across both frequency-domain and spatial-domain analyses. However, supervised approaches require labelled datasets at scale and do not directly support unsupervised pattern discovery. Zhang et al. [16] report that approximately 70% of EEG deep-learning studies addressed classification tasks, with comparatively limited attention to unsupervised methods.

Variational autoencoders (VAEs) have demonstrated the capacity to learn compressed EEG representations while preserving reconstruction quality. Bethge et al. [7] introduced EEG2Vec for affective representations. Zhao [17] proposed VAEEG for cross-subject emotion recognition. Cisotto et al. [18] extended this line with hvEEG-Net, a hierarchical VAE achieving high-fidelity multi-channel reconstruction via a Dynamic Time Warping (DTW) loss that outperforms standard mean-squared-error objectives. Despite these advances, existing VAE applications to EEG predominantly serve as feature extractors for downstream supervised tasks, rather than as the basis for unsupervised cluster discovery with statistical validation. An initial step toward unsupervised EEG representation learning appears in [19, 20], where person-specific VAEs minimised latent dimensionality while preserving reconstruction quality.

Recent deep generative models have begun to address recurrent pattern discovery in EEG. Chen et al. [21] proposed a Double Discrete Variational Autoencoder (D2-VAE) for 'deep discretisation', compressing epileptic signals into low-dimensional discrete codes that capture recurring underlying patterns. In resting-state analysis, [22] applied an explainable convolutional autoencoder to specific pathologies, recovering eight distinct EEG states differentiated by varying activation levels, validating the autoencoder's capacity for natural cluster discovery. Yue et al. [23] pursued related unsupervised feature extraction. Bollens et al. [24] highlighted the importance of subject-invariant representations for ensuring discovered patterns generalise

rather than reflecting idiosyncratic artefacts. This line was extended in [25] with clustering over emergent latent dimensions to support interpretability, and Criscuolo et al. [26] trained person-specific VAEs whose latent dimensions were interpreted for semi-automated eye-blink artefact detection. However, in these works clustering was applied *post hoc*, typically using K-means or similar non-gradient-based methods, rather than jointly during manifold learning, leaving the embedding-clustering objective decoupled.

Joint optimisation of representation learning and clustering addresses this decoupling. Ren et al. [27] surveyed deep clustering across single-view, semi-supervised, multi-view, and transfer-learning categories, reporting that joint learning of representations and clusters consistently outperforms sequential approaches. Guo et al. [28] proposed VAE-based deep clustering with gamma-mixture latent embeddings, effective across several domains but not yet applied to neurophysiological signals, closely related to the Gaussian-mixture latent prior used in the present study. Lin et al. [29] applied Gaussian Mixture Models (GMMs) to EEG artefact detection via canonical correlation analysis, demonstrating GMM viability on EEG features though in a non-clustering context. However, the application of GMM priors over learned EEG latent spaces, combined with participant-level statistical validation of the optimal cluster count, remains largely unexplored.

The microstate literature establishes that four canonical topographic classes (A to D) account for roughly 80% of resting-state variance [4, 5, 30–32], with up to 15 microstates required for comparable variance explanation in active-state recordings [33]. These findings derive from K-means clustering on flattened topographies, leaving open whether modern deep generative frameworks, capable of learning nonlinear topographic manifolds, yield complementary or improved cluster solutions. Most directly relevant to the present work, Dinov and Leech [34] performed probabilistic microstate clustering using softmax multi-layer perceptron (MLP) classifiers trained on K-means (KM) and Fuzzy C-Means (FCM) cluster assignments as target labels, identifying $K = 5$ as optimal via within-cluster distances and silhouette analysis. Eskandari Nasab et al. [35] combined Gated Recurrent Unit and CNN (GRU-CNN) architectures with microstate and recurrence-quantification features for auditory-attention detection, identifying $K = 10$ in their setting and diverging from earlier K-means-based studies [36, 37] that relied on cross-validation metrics. The K-selection variability across this literature (4, 5, 8, 10, 15) underscores the absence of a unified statistical framework for participant-level cluster-count selection.

The work most methodologically adjacent to the present study includes Fan et al. [38], who demonstrated joint VAE and GMM optimisation for video anomaly detection. Csore et al. [39] applied a convolutional VAE with GMM latent prior to medical-image segmentation, establishing the architectural template that Conv-VaDE adapts to EEG topographies. Tabejaamaat et al. [8] integrated latent-space clustering as a regulariser for EEG classification, achieving 97.9% accuracy on EEG-ImageNet

under limited-data conditions. Han et al. [40] introduced noise-factorised representations for motor-imagery EEG, separating subject-specific, task-specific, and noise components. Across this literature, VAE-based EEG representation learning has prioritised reconstruction and supervised downstream tasks over unsupervised cluster discovery with statistical validation. Microstate clustering relies predominantly on K-means, which assumes linear cluster boundaries and struggles with the nonlinear geometry of topographic data. K-selection also varies widely across studies, reflecting the absence of a unified validation framework. Joint Conv-VAE/GMM architectures have shown promise in adjacent domains but have not yet been applied to multichannel EEG topographic maps. Such adaptation would need to account for the geometry, dimensionality, and polarity invariance specific to scalp topographies. The present study addresses this convergence by adapting the VaDE framework to EEG topographic maps, integrating a participant-level rank-based meta-criterion across canonical microstate cluster-validity indices for cluster-count selection, and benchmarking against the field-standard modified k-means pipeline in a methodologically matched evaluation.

3 Methods

This study evaluates the Convolutional Variational Deep Embedding (Conv-VaDE) framework on the LEMON resting-state EEG sample and compares it against the field-standard modified k-means baseline (ModKMeans, implemented via the pycrostates pipeline [11]), with each method evaluated in its native framework. The analysis is structured around three objectives: (i) scaling the architectural configuration validated in earlier pilot work [12] from a 10-participant pilot to the full LEMON sample (N = 203); (ii) the Multi-Quadrant Evaluation framework, which reports cluster-validity indices across four (representation × distance) conditions and explicitly separates Conv-VaDE's native evaluation frame (latent / sklearn-Euclidean) from supplementary robustness frames; and (iii) the participant-scale rank-based meta-criterion across the canonical microstate cluster-validity indices, applied within each method's native framework to identify the optimum number of clusters $K^\star$.

The hypothesis evaluated is whether, in its native evaluation framework, Conv-VaDE achieves cluster-validity scores comparable to those of ModKMeans. The null hypothesis (H_0) is that Conv-VaDE and ModKMeans yield equivalent cluster-validity scores at the optimum $K^\star$ in their respective native frames. Conv-VaDE and ModKMeans are positioned as two points in a methodological design space rather than ranked on a single cluster-validity score. ModKMeans optimises directly in the 61-dimensional electrode geometry against which the classical indices are computed, whereas Conv-VaDE incurs the structural cost of a learned latent compression and decoder reconstruction; the classical cluster-validity indices are accordingly expected to favour ModKMeans, and the comparison is framed as a design trade-off rather than a contest of scores. Conv-VaDE is assessed additionally at the level of capability: the reproducibility of its learned latent geometry across participants, decoded GMM centroids inspectable as scalp topographies, and soft probabilistic assignments

for transitional states, none of which the hard-assignment ModKMeans pipeline provides. Because the two methods are scored in different native spaces, their absolute values are not directly comparable, and all cross-method comparison uses rank-based tests, which are invariant to the differing scales.

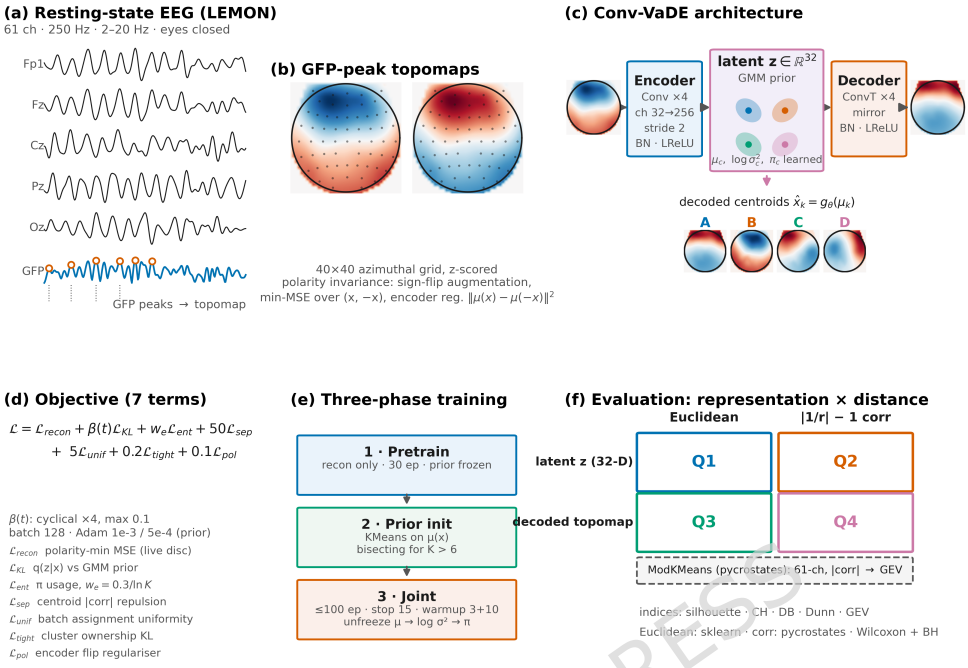

Fig. 1 Conv-VaDE pipeline overview, six panels: (a) LEMON dataset and preprocessing (eyes-closed resting state, Global Field Power (GFP) peak topographic-map extraction, z-score normalisation); (b) three-level polarity invariance scheme (sign-flip augmentation, polarity-invariant reconstruction loss, encoder regularisation); (c) Conv-VaDE encoder and decoder architecture with the Gaussian Mixture Model prior on the latent space; (d) seven-component training objective (reconstruction, KL with cyclical β-annealing, entropy, separation, batch-uniformity, cluster-tightness, polarity-invariance); (e) training pipeline (pretraining, GMM initialisation via K-means ($n_{\text{init}} = 20$), end-to-end joint training); (f) Multi-Quadrant Evaluation framework reporting cluster-validity indices in four (representation × distance) conditions.

3.1 Dataset and Data Preparation

The Leipzig Study for Mind-Body-Emotion Interactions (LEMON) dataset [41] provides a publicly available resting-state EEG sample widely used in EEG benchmarking. The original dataset comprises 228 healthy adults; the 203-participant subset analysed here corresponds to the participants released in the publicly distributed

`EEG_Preprocessed_BIDS_ID/EEG_Preprocessed/` archive[1] for whom complete eyes-closed recordings with consistent electrode coverage are available. The full list of participant identifiers is provided in Appendix B. Recordings use 61 scalp electrodes plus one vertical electro-oculographic (VEOG) channel arranged in the 10–10 montage, sampled at 250 Hz, in the eyes-closed resting condition. A common-average reference and a zero-phase finite impulse response (FIR) bandpass filter (2–20 Hz, consistent with standard microstate preprocessing) are applied to preserve temporal alignment and prevent phase-induced microstate-boundary blurring.

Topographic maps are extracted using the pycrostates library [11] at local maxima of the Global Field Power (GFP) signal, with a minimum inter-peak distance of three samples to prevent oversampling of temporally adjacent peaks. The peak detection identifies time points $t^\star$ for which $\mathrm{GFP}(t^\star) > \mathrm{GFP}(t^\star \pm 1)$, subject to the inter-peak constraint $|t_i^\star - t_{i-1}^\star| \geq 3$. Each extracted GFP peak yields a 61-dimensional voltage vector across the scalp; electrode positions are converted from three-dimensional Cartesian to two-dimensional coordinates via the azimuthal-equidistant projection, with cubic-spline interpolation producing a 40×40 topographic image (resolution chosen to balance interpolation fidelity against model capacity, consistent with [12]). Grid points outside the convex hull of electrode positions are assigned zero values.

Within each participant, the extracted GFP-peak maps are partitioned 90/10 into a training set and an evaluation set by a fixed random permutation (seed 42), drawn independently for each participant, and the evaluation set is left unaugmented. The split is within-participant; cross-participant generalisation lies beyond the scope of this study. Topographic maps are normalised by per-participant global z-score standardisation with $\pm 5\sigma$ clipping. A single scalar mean $\bar{x}_{\mathrm{train}}$ and standard deviation σ_{train} are computed by pooling every pixel of every training-split map for that participant, that is, global statistics over the training set rather than per-pixel or per-map statistics, and are applied uniformly to every pixel of every map:

$$x_{\mathrm{norm}} = \mathrm{clip}\left(\frac{x - \bar{x}_{\mathrm{train}}}{\sigma_{\mathrm{train}}}, -5, +5\right).$$

The statistics are estimated on the training split alone so that no evaluation-set information enters the normalisation constants.

3.2 Conv-VaDE Architecture

The Conv-VaDE architecture (Figure 1, panels c–e) consists of a convolutional encoder, a probabilistic latent space organised by a Gaussian Mixture Model prior, and a transposed-convolutional decoder. The configuration evaluated here (latent dimension $d_z = 32$, network depth $L = 4$, channel width $n_f = 32$) is selected as the cross-K robust architecture in the architecture-search pilot of [12], where 486 architectural configurations were swept across $K \in [3, 20]$ on a 10-participant pilot subset of the same

[1] https://ftp.gwdg.de/pub/misc/MPI-Leipzig_Mind-Brain-Body-LEMON/EEG_MPILMBB_LEMON/EEG_Preprocessed_BIDS_ID/EEG_Preprocessed/, accessed during data preparation. The directory contains paired eyes-closed (`_EC`) and eyes-open (`_EO`) EEGLAB `.set`/`.fdt` files for 203 subjects.

LEMON dataset. The transfer of these hyperparameters to the full participants carries an assumption of architectural stability across participants scale. Training the full set of person-specific models is computationally substantial: the 18 values of $K \in [3, 20]$ among 203 participants amount to $18 \times 203 = 3654$ separately-trained Conv-VaDE models, each optimised for up to 100 epochs with early stopping (a median of 19 epochs before stopping). On multiple NVIDIA GPUs this required approximately 30 days of total compute.

The encoder progressively downsamples the 40×40 topographic maps through convolutional blocks of increasing channel width, with Leaky ReLU activations (slope $\alpha = 0.2$), batch normalisation, and dropout ($p = 0.2$), culminating in adaptive average pooling and a fully connected layer that produces the latent distribution parameters $(\boldsymbol{\mu}_\phi, \log \boldsymbol{\sigma}^2_\phi) \in \mathbb{R}^{d_z}$. In plain terms, the encoder maps each topographic map not to a single latent point but to a d_z-dimensional Gaussian: $\boldsymbol{\mu}_\phi$ is its mean (the central latent code for that map) and $\log \boldsymbol{\sigma}^2_\phi$ is its log-variance (the encoder's uncertainty along each latent axis); the log-variance, rather than the variance, is predicted so that the value is unconstrained in sign and numerically stable to optimise. The decoder mirrors the encoder via transposed convolutions with a linear final activation.

Microstate topographies are polarity-invariant: a map $\mathbf{x}$ and its sign-inverted counterpart $-\mathbf{x}$ index the same microstate, because the sign of an EEG topography depends on the arbitrary recording reference rather than on the underlying cortical generators. Conv-VaDE enforces this property at three complementary points in the pipeline. First, *sign-flip augmentation* presents every training map both as $\mathbf{x}$ and as $-\mathbf{x}$, exposing the encoder to both polarities during training. Augmentation supplies the data, while the two loss terms below do the actual enforcing. Second, a *polarity-invariant reconstruction loss*, $\mathcal{L}_{\text{recon}} = \min(\text{MSE}(\hat{\mathbf{x}}, \mathbf{x}), \text{MSE}(\hat{\mathbf{x}}, -\mathbf{x}))$, scores the decoder output $\hat{\mathbf{x}}$ against whichever of $\mathbf{x}$ or $-\mathbf{x}$ it matches more closely. Taking the smaller of the two errors judges the network on reproducing the *shape* of the topography and does not penalise it for returning the map at the opposite polarity. Third, an *encoder regularisation term*, $\mathcal{L}_{\text{pol}} = \text{MSE}(\boldsymbol{\mu}_\phi(\mathbf{x}), \boldsymbol{\mu}_\phi(-\mathbf{x}))$, penalises any difference between the latent means assigned to a map and to its sign-flip, pulling the two polarities to almost the same location in latent space so that they are not split across different clusters. The three mechanisms act on distinct levels, the input data, the decoder output, and the latent representation, so polarity invariance is imposed across the whole model rather than relying on any single loss to do all the work.

The Gaussian Mixture Model (GMM) integration replaces the standard normal VAE prior with learnable Gaussian mixture components. For each tested cluster count $k \in [3, 20]$, the model learns mixture weights $\boldsymbol{\pi}$ (softmax-constrained), component means $\boldsymbol{\mu}_c$ (a $k \times 32$ matrix), and component variances $\boldsymbol{\sigma}^2_c$ (log-parameterised, clamped to $[-4.0, 2.0]$). The range $k \in [3, 20]$ is selected to cover the canonical four-microstate solution [4] alongside the wider range reported in the microstate literature: reported optima range from 3 to 7 clusters [10] and up to 15 clusters for comparable variance explanation in active states [33]. The reparameterisation trick enables differentiable sampling: $\boldsymbol{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$. Detailed layer specifications for the complete encoder-decoder pipeline are provided in Table E3 in Appendix E.

Conv-VaDE inherits the VaDE family's joint encoder–decoder–GMM training scheme and its mixture-augmented evidence lower bound [9]. The evidence lower bound is the objective a variational autoencoder maximises in place of the intractable data log-likelihood; *mixture-augmented* means the usual single-Gaussian prior is replaced by the Gaussian mixture, so the bound also accounts for the soft assignment of each map to the mixture components, and maximising it jointly fits the encoder, decoder and mixture so that the latent codes both reconstruct the input and organise into clusters. Conv-VaDE introduces six microstate-specific extensions that distinguish it from the original VaDE formulation:

1. a convolutional encoder and decoder operating on 40×40 azimuthal-projected topographic images of multi-channel scalp potentials, in place of the fully-connected layers of the original VaDE;
2. the three-level polarity-invariance scheme described above and formalised in Section 3.3;
3. the seven-component composite training objective specified in Section 3.3, comprising a polarity-invariant reconstruction loss, a mixture-weighted KL divergence under cyclical β-annealing, and four auxiliary clustering regularisers;
4. per-participant training to respect subject-level heterogeneity, consistent with prior person-specific VAE work in EEG [19, 20, 26];
5. the Multi-Quadrant Evaluation framework for cross-paradigm cluster-validity reporting (Section 3.5);
6. the participant-level rank-based meta-criterion for K-selection (Section 3.6).

Flattening the 40×40 map to a 1600-dimensional vector would let vanilla VaDE ingest it, but this discards the spatial structure the convolutional encoder is built to exploit; a comparison would then conflate the topographic adaptation with the change of input format and yield an uninformative baseline. ModKMeans, by contrast, is the field-standard method for clustering scalp topographies and is applied to the same GFP-peak maps in their native 61-channel representation; it is therefore the meaningful baseline and is reported throughout Section 4.

The simplified two-term loss balances reconstruction fidelity and posterior regularisation:

$$\mathcal{L} = \mathcal{L}_{\text{MSE}} + \beta(t) \cdot \mathcal{L}_{\text{KLD}},$$

where the mean squared error (MSE) measures pixel-wise reconstruction fidelity, and the Kullback-Leibler divergence (KLD) regularises encoder distributions against the GMM prior through a mixture-weighted divergence:

$$\mathcal{L}_{\text{KLD}} = \sum_{i=1}^{N} \sum_{c=1}^{k} \gamma_{ic}\, \text{KL}\big[q(\boldsymbol{z}_i \mid \boldsymbol{x}_i) \,\|\, \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c^2)\big]\,, \tag{1}$$

with responsibilities γ_{ic} computed from mixture components, N the batch size, and k the number of Gaussian mixture components. The responsibility γ_{ic} is the posterior

probability that sample i belongs to component c, given by:

$$\gamma_{ic} = \frac{\pi_c \mathcal{N}(\boldsymbol{z}_i \mid \boldsymbol{\mu}_c, \boldsymbol{\sigma}_c^2)}{\sum_{j=1}^{k} \pi_j \mathcal{N}(\boldsymbol{z}_i \mid \boldsymbol{\mu}_j, \boldsymbol{\sigma}_j^2)}, \tag{2}$$

where π_c are the mixture weights, softmax-constrained to sum to one. The encoder's approximate posterior for latent variable $\boldsymbol{z}_i$ given input map $\boldsymbol{x}_i$ is denoted $q(\boldsymbol{z}_i \mid \boldsymbol{x}_i)$, while $\mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c^2)$ is the c-th Gaussian component of the GMM prior with learnable mean $\boldsymbol{\mu}_c$ and variance $\boldsymbol{\sigma}_c^2$. The full seven-component training objective used in joint training is specified in Section 3.3.

The KL weighting coefficient $\beta(t)$ employs batch-level cyclical annealing [42] to prevent posterior collapse and maintain balance between reconstruction and regularisation. The base value of β is $\beta_{\min} = 0.01$, and the schedule cycles β across the full training run (cycle count and $\beta_{\max}$ specified in Section 3.3 and Table E2).

Person-specific models are trained for each participant across all k values through a three-phase protocol; participant-level reporting in Sections 3.5 and 3.6 aggregates per-subject cluster-validity statistics rather than pooling latent embeddings across subjects:

- Phase 1 (Pretraining): pretraining with fixed low β ($\beta_{\min} = 0.01$) emphasising reconstruction, with auxiliary losses zeroed.
- Phase 2 (GMM Initialisation): K-means clustering ($n_{\text{init}} = 20$) on encoded latents followed by scikit-learn Gaussian Mixture fitting (diagonal covariance, 1000 iterations, 10^{-4} regularisation), with parameter transfer to the model.
- Phase 3 (Joint Training): end-to-end optimisation using Adam (encoder and decoder learning rate 10^{-3}, weight decay 10^{-5}; GMM learning rate 5×10^{-4}) with ReduceLROnPlateau scheduling and gradient clipping (max_norm = 5.0). Early stopping monitors the validation loss to prevent overfitting.

The complete hyperparameter specifications, including epoch counts, scheduler patience values, and annealing parameters, are provided in Table E2.

3.3 Polarity Invariance and the Composite Training Objective

EEG microstates are conventionally treated as polarity-invariant: a topographic configuration $\mathbf{x}$ and its sign-inverted counterpart $-\mathbf{x}$ correspond to the same underlying configuration of cortical generators because the recorded scalp polarity reflects the instantaneous direction of dipolar sources rather than a separate brain state [4, 10]. The standard ModKMeans pipeline absorbs this invariance through an absolute-value cosine (correlation) similarity in 61-dimensional electrode space [11]. Conv-VaDE preserves the same invariance in its learned representation through three coordinated mechanisms (panel b of Figure 1): (i) *sign-flip augmentation* of the training set, doubling each participant's GFP-peak set with $-\mathbf{x}$ copies so that both polarities are observed during training; (ii) a *polarity-invariant reconstruction loss*, which compares the decoded map against whichever of $\mathbf{x}$ or $-\mathbf{x}$ it is closer to, removing the incentive for the encoder to preserve sign information; and (iii) an *encoder regulariser* penalising disagreement between $\boldsymbol{\mu}_\phi(\mathbf{x})$ and $\boldsymbol{\mu}_\phi(-\mathbf{x})$, anchoring the latent embedding to be

sign-symmetric. Combined with the mixture-weighted KL divergence and three auxiliary clustering regularisers that prevent posterior collapse, mixture degeneracy, and centroid drift, the joint-training objective is the seven-component composite:

$$\mathcal{L} \;=\; \mathcal{L}_{\text{recon}} + \beta(t)\,\mathcal{L}_{\text{KLD}} + \lambda_{\text{ent}}\,\mathcal{L}_{\text{ent}} + \lambda_{\text{sep}}\,\mathcal{L}_{\text{sep}} + \lambda_{\text{batch}}\,\mathcal{L}_{\text{batch}} + \lambda_{\text{tight}}\,\mathcal{L}_{\text{tight}} + \lambda_{\text{pol}}\,\mathcal{L}_{\text{pol}}. \tag{3}$$

Polarity-invariant reconstruction $\mathcal{L}_{\text{recon}}$. Per-sample minimum of the mean-squared reconstruction error against the original and sign-flipped target:

$$\mathcal{L}_{\text{recon}} \;=\; \frac{1}{N}\sum_{i=1}^{N} \min\Big(\text{MSE}(\hat{\mathbf{x}}_i, \mathbf{x}_i),\; \text{MSE}(\hat{\mathbf{x}}_i, -\mathbf{x}_i)\Big), \tag{4}$$

with $\hat{\mathbf{x}}_i = g_\theta(\boldsymbol{z}_i)$ the decoded map for sample i.

Mixture-weighted KL divergence $\mathcal{L}_{\text{KLD}}$. Defined as in equations (10)–(11); algebraically, it decomposes into a cluster-specific divergence, a mixture-assignment regularisation, and a posterior-entropy term [9] (Appendix A). Its weight $\beta(t)$ follows the cyclical-annealing schedule of [42] with $\beta_{\min} = 0.01$, $\beta_{\max} = 0.1$, and four cycles over the full training run. The low-β phases favour reconstruction whilst the high-β phases enforce prior alignment, preventing posterior collapse.

Prior-entropy regulariser $\mathcal{L}_{\text{ent}}$. Discourages dead clusters in the global mixture by penalising deviation of the softmax-normalised mixture weights $\tilde{\boldsymbol{\pi}} = \text{softmax}(\boldsymbol{\pi})$ from the uniform distribution:

$$\mathcal{L}_{\text{ent}} \;=\; \log K \;+\; \sum_{c=1}^{K} \tilde{\pi}_c \log \tilde{\pi}_c, \tag{5}$$

which is zero when $\tilde{\boldsymbol{\pi}}$ is uniform and strictly positive otherwise.

Centroid-separation regulariser $\mathcal{L}_{\text{sep}}$. Continuous repulsion between decoded cluster centroids in topographic space, computed as the mean absolute Pearson correlation between centred decoded centroids over all distinct cluster pairs:

$$\mathcal{L}_{\text{sep}} \;=\; \frac{2}{K(K-1)} \sum_{1 \le c < c' \le K} \big|\, \rho(g_\theta(\boldsymbol{\mu}_c),\, g_\theta(\boldsymbol{\mu}_{c'}))\big|, \tag{6}$$

where $\rho(\cdot,\cdot)$ is the spatial correlation between two flattened 40×40 topographic maps and g_θ is the decoder. The absolute value preserves polarity invariance. The formulation provides a non-vanishing gradient that maintains gentle pressure even after centroids have spread, in contrast to log-barrier alternatives that go to zero past a separation threshold.

Batch-uniformity regulariser $\mathcal{L}_{\text{batch}}$. Encourages each mini-batch to use all clusters approximately equally, complementing the prior-entropy regulariser by acting

on the data-driven posterior rather than the prior weights. With per-sample responsibilities $\boldsymbol{\gamma}_i \in \Delta^{K-1}$ from equation (11) and the batch-mean assignment $\bar{\boldsymbol{\gamma}} = \frac{1}{N}\sum_i \boldsymbol{\gamma}_i$:

$$\mathcal{L}_{\text{batch}} = \sum_{c=1}^{K} \bar{\gamma}_c \log \frac{\bar{\gamma}_c}{1/K}. \tag{7}$$

Cluster-tightness term $\mathcal{L}_{\text{tight}}$. Augmented-ELBO term tightening each sample's posterior to its hard-assigned mixture component $c_i^\star = \arg\max_c \gamma_{ic}$:

$$\mathcal{L}_{\text{tight}} = \frac{1}{N}\sum_{i=1}^{N} \text{KL}\Big[\mathcal{N}(\boldsymbol{\mu}_{\phi,i}, \boldsymbol{\sigma}^2_{\phi,i}) \,\|\, \mathcal{N}(\boldsymbol{\mu}_{c_i^\star}, \boldsymbol{\sigma}^2_{c_i^\star})\Big]. \tag{8}$$

This per-sample component-specific KL is the standard tightening step shared across the VaDE, MFCVAE, and DCGMM families and improves cluster compactness.

Encoder polarity regulariser $\mathcal{L}_{\text{pol}}$. Penalises latent disagreement between a sample and its sign-inverted counterpart, forming the third level of the polarity-invariance scheme alongside the augmentation and the reconstruction loss:

$$\mathcal{L}_{\text{pol}} = \frac{1}{N}\sum_{i=1}^{N} \big\| \boldsymbol{\mu}_\phi(\mathbf{x}_i) - \boldsymbol{\mu}_\phi(-\mathbf{x}_i)\big\|_2^2. \tag{9}$$

Whereas $\mathcal{L}_{\text{recon}}$ removes the *incentive* for the encoder to preserve sign, $\mathcal{L}_{\text{pol}}$ explicitly *penalises* any residual sign sensitivity that survives the reconstruction objective.

Loss-component weights and warm-up. The auxiliary-term weights $\{\lambda_{\text{ent}}, \lambda_{\text{sep}}, \lambda_{\text{batch}}, \lambda_{\text{tight}}\}$ are zeroed during pretraining (the encoder must learn topographic features before clustering pressure is applied) and ramped linearly over the first ten epochs of joint training via a warmup$(t) \in [0, 1]$ scale factor. The polarity weight λ_{pol} and the cyclical $\beta(t)$ are active throughout. Numerical values are listed in Table E2. An ablation of these weights and of the β schedule is identified as a limitation in Section 5.1.

3.4 Gaussian Mixture Model Inference

For inference and cluster assignment, the GMM prior is parameterised as:

$$p(\mathbf{z}) = \sum_{c=1}^{K} \pi_c \mathcal{N}\big(\mathbf{z} \mid \boldsymbol{\mu}_c, \text{diag}(\boldsymbol{\sigma}_c^2)\big), \tag{10}$$

where K is the number of mixture components (clusters), and π_c, $\boldsymbol{\mu}_c$, and $\boldsymbol{\sigma}_c^2$ are the mixture weight, mean vector, and diagonal covariance for component c, respectively [9]. The KL term in the training loss decomposes into three interpretable

contributions: a cluster-specific divergence, a mixture-assignment regularisation, and a posterior-entropy term. Detailed derivations of $\gamma_c(\mathbf{z})$, numerical-stability constraints, and parameter-initialisation strategies are provided in Appendix A. During inference, the deterministic mean of the approximate posterior, $\boldsymbol{\mu}_{\mathbf{x}}$, is used to obtain stable cluster assignments. The predicted cluster $\hat{c}$ is selected by maximising the log-posterior probability:

$$\hat{c} = \arg\max_{c} \left[\log \pi_c + \log p(\boldsymbol{\mu}_{\mathbf{x}} \mid \text{component } c)\right]. \tag{11}$$

3.5 Multi-Quadrant Evaluation Framework

A central methodological contribution of this work is the explicit separation of *within-method* cluster characterisation from *cross-method* statistical comparison. Each cluster-validity index is computed in two representations (the 32-dimensional VAE latent embedding, or the 1600-dimensional decoded topomap, with $1600 = 40 \times 40$) and under two distance metrics (sklearn Euclidean distance, or pycrostates absolute spatial correlation), producing four representation $\times$ distance evaluation conditions, or 'quadrants':

- *Latent / sklearn-Euclidean*: latent embedding evaluated with sklearn Euclidean distance.
- *Latent / pycrostates*: latent embedding evaluated with pycrostates correlation distance.
- *Topomap / Euclidean*: decoded topomap evaluated with sklearn Euclidean distance.
- *Topomap / pycrostates*: decoded topomap evaluated with pycrostates correlation distance.

The ModKMeans baseline operates in the 61-dimensional electrode space using PyCrostates absolute correlation distance. In this research, *each method is evaluated in its own native framework*: Conv-VaDE in the latent embedding (sklearn-Euclidean distance, the geometry the model is effectively trained against, with the multivariate Gaussian GMM prior), and ModKMeans in its native pycrostates pipeline. This choice is methodologically consistent with the architecture-search pilot [12], which uses latent-Euclidean as the headline cluster-quality measure, and with the meta-criterion that identifies the per-method optimal K (Section 3.6). The remaining quadrants are reported as supplementary robustness checks. In particular, the decoded-topomap / pycrostates condition (Conv-VaDE assessed under the ModKMeans evaluation pipeline) is reported in Section 4.2 as a methodological robustness disclosure.

Cross-method statistical comparisons are conducted with paired Wilcoxon signed-rank tests, which are rank-based and therefore valid across the absolute-scale differences between the two methods' native cluster-validity computations. The test does not claim that one method's absolute silhouette (or Davies-Bouldin, or Dunn) value is higher than the other's. The two computations live in different geometries (32-dimensional latent versus 61-dimensional electrode) under different distance metrics (sklearn Euclidean versus pycrostates correlation), and their absolute magnitudes are not directly commensurable. The test claims, instead, that one method's per-subject ranking is consistently distinguishable from the other's at FDR-corrected significance,

which is the appropriate non-parametric inference for cross-paradigm comparison when the paradigms operate in different spaces. Global Explained Variance (GEV) is the one cluster-validity measure that supports a direct, scale-aligned numerical comparison between methods, because GEV is computed by backfitting cluster centres onto the original 61-channel electrode space in both methods and is therefore matched by construction. The 4–5 percentage-point GEV gap reported in Section 4.1 is interpretable in absolute terms; the silhouette, Davies-Bouldin, and Dunn comparisons are interpretable only as rank-based signals.

3.6 Optimal Cluster Determination

The optimum number of clusters $K^\star$ is determined through a rank-based meta-criterion across the canonical microstate cluster-validity indices (silhouette, Calinski-Harabasz, Davies-Bouldin) computed in each method's native evaluation framework. For each subject, the eighteen K values are ranked separately for each metric (silhouette and Calinski-Harabasz ranked descending so that better = rank 1; Davies-Bouldin ranked ascending). The metric ranks are summed per $(K, \text{subject})$ cell, and the per-subject optimum is the K that minimises this sum. The sample-level optimum is the mode of the per-subject optima across the 203 LEMON participants. Sum-of-ranks is selected for its simplicity and invariance to monotone metric transformations; alternatives such as Borda count or the geometric mean of ranks yielded qualitatively identical optima in pilot work [12]. For full transparency, four variants of the meta-criterion are computed: the three-cluster-validity-index (three-CVI) and four-CVI sets, each in the native-frame and matched-frame settings. The three-CVI native-frame variant is reported as primary because it (i) matches the canonical microstate validity-index set, (ii) evaluates each method in its own native frame consistent with the rest of the analysis, and (iii) follows the reporting convention of the architecture-search pilot [12]. The rank-based meta-criterion is complemented by the GEV elbow, defined as the first K at which the median per-K GEV gain falls below one percentage point (a parsimony heuristic also used in pilot work [12]). The elbow check is independent of the meta-criterion and provides a parsimony anchor against which the meta-criterion outcome can be interpreted.

Statistical inference is conducted with non-parametric tests appropriate to the heavy-tailed and scale-discrepant distributions of the cluster-validity indices in pycrostates correlation space (distributional diagnostics in Appendix C). Friedman tests assess the omnibus K-effect per (metric, method) family, with Kendall's W as the effect size and a non-parametric subject-bootstrap ($B = 2000$) confidence interval on W. Pairwise Wilcoxon signed-rank tests compare K values across the eighteen-value range, with Benjamini-Hochberg false-discovery-rate correction applied per (metric, method) family of 153 pairs ($153 = \binom{18}{2}$). Bias-corrected and accelerated (BCa) bootstrap confidence intervals on the sample mean of every metric at every K are computed with $B = 5000$ subject-resamples per cell, falling back to percentile-bootstrap on cells where BCa estimation fails (failure rate reported in Appendix D).

3.7 Implementation and training strategy

Conv-VaDE is trained per-participant with the configuration $(d_z, L, n_f) = (32, 4, 32)$ adopted from [12] and a fixed random seed of 20 for reproducibility. For each participant, eighteen models are trained at $K \in [3, 20]$, yielding $18 \times 203 = 3654$ person-specific models across the sample. Pretraining runs for one epoch (approximately 200 batch steps depending on participant peak count) with $\beta_{\text{min}} = 0.01$ and auxiliary losses zeroed. The GMM prior is initialised via K-means ($n_{\text{init}} = 20$) on unaugmented latent means. Main training proceeds for up to 100 epochs with early stopping (patience = 15 epochs on validation loss), using dual Adam optimisers (encoder and decoder learning rate 10^{-3}, weight decay 10^{-5}; GMM learning rate $5{\times}10^{-4}$), gradient clipping at 5.0, and ReduceLROnPlateau scheduling (patience = 10, factor = 0.5). Joint training optimises the seven-component composite of equation (3); per-term definitions, weights, and the warm-up schedule are given in Section 3.3 and Table E2. The implementation uses PyTorch 2.5, MNE-Python 1.8, scikit-learn 1.5, and pycrostates [11]. All architectural specifications and training protocols are detailed in Tables E3 and E5 in Appendix E.

4 Results

The Conv-VaDE architecture was trained on the full LEMON sample (N = 203 healthy adults, eyes-closed condition) at every $K \in [3, 20]$, yielding 3654 person-specific models in total. The modified k-means baseline was computed in parallel using the pycrostates pipeline at the same K values for the same 203 participants. Throughout this section, each method is evaluated in its native framework: Conv-VaDE in the latent space (sklearn-Euclidean distance) and ModKMeans in its native pycrostates pipeline (per Section 3.5). The two methods sit at distinct points of the methodological design space. ModKMeans optimises classical cluster validity directly in 61-dimensional electrode space, with no representation-learning overhead. Conv-VaDE trades part of that direct optimisation for a learned generative latent manifold and the downstream capabilities it enables. The cross-method comparisons reported below characterise this design trade-off rather than adjudicating method superiority. Cross-method statistical comparisons use paired Wilcoxon signed-rank tests, which are rank-based and valid across the absolute-scale differences between the two methods' native computations (Section 3.5).

4.1 Cross-Method Comparison

In its native pycrostates pipeline, ModKMeans achieves higher classical cluster validity than Conv-VaDE at every $K \in [3, 20]$ across every classical metric tested. This pattern is expected from the design trade-off introduced in the previous paragraph: ModKMeans optimises directly in the electrode-space geometry against which the classical cluster-validity indices are computed, whereas Conv-VaDE traverses an encoder–decoder pathway that necessarily imposes a representational cost on classical metrics in exchange for the generative latent manifold characterised in Section 4.3. Table 1

reports the per-subject median values at three representative K. Figure 2 shows the full K trajectory with quartile bands and per-K FDR-corrected significance markers.

Table 1 Cross-method comparison at representative K values. Each method is evaluated in its native framework: Conv-VaDE in the latent space (sklearn Euclidean distance) and ModKMeans in its native pycrostates pipeline (electrode/correlation distance). Cells report per-subject median of each metric (N = 203 LEMON, eyes-closed). Δ is the per-subject median of paired (Conv-VaDE minus ModKMeans) differences (negative values indicate ModKMeans is higher; lower is better for Davies-Bouldin). All cells are significant at FDR-adjusted $p < 10^{-8}$ under paired Wilcoxon signed-rank with Benjamini-Hochberg correction within each metric family.

K	Metric	Conv-VaDE	ModKMeans	Δ	p_{FDR}
3	Silhouette	0.232	0.588	-0.351	1.6×10^{-33}
3	Calinski–Harabasz	3475	4301	-779	9.2×10^{-9}
3	GEV	0.658	0.713	-0.037	2.4×10^{-34}
4	Silhouette	0.225	0.495	-0.268	1.8×10^{-33}
4	Calinski–Harabasz	3206	4038	-698	1.0×10^{-9}
4	GEV	0.691	0.747	-0.044	6.0×10^{-35}
5	Silhouette	0.189	0.415	-0.225	1.8×10^{-33}
5	Calinski–Harabasz	2879	3558	-632	1.0×10^{-9}
5	GEV	0.715	0.771	-0.052	6.0×10^{-35}

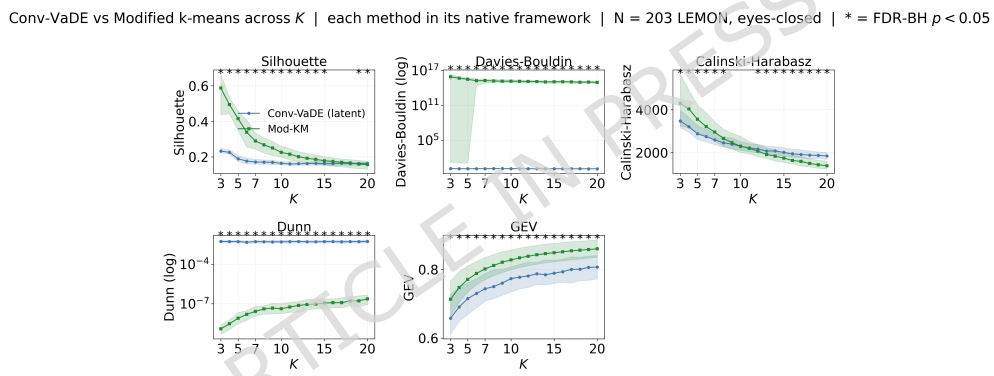

Fig. 2 Cross-method comparison of Conv-VaDE (latent / sklearn Euclidean, the model's native frame) and modified k-means (electrode / pycrostates correlation distance, its native frame) across $K \in [3, 20]$. Lines show per-subject medians; shaded bands show interquartile range across the 203-participant LEMON participants. Asterisks above each K indicate Benjamini-Hochberg FDR-corrected $p < 0.05$ under paired Wilcoxon signed-rank within each metric family. Davies-Bouldin and Dunn use a logarithmic vertical axis because their absolute scales differ between methods (32-D latent versus 61-D electrode evaluation space).

The Davies-Bouldin and Dunn comparisons in Table 1 and Figure 2 reach extreme significance and large absolute magnitudes. However, these two metrics live on incommensurable absolute scales between the two methods. ModKMeans (in 61-dimensional

electrode space with pycrostates correlation distance) yields Davies-Bouldin values on the order of 10^{15} and Dunn values on the order of 10^{-8}. These magnitudes reflect the very small denominators of the correlation-distance computation in this implementation rather than a clustering pathology. Conv-VaDE (in 32-dimensional latent space with sklearn Euclidean distance) yields Davies-Bouldin on the order of 10^{0} and Dunn on the order of 10^{-3}, in their conventional ranges. The Wilcoxon rank-based test remains valid regardless of scale, and the direction of effect is preserved within each method's native frame. We therefore retain Davies-Bouldin and Dunn as within-method ranking signals, but make no cross-method absolute-magnitude claim on these two metrics. Bias-corrected and accelerated bootstrap 95% confidence intervals on the sample mean of every metric are reported in the supplementary material.

Global Explained Variance (GEV), the one metric matched between methods by construction (both compute GEV by backfitting cluster centres onto the original 61-channel electrode space), shows ModKMeans approximately five percentage points higher than Conv-VaDE at every K tested. As discussed in the architecture-search pilot [12], ModKMeans templates are native electrode vectors optimised for spatial correlation, giving a structural advantage on GEV. Conv-VaDE decoded centroids must transit from the 40×40 pixel grid back to the 61-electrode space before GEV is computed, attenuating the spatial correlation that GEV weights quadratically.

4.2 Optimal Cluster Count under the Native-Frame Meta-Criterion

The rank-based meta-criterion across the canonical microstate cluster-validity indices (silhouette, Calinski-Harabasz, Davies-Bouldin), applied within each method's native evaluation framework per Section 3.6, identifies $K^{\star} = 4$ as the modal per-subject optimum for Conv-VaDE (39.4% subject support, in the latent / sklearn Euclidean frame). For ModKMeans, the modal optimum is $K = 3$ (29.6% subject support), with $K = 4$ a close second at 23.2% in a relatively flat distribution. The remaining per-subject optima distribute across $K \in \{5, 6, 7, \ldots\}$ in both methods. This per-participant heterogeneity is itself substantive: it indicates that microstate organisation is participant-specific and that any single global K is a reporting convention chosen for cross-study comparability rather than a biologically invariant quantity recoverable from the data alone. Table 2 reports the four-variant analysis; Figure 3 shows the per-subject distribution of optima for the primary variant.

The Conv-VaDE native-frame optimum $K = 4$ aligns with the canonical four-microstate literature (Lehmann A/B/C/D; [4]) and with the architecture-search pilot [12]. The ModKMeans native-frame optimum is $K = 3$, with $K = 4$ a close second (23.2% support); the per-subject distribution is flat across $K \in [3, 7]$, indicating that ModKMeans does not strongly prefer one canonical-range K over another in this sample. The GEV elbow, computed independently as the first K at which the median GEV gain falls below one percentage point, sits at $K = 8$ in both methods. The discrepancy between the meta-criterion optimum (low K) and the GEV elbow ($K = 8$) reflects the conventional tension between cluster-validity criteria, which favour parsimony, and variance-explanation criteria, which reward additional components. Both

Table 2 Optimal K under the rank-based meta-criterion across four (representation × distance) evaluation variants. Variant A (each method evaluated in its native frame, three cluster-validity indices) is the primary report; Variants B–D are reported for transparency. Conv-VaDE shifts from $K = 4$ in the native-frame variants to $K = 3$ in the matched-frame variants, reflecting the difference between the model's native latent geometry and its decoded-topomap evaluation under the field-standard distance metric. ModKMeans is in its native frame in all variants because the 61-dimensional electrode space is the only space ModKMeans operates in. Modal K is followed by the percentage of subjects whose per-subject optimum equals that K.

Variant	Configuration	Conv-VaDE $K^\star$	ModKMeans $K^\star$
A	**3 CVIs, native (PRIMARY)**	**4 (39.4%)**	**3 (29.6%)**
B	4 CVIs, native	4 (34.0%)	3 (18.7%)
C	3 CVIs, matched (decoded)	3 (62.1%)	3 (29.6%)
D	4 CVIs, matched (decoded)	3 (41.4%)	3 (18.7%)

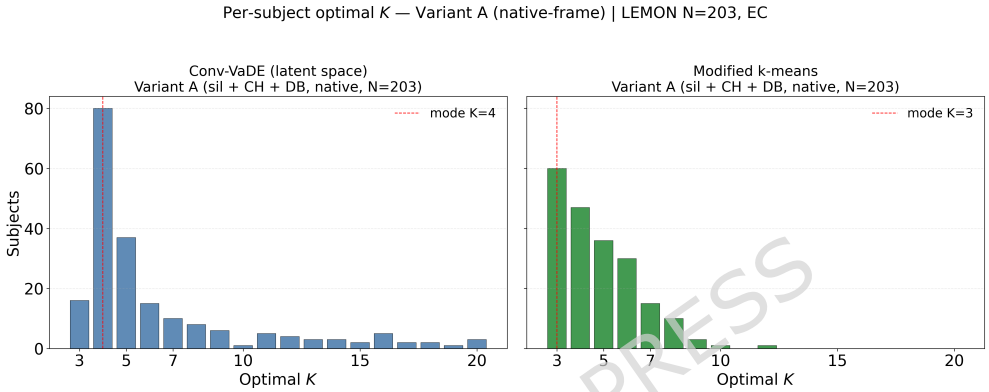

Fig. 3 Distribution of per-subject optimal K under the primary meta-criterion (Variant A: silhouette + Calinski-Harabasz + Davies-Bouldin computed in each method's native evaluation framework). Conv-VaDE (latent) (left) shows the modal optimum at $K = 4$ (80 of 203 subjects, 39.4%); ModKMeans (right) shows the modal optimum at $K = 3$ (60 of 203, 29.6%, with $K = 4$ a close second at 23.2% in a flat distribution). The dashed red line marks the modal optimum for each method.

methods exhibit the same pattern, supporting interpretive convergence.

LEMON was selected because its 203-participant resting-state design matches the canonical microstate analytical framework, which has long anchored at $K = 4$ [4]. Active-state and task-based EEG studies have reported larger optimal K values [33, 35], consistent with task-driven mobilisation of additional functional networks beyond the canonical resting-state set.

Under matched cross-method conditions (Variants C and D, where Conv-VaDE is evaluated on the decoded topomap with pycrostates correlation distance), the Conv-VaDE optimum shifts to $K = 3$ (62.1% in Variant C, 41.4% in Variant D). This shift is reported as a methodological robustness finding: it indicates that decoder-induced

attenuation of latent cluster separation in electrode-space correlation distance pulls the matched-conditions optimum to a smaller K, consistent with the GEV-attenuation discussion in the architecture-search pilot [12]. We acknowledge that the matched-frame plurality at $K = 3$ (62.1% in Variant C) is more concentrated than the native-frame plurality at $K = 4$ (39.4% in Variant A). However, the native frame is the geometry against which the Conv-VaDE objective is optimised, and is therefore the appropriate basis for declaring the model's intrinsic cluster structure. The operational $K = 4$ choice is anchored to the model's native-frame optimum, the canonical microstate literature, and the architecture-search pilot.

4.3 Per-Subject Distributions

Per-subject value distributions are presented as raincloud plots (Figure 4), following the visualisation recommendation of [43]. Each panel overlays the per-subject value distribution for Conv-VaDE (latent, native frame) and ModKMeans (electrode, native frame) across all K values, providing simultaneous access to distributional shape, median and interquartile range, and the underlying $N = 203$ sample. Davies-Bouldin and Dunn use logarithmic vertical axes for the reasons given in Section 4.1. For completeness, Table 3 reports the Conv-VaDE native-latent silhouette (32-dimensional latent embedding, sklearn Euclidean distance) across $K \in [3, 7]$. These values are descriptive of the model's intrinsic latent geometry and are not statistically compared against ModKMeans, because the latent evaluation framework uses sklearn Euclidean distance on a 32-dimensional learned latent, a different evaluation framework from the 61-dimensional electrode-space pycrostates pipeline that defines the ModKMeans silhouette.

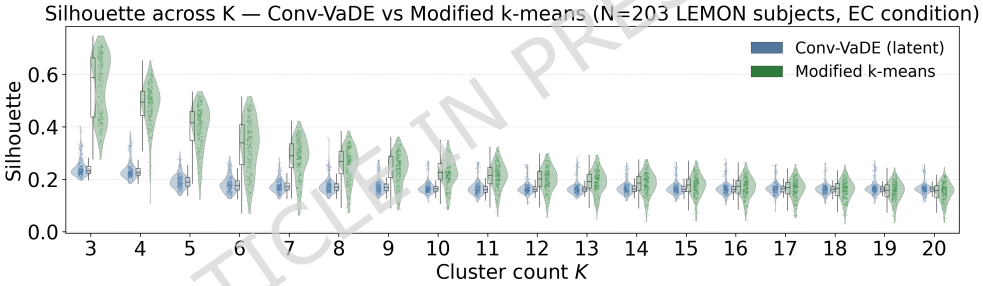

Fig. 4 Per-subject silhouette distributions across $K \in [3, 20]$ with each method evaluated in its native framework. The raincloud format combines a half-violin density estimate with an overlaid box-plot summary and jittered subject-level points (N = 203). At $K = 3$, ModKMeans achieves median silhouette 0.588 (interquartile range $[0.438, 0.663]$) versus Conv-VaDE (latent) median 0.232 (interquartile range $[0.223, 0.249]$). Conv-VaDE latent silhouette is positive across the full K range and shows very narrow cross-subject spread (interquartile-range width ≈ 0.02 to 0.04), indicating that the model's learned latent geometry is reproducible across the sample. Equivalent raincloud panels for Davies-Bouldin, Calinski-Harabasz, Dunn, and GEV are provided in the supplementary material.

Table 3 Conv-VaDE native-latent silhouette (32-dimensional latent embedding, sklearn Euclidean distance) across the canonical microstate range. Cells report per-subject median followed by interquartile range width. The narrow latent-Euclidean spreads (≈ 0.025 to 0.034) indicate that the Conv-VaDE latent geometry is reproducible across the 203 participants. These values are not directly comparable to ModKMeans silhouette because the two computations evaluate cluster quality in different spaces under different distance metrics.

K	**3**	**4**	**5**	**6**	**7**
Median	0.232	0.225	0.189	0.177	0.171
IQR width	0.026	0.025	0.034	0.034	0.026

4.4 Stability of the Modal Optimum

Pairwise paired Wilcoxon signed-rank tests across the eighteen K values are reported in Figure 5 as a Benjamini-Hochberg FDR-corrected p-value matrix per (metric, method) family of all $\binom{18}{2} = 153$ pairwise K comparisons. The operational reporting $K = 4$ (anchored to Conv-VaDE's native-frame optimum) is significantly distinct from each of $K \in \{3, 5, 6, 7, 10\}$ in Conv-VaDE (FDR-adjusted $p < 10^{-27}$ for $K = 4$ versus $K \geq 5$). For ModKMeans, $K = 4$ is statistically distinct from $K \in \{5, 6, 7, 10\}$ but lies within the flat $K \in [3, 4]$ region of the ModKMeans optimum distribution. These results show that $K = 4$ is a genuine peak for Conv-VaDE, scoring significantly better than its neighbouring K values rather than being picked out of a flat range where several K perform about equally well. For ModKMeans, the best K lies in the 3-to-4 range, where neighbouring values are statistically close.

Rank stability across the sample is characterised by Kendall's W with non-parametric subject-bootstrap 95% confidence intervals ($B = 2000$ resamples; Table 4). In other words, Kendall's W measures how consistently the 203 participants agree on the ordering of the K values for a given metric: $W = 1$ means every participant ranks the K values in the same order, and W near 0 means there is no agreement. For silhouette, ModKMeans achieves $W = 0.914$, 95% CI $[0.893, 0.932]$, near-perfect rank agreement across subjects. In contrast, Conv-VaDE in the native-latent quadrant achieves $W = 0.367$, 95% CI $[0.352, 0.390]$, thus a moderate agreement. The disparity is itself informative: Conv-VaDE's learned latent geometry is more participant-specific than ModKMeans's shared electrode-space evaluation, reflecting the per-participant training scheme. We adopt $K = 4$ as the operational $K^\star$ for reporting, consistent with the canonical literature. Per-participant $K^\star$ values are retained in the supplementary material for downstream applications. For Global Explained Variance, both methods show very high rank stability (Table 4: ModKMeans $W = 1.000$, consistent with the monotone reduction of within-cluster variance with increasing K in K-means; Conv-VaDE $W = 0.820$, 95% CI $[0.804, 0.836]$).

For completeness, Conv-VaDE latent silhouette shows no significant difference between $K = 10$ and $K = 12$ (Figure 6, Wilcoxon $p = 0.080$, $n = 203$, median $\Delta = -0.002$), while ModKMeans silhouette shows a small but significant difference

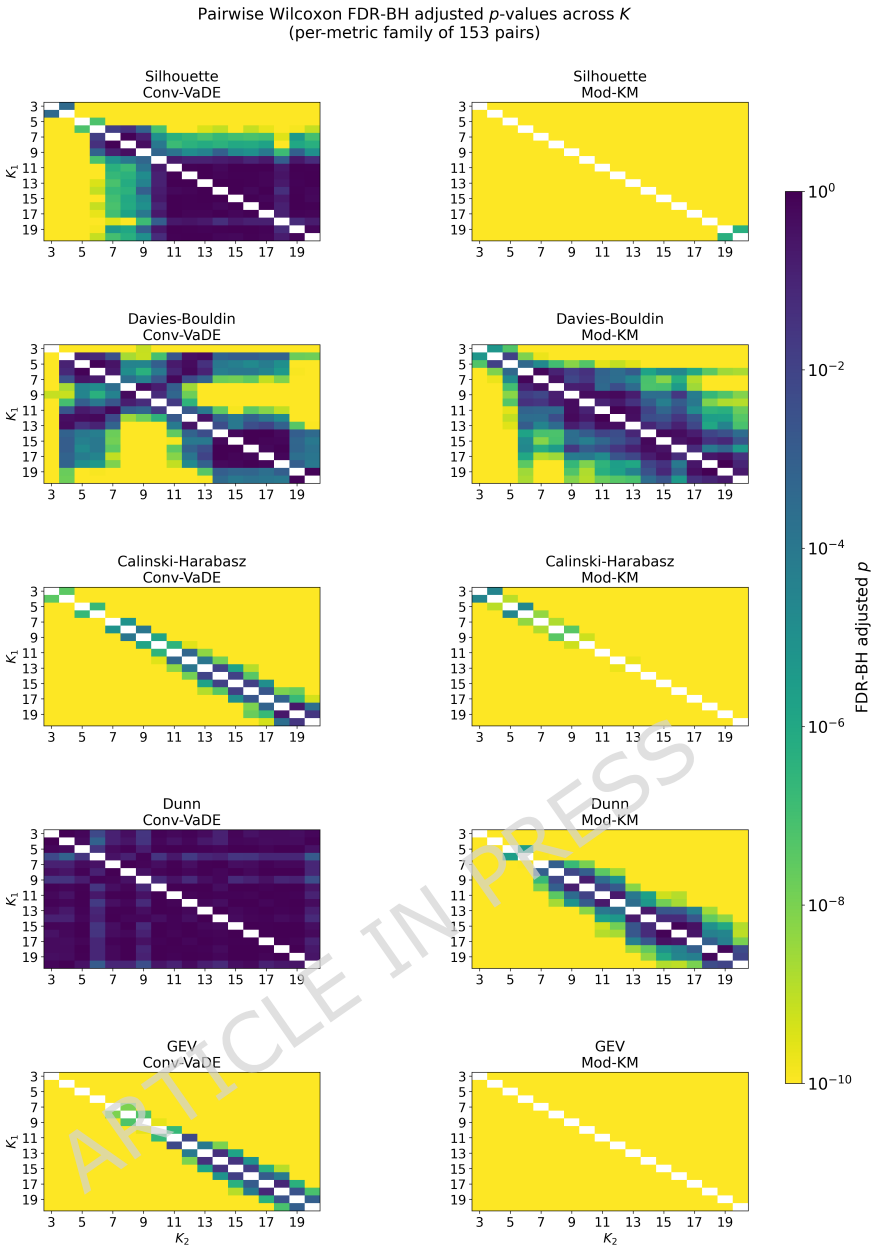

Fig. 5 Pairwise Wilcoxon signed-rank significance (p-value heatmap) across K values per (metric, method) family of all $\binom{18}{2} = 153$ pairwise K comparisons, with Benjamini-Hochberg false-discovery-rate correction. Darker cells indicate greater statistical separation between K_1 and K_2. The operational reporting $K = 4$ is statistically distinct from every adjacent K in Conv-VaDE; in Mod-KMeans, $K = 4$ lies within the flat $K \in [3, 4]$ region of the optimum distribution.

($p = 2.4 \times 10^{-29}$, median $\Delta = -0.026$). This means that neither K value is the operational optimum under the meta-criterion for the sample under evaluation.

Table 4 Rank stability of each cluster-validity index across the 203 participants, measured by Kendall's coefficient of concordance W with 95% subject-bootstrap confidence intervals ($B = 2000$). $W = 1$ denotes identical per-participant orderings of the eighteen K values; W near 0 denotes no agreement. Conv-VaDE is scored in its native latent / Euclidean frame, ModKMeans in its native electrode / correlation frame.

Metric	Method	Kendall's W	95% CI
Silhouette	ModKMeans	0.914	$[0.893, 0.932]$
	Conv-VaDE	0.367	$[0.352, 0.390]$
Calinski–Harabasz	ModKMeans	0.834	$[0.808, 0.859]$
	Conv-VaDE	0.703	$[0.664, 0.746]$
Davies–Bouldin	ModKMeans	0.067	$[0.053, 0.094]$
	Conv-VaDE	0.233	$[0.206, 0.265]$
Dunn	ModKMeans	0.517	$[0.484, 0.556]$
	Conv-VaDE	0.012	$[0.010, 0.025]$
GEV	ModKMeans	1.000	$[1.000, 1.000]$
	Conv-VaDE	0.820	$[0.804, 0.836]$

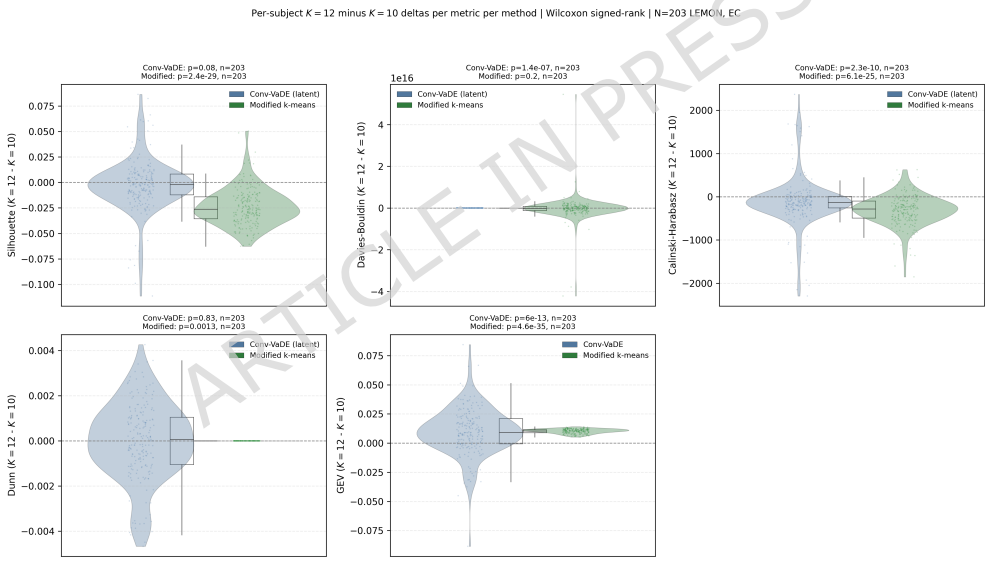

Fig. 6 Per-participant change in silhouette between $K = 10$ and $K = 12$ for each method (N = 203). Conv-VaDE shows no significant difference (Wilcoxon $p = 0.080$, median $\Delta = -0.002$). ModKMeans shows a small but significant difference ($p = 2.4 \times 10^{-29}$, median $\Delta = -0.026$).

4.5 Decoded Cluster Centroids at the Operational Optimum

Unlike the standard ModKMeans pipeline, Conv-VaDE can pass each cluster centre back through its trained decoder to generate the corresponding scalp topography. This opens up generative analyses, such as interpolating between clusters, sampling new topographies, and conditional generation, that are not available from cluster-mean prototypes alone. Figure 7 renders the four decoded GMM centroids at the operational reporting $K = 4$ for a representative LEMON subject. Participants' aggregate centroid analysis (cluster-template averaging across subjects after polarity-aligned matching) is reported in the supplementary material.

A spatial-correlation analysis between decoded Conv-VaDE centroids at $K = 4$ and the canonical Lehmann A/B/C/D templates was conducted across the full LEMON cohort ($N = 203$ subjects with available model weights). Per subject, the four decoded GMM centroids were mapped from the 40×40 pixel grid to the 61 electrode locations via bilinear interpolation, then matched one-to-one against the four canonical templates (generated as $x \pm y$ spatial-weighting combinations from the MNE standard montage electrode positions [4]) using the Hungarian algorithm with cost $1 - |\text{Pearson } r|$. The mean absolute spatial correlation across the four templates and all subjects was $|r| = 0.41$ (median 0.41, SD 0.11, 95% CI $[0.21, 0.71]$). Per-template means were $|r|_{\text{A}} = 0.36$, $|r|_{\text{B}} = 0.35$, $|r|_{\text{C}} = 0.48$, $|r|_{\text{D}} = 0.46$. The C/D lateral-asymmetry maps show systematically higher correspondence than the A/B anterior-posterior maps ($|r|_{\text{C/D best}} = 0.53$ median vs. $|r|_{\text{A/B best}} = 0.36$ median). The best single-template match per subject was $|r| = 0.57$ (median 0.58, SD 0.13). Twenty-four of 203 subjects (11.8%) showed strong recovery (mean $|r| \geq 0.50$); four subjects (2.0%) showed essentially no recovery (mean $|r| < 0.15$). Template A was the best-matching template for 61 subjects (30.0%), B for 16 (7.9%), C for 116 (57.1%), and D for 10 (4.9%). The full per-subject correlation table is provided in the supplementary material.

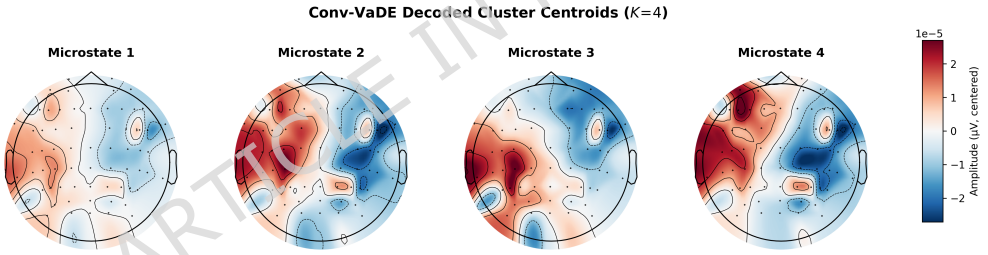

Fig. 7 Conv-VaDE decoded cluster centroids at the operational reporting $K = 4$ for a representative LEMON subject. Each panel shows the decoder output for one Gaussian Mixture Model component k, rendered as a circular scalp topography. Colour encodes amplitude in microvolts (RdBu_r colour map, zero-centred).

Taken together, the findings show that the established method, modified k-means, produces numerically tighter clusters on the standard validity scores, because it works directly in the sensor space where those scores are computed. Conv-VaDE does not

beat it on those numbers, and this study does not claim it should. Its value lies elsewhere: it learns a structured internal representation of brain-state topographies, its cluster centres can be turned back into readable scalp maps, it assigns each moment a soft (probabilistic) membership rather than a hard label, and it provides a generative model that supports further analysis. Across the 203 participants the model settles most often on four clusters, matching the four canonical microstates long established in the literature. The two approaches are therefore complementary rather than competing: one is stronger on classical cluster scores, the other adds interpretability and generative capability.

Main outcomes of this study. The principal findings can be summarised as follows.

- **Design trade-off, not superiority.** In its native pycrostates pipeline, ModKMeans attains higher classical cluster validity at every $K \in [3, 20]$, including a 4–5 percentage-point Global Explained Variance advantage (Wilcoxon, Benjamini-Hochberg corrected $p < 10^{-30}$). This reflects its direct optimisation in the electrode space against which the classical indices are computed.
- **Reproducible latent geometry.** Conv-VaDE achieves positive silhouette across the full K range with very narrow cross-subject spread (interquartile-range width ≈ 0.02 across $N = 203$ participants), indicating that the learned latent manifold is highly reproducible across the sample.
- **Participant-level K selection.** The rank-based meta-criterion identifies modal $K^\star = 4$ for Conv-VaDE (39.4% subject support) and $K = 3$ for ModKMeans (29.6%, with $K = 4$ a close second at 23.2%). The operational $K = 4$ aligns with the canonical four-microstate organisation (Lehmann A/B/C/D).
- **Statistical robustness of the optimum.** For Conv-VaDE, $K = 4$ is significantly distinct from every neighbouring K tested (FDR-adjusted $p < 10^{-27}$ versus $K \geq 5$). For ModKMeans, the optimum lies in a flat $K \in [3, 4]$ region. The GEV elbow sits at $K = 8$ in both methods, reflecting the shared tension between parsimony and variance explanation.
- **Generative interpretability.** Decoded GMM cluster centroids render as inspectable scalp topographies, with cohort-level correspondence to the canonical A/B/C/D templates of mean $|r| = 0.41$ (strongest for the C/D lateral-asymmetry maps).
- **Complementary capabilities.** Conv-VaDE provides a learned generative latent manifold, probabilistic soft assignment for transitional states, and a trained decoder supporting generative analyses, none of which the hard-assignment ModKMeans pipeline offers. The two methods occupy complementary niches rather than competing positions.

5 Discussion

The native-frame analysis places the two methods at distinct points of one design space rather than on a single quality ranking. ModKMeans achieves higher classical cluster validity at every K tested, including a Global Explained Variance gap of approximately four to five percentage points across the entire cluster-count range, reflecting its direct optimisation in the 61-dimensional electrode-space geometry against which

the classical metrics are computed. The study adopts $K = 4$ as the operational reporting K to align with Conv-VaDE's native-frame optimum, the canonical four-microstate literature, and the architecture-search pilot [12]. Under the matched-frame robustness variants (Variants C and D, where Conv-VaDE is evaluated on the decoded topomap with pycrostates correlation distance), the Conv-VaDE optimum shifts to $K = 3$, reported as a methodological disclosure reflecting decoder-induced attenuation of latent cluster separation in electrode-space correlation distance. The impact of Conv-VaDE, therefore, lies not in numerical superiority on classical metrics, but in capabilities the standard ModKMeans pipeline does not provide. In its native latent geometry (32-dimensional latent embedding, sklearn Euclidean distance), Conv-VaDE achieves positive silhouette throughout the cluster-count range (median 0.232 at $K = 3$, 0.225 at $K = 4$) with very narrow cross-subject spread (interquartile-range width ≈ 0.02), indicating that the model's learned manifold is highly reproducible across the sample. The Gaussian Mixture Model over this latent space supports probabilistic soft assignment for transitional states. The trained decoder can also be applied to GMM cluster centres to render them as generated scalp topographies, a generative interpretability route that the ModKMeans pipeline, which assigns each data point to exactly one cluster, does not provide. In practice, researchers seeking generative interpretability, soft probabilistic assignment, or representation-level downstream analyses can adopt Conv-VaDE as a framework. The two methods, modified k-means and Conv-VaDE, are not in direct competition. They occupy different niches in EEG analysis.

5.1 Limitations and Technical Extensions

Some limitations of the present study should be addressed in future work.

1. **Single architectural configuration.** The study evaluates a single architectural configuration ($d_z = 32$, $L = 4$, $n_f = 32$) selected for cross-K robustness in the architecture-search pilot of [12]. A systematic ablation of the cyclical β-annealing schedule and the seven-component loss weights, while informative, falls outside the validation scope of the present study and is left to future work.
2. **Modest canonical-template correspondence.** The spatial-correspondence analysis between decoded Conv-VaDE centroids and canonical templates (Section 4.5) yielded a cohort-level mean $|r| = 0.41$, which is modest. The decoded centroids carry individualised subject-level features that the same four A/B/C/D templates, constructed as generic spatial-weighting maps from a standard montage, do not capture. Per-subject template matching at this level of correspondence does not provide a strong quantitative grounding for interpreting Conv-VaDE cluster identities in terms of the canonical four. Alternative template-matching strategies, such as generating subject-specific templates from the ModKMeans centroids of the same subject, may yield a tighter correspondence and are left to future work.
3. **Eyes-closed condition only.** The eyes-closed condition was selected because resting-state microstate analysis is canonically performed in EC, where alpha-rhythm modulation provides the cleanest topographic-clustering substrate [4]. The eyes-open subset of LEMON is publicly available, and the preprocessing

pipeline released with this study supports direct EO replication. EO replication is identified as future work to test whether the design-trade-off characterisation generalises across resting-state subtypes, including replication under varied experimental conditions such as different recording paradigms and participant demographics.

4. **Transition dynamics not modelled.** Analysis of microstate state-transition dynamics, as commonly performed with Markov chain modelling in the classical microstate literature, has not been explored within the Conv-VaDE framework. Investigating transition probabilities between learned cluster assignments could reveal temporal dependencies not captured by the current per-timepoint classification approach.
5. **Distance metric.** The Euclidean distance metric used in the present clustering and evaluation may not optimally capture topographic similarity. Alternative distance measures, such as cosine distance or geodesic metrics on the scalp manifold, could yield improved cluster separation and are worth systematic comparison.
6. **Explainability and regularisation.** Integration of explainable artificial intelligence techniques (DeepSHAP, integrated gradients) and advanced VAE regularisation schemes (β-TCVAE, posterior-collapse mitigation) would extend the interpretability layer that the present architecture provides, supporting mechanistic interpretation of the learned latent space representations.
7. **Loss-term weighting.** The effect of modifying individual terms in the composite loss function on cluster separation quality remains unexplored. Targeted reweighting or reformulation of specific loss components could improve the balance between reconstruction fidelity and cluster compactness.
8. **Topographic interpolation attenuation.** The Conv-VaDE pipeline interpolates scalp potentials from 61 electrodes onto a 40×40 pixel grid via cubic spline, and decoded centroids are mapped back to electrode space for GEV computation and temporal backfitting. This grid-to-electrode round-trip introduces interpolation attenuation: the decoded centroid is sampled at 61 electrode locations, and bicubic interpolation between grid points smooths the spatial profile relative to a native electrode-space centroid. Spatial correlation, and therefore GEV (which weights correlation quadratically), is lower for Conv-VaDE than for ModKMeans by construction, even when the underlying topographic configuration is identical. A direct 61-channel decoder head that bypasses the grid for evaluation would isolate the clustering contribution from the interpolation cost. This is left to future work.
9. **Circular mask and corner-zero padding.** The 40×40 topographic grid inscribes a circular head-disc within a square frame. Roughly 30% of pixels fall outside the disc, are constant zero from the griddata interpolation (fill_value = 0), and carry no brain signal. A circular binary mask excludes these corner pixels from every loss, per-pixel metric, and pairwise topomap-space distance computation in the current pipeline. The mask radius factor (0.95 of the grid half-width, fixed) has not been calibrated against the electrode coverage of the LEMON 61-channel montage. A mask that is too conservative excludes live pixels near the scalp periphery, and biases cluster separation metrics downwards. A mask that is too permissive includes zero-valued corners that inflate the spatial correlation between centroids.

A sensitivity analysis over mask radii, compared against an unmasked baseline, would quantify this bias and is deferred to future work.

6 Conclusion

The central research problem addressed by this study is to characterise the design trade-off between a deep generative clustering framework (Conv-VaDE) and the field-standard ModKMeans pipeline on EEG microstate clustering, when each method is evaluated in its native frame. The study addressed this on the LEMON resting-state sample (N = 203, eyes-closed) using the pycrostates pipeline for ModKMeans. Two methodological contributions structured the comparison. First, the Multi-Quadrant Evaluation framework reports cluster-validity indices across four representation × distance conditions, separating Conv-VaDE's native evaluation frame from supplementary robustness frames. Second, a participant-scale rank-based meta-criterion across the canonical microstate cluster-validity indices is applied in each method's native frame to identify the optimal cluster count. The findings show that ModKMeans attains higher classical cluster validity at every K tested, whereas Conv-VaDE provides a reproducible learned latent manifold, a modal optimum of $K = 4$ aligned with the canonical four-microstate literature, probabilistic soft assignment for transitional states, and decodable cluster centroids inspectable as scalp topographies. The two methods are complementary rather than competing. ModKMeans remains the stronger choice on classical cluster-validity scores, and Conv-VaDE adds generative interpretability and representation-level capabilities that the standard pipeline does not provide.

Declarations

- **Funding:** The author received no specific funding for this research.
- **Conflict of interest:** The author declares that they have no conflict of interest.
- **Ethics approval:** This study is a secondary analysis of the publicly available and anonymised LEMON (Leipzig Mind-Brain-Body) dataset distributed by the Max Planck Institute for Human Cognitive and Brain Sciences. The original data collection was conducted with appropriate ethical oversight, and all participants provided informed consent.
- **Consent for publication:** Not applicable, as this study utilised a pre-existing, anonymised public dataset where individual-level data is not identifiable.
- **Data availability:** The LEMON dataset analysed during the current study is publicly available from the Max Planck Institute for Human Cognitive and Brain Sciences [41]. The specific preprocessed subset used in this study (203 participants, eyes-closed and eyes-open conditions, BIDS-formatted EEGLAB `.set`/`.fdt` files) is distributed at https://ftp.gwdg.de/pub/misc/MPI-Leipzig_Mind-Brain-Body-LEMON/EEG_MPILMBB_LEMON/EEG_Preprocessed_BIDS_ID/EEG_Preprocessed/. The full list of analysed participant identifiers is provided in Appendix B for exact reproducibility.
- **Code availability:** The code developed for the analysis in this study is available from the contributing author upon reasonable request.

- **Author contribution:** S.F. and L.L. contributed to conceptualization, methodology, validation, investigation, and supervision. S.F. conducted implementation, formal analysis, data curation, writing of the original draft, and visualisation. All authors reviewed and approved the final manuscript.

Appendix A GMM Implementation Details and Derivations

A.1 Parameterisation and Stability

To ensure numerical stability and satisfy the simplex constraint automatically, the mixture weights are computed in the log-space using a softmax function:

$$\pi_c = \frac{\exp(\theta_c)}{\sum_{j=1}^{K} \exp(\theta_j)} \tag{A1}$$

where $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_K]$ are unconstrained log-mixture weights initialised uniformly. Similarly, variances are specified in log-space to enforce positivity. To maintain a minimum level of uncertainty and prevent numerical instability, the variances are clipped:

$$\sigma_{c,i}^2 = \text{clip}\big(\exp(\log \sigma_{c,i}^2), \exp(-4.0), \exp(2.0)\big) \tag{A2}$$

A.2 Responsibility Computation

The term $\gamma_c(\mathbf{z})$ in the main objective function denotes the responsibility of cluster c for sample $\mathbf{z}$. This is computed via Bayes' theorem:

$$\gamma_c(\mathbf{z}) = \frac{\pi_c \cdot p(\mathbf{z}|\text{component } c)}{\sum_{j=1}^{K} \pi_j \cdot p(\mathbf{z}|\text{component } j)} \tag{A3}$$

The log-probability of a sample under a specific Gaussian component c is given by:

$$\log p(\mathbf{z}|\text{component } c) = -\frac{1}{2} \sum_{i=1}^{d} \left[\log(2\pi) + \log \sigma_{c,i}^2 + \frac{(z_i - \mu_{c,i})^2}{\sigma_{c,i}^2} \right] \tag{A4}$$

where the term $\frac{(z_i - \mu_{c,i})^2}{\sigma_{c,i}^2}$ represents the squared Mahalanobis distance.

A.3 Initialisation Strategy

To ensure the initial cluster positions reflect the actual data distribution, K-means clustering is applied to the encoded representations of a subset of training data. Multiple random initialisations ($n_{\text{init}} = 20$) provide the centroids, which then initialise the GMM parameters: $\boldsymbol{\theta} \leftarrow \log(\boldsymbol{\pi}_{\text{GMM}})$, $\boldsymbol{\mu}_c \leftarrow \boldsymbol{\mu}_{c,\text{GMM}}$, and $\log \boldsymbol{\sigma}_c^2 \leftarrow \log(\boldsymbol{\Sigma}_{c,\text{GMM}})$.

Appendix B Analysed Participant Identifiers

For exact reproducibility, Table B1 lists the 203 LEMON participant identifiers analysed in this study, drawn from the publicly distributed `EEG_Preprocessed_BIDS_ID/EEG_Preprocessed/` archive (https://ftp.gwdg.de/pub/misc/MPI-Leipzig_Mind-Brain-Body-LEMON/EEG_MPILMBB_LEMON/EEG_Preprocessed_BIDS_ID/EEG_Preprocessed/). All identifiers retain the original BIDS `sub-` prefix used in the public release; only the eyes-closed (`_EC`) recordings were used for the analyses reported in Section 4 (eyes-open replication is identified as a future-work item in Section 5.1).

Table B1 The 203 LEMON participant identifiers analysed in this study, ordered numerically (prefix `sub-0` omitted; for example, `10002` denotes `sub-010002`). Each entry corresponds to a paired `_EC.set`/`_EC.fdt` EEGLAB file in the public LEMON preprocessed archive.

10002	10003	10004	10005	10006	10007	10010	10012	10016	10017	10019	10020
10021	10022	10023	10024	10026	10027	10028	10029	10030	10031	10032	10033
10034	10035	10036	10037	10038	10039	10040	10041	10042	10044	10045	10046
10047	10048	10049	10050	10051	10052	10053	10056	10059	10060	10061	10062
10063	10064	10065	10066	10067	10068	10069	10070	10071	10072	10073	10074
10075	10076	10079	10080	10081	10083	10084	10085	10086	10088	10089	10090
10091	10092	10093	10094	10104	10126	10134	10136	10137	10138	10141	10142
10146	10148	10150	10152	10155	10157	10162	10163	10164	10165	10166	10168
10170	10176	10183	10191	10192	10194	10195	10196	10197	10199	10200	10201
10202	10204	10207	10210	10213	10214	10215	10218	10219	10220	10222	10223
10224	10226	10227	10228	10229	10230	10231	10232	10233	10234	10236	10238
10239	10240	10241	10242	10243	10244	10245	10246	10247	10248	10249	10250
10251	10252	10254	10255	10256	10257	10258	10260	10261	10262	10263	10264
10265	10266	10267	10268	10269	10270	10271	10272	10273	10274	10275	10276
10277	10283	10284	10286	10287	10288	10289	10290	10291	10292	10294	10295
10296	10297	10298	10299	10300	10301	10302	10303	10304	10305	10306	10307
10308	10309	10310	10311	10314	10315	10316	10317	10318	10319	10321	

Appendix C Distributional Diagnostics

The cluster-validity indices computed in pycrostates correlation space exhibit heavy-tailed and scale-discrepant distributions across the participants, particularly for Davies-Bouldin and Dunn (Section 4.1). These distributional properties motivate the use of non-parametric inference (Friedman test, paired Wilcoxon signed-rank) rather than parametric ANOVA. Per-metric per-K histograms and quantile-quantile plots are retained in the supplementary analysis pipeline outputs.

Appendix D Bootstrap Diagnostics

Bias-corrected and accelerated (BCa) bootstrap 95% confidence intervals on the participants' mean of every cluster-validity index at every K were computed with $B = 5000$ subject-resamples per cell. In a small minority of cells, BCa estimation failed (for example, due to degenerate acceleration estimates on near-constant resamples); these cells fall back to percentile-bootstrap intervals computed on the same resample set. The BCa-failure rate across the full grid is below 5% and does not concentrate at any particular K or metric.

Appendix E Comprehensive Performance Tables

This appendix presents architectural specifications and computational resource information for the 3654 person-specific Conv-VaDE models trained across the 203-participant LEMON participants at $K \in [3, 20]$. Per-subject metric distributions and statistical-test outputs are reported in Section 4 and in the supplementary analysis pipeline outputs (`analysis/` directory).

Table E2 Conv-VaDE Training Hyperparameters and Configuration

Category	Parameter	Value
Architecture	Latent Dimensions	32
	Encoder Channels	$32 \to 64 \to 128 \to 256$
	Decoder Channels	$256 \to 128 \to 64 \to 32 \to 1$
	Activation Function	Leaky ReLU (slope = 0.2)
Optimisation	Optimiser	Adam (dual: encoder/decoder + GMM)
	Encoder/Decoder Learning Rate	1×10^{-3}
	GMM Learning Rate	5×10^{-4}
	Weight Decay	1×10^{-5}
	Batch Size	128
	Training Epochs (max)	100
	LR Scheduler	ReduceLROnPlateau (patience = 10, factor = 0.5)
Regularisation	Dropout Rate	0.2 (encoder only)
	Batch Normalisation	All conv layers except first
	Gradient Clipping	max_norm = 5.0
	Early Stopping	patience = 15 epochs
Loss Function	Reconstruction Loss	MSE (polarity-invariant)
	Regularisation Loss	KL divergence vs. GMM prior
	β Annealing	Cyclical, $\beta_{\min} = 0.01$, $\beta_{\max} = 0.1$, 4 cycles over training
	λ_{recon}	1.0 (fixed)
	λ_{ent}	$0.3/\log K \cdot \text{warmup}(t)$
	λ_{sep}	$50.0 \cdot \text{warmup}(t)$
	λ_{batch}	$5.0 \cdot \text{warmup}(t)$
	λ_{tight}	$0.2 \cdot \text{warmup}(t)$
	λ_{pol}	0.1 (active throughout)
	Warm-up schedule	Linear ramp $\text{warmup}(t) \in [0, 1]$ over 10 epochs of joint training; 0 during pretraining
GMM Prior	K Range Tested	3–20 (18 configurations)
	Covariance Type	Diagonal
	Initialisation	K-means ($n_{\text{init}} = 20$) on encoded latents
Data	Train/Evaluation Split	90 / 10 (within-subject GFP peaks per participant)
	Random Seed	42 (reproducibility)

Table E3 CNN-VAE Encoder and Decoder Layer Specifications

ID	Type	Input Shape	Output Shape	K	S	Activation
Encoder Pipeline						
E1	Conv2d	$1 \times 40 \times 40$	$32 \times 20 \times 20$	4	2	L.ReLU + Drop(0.2)
E2	Conv2d + BN	$32 \times 20 \times 20$	$64 \times 10 \times 10$	4	2	L.ReLU
E3	Conv2d + BN	$64 \times 10 \times 10$	$128 \times 5 \times 5$	4	2	L.ReLU + Drop(0.2)
E4	Conv2d + BN	$128 \times 5 \times 5$	$256 \times 2 \times 2$	4	2	L.ReLU
E5	AdaptAvgPool	$256 \times 2 \times 2$	$256 \times 1 \times 1$	–	–	–
E6	Conv2d (1×1)	$256 \times 1 \times 1$	$1024 \times 1 \times 1$	1	1	L.ReLU
E7	Flatten	$1024 \times 1 \times 1$	1024	–	–	–
E8	Linear	1024	512	–	–	L.ReLU
E9	Linear (μ)	512	32	–	–	Linear
E10	Linear ($\log \sigma^2$)	512	32	–	–	Linear
Decoder Pipeline						
D1	Linear	32	512	–	–	L.ReLU
D2	Linear	512	1024	–	–	L.ReLU
D3	Reshape	1024	$1024 \times 1 \times 1$	–	–	–
D4	ConvTrans + BN	$1024 \times 1 \times 1$	$256 \times 4 \times 4$	4	1	L.ReLU + Drop(0.2)
D5	ConvTrans + BN	$256 \times 4 \times 4$	$128 \times 8 \times 8$	4	2	L.ReLU
D6	ConvTrans + BN	$128 \times 8 \times 8$	$64 \times 16 \times 16$	4	2	L.ReLU
D7	ConvTrans + BN	$64 \times 16 \times 16$	$32 \times 32 \times 32$	4	2	L.ReLU
D8	ConvTrans	$32 \times 32 \times 32$	$1 \times 64 \times 64$	4	2	Linear

Description: The Encoder compresses the 40×40 input maps into a 32-dim latent distribution using increasing channel depth (32 to 1024). The Decoder symmetrically mirrors this using transposed convolutions; its final transposed-convolution output (64×64) is interpolated to the 40×40 target. *Legend:* **K** = Kernel Size; **S** = Stride; **BN** = Batch Normalisation; **Drop** = Dropout; **L.ReLU** = LeakyReLU; **ConvTrans** = ConvTranspose2d.

Table E4 EEG Preprocessing Pipeline for the LEMON Dataset (eyes-closed condition)

Phase	Step	Input	Output	Details
Data loading	1. Load LEMON	.fif/.edf	RawArray	MNE-Python; 61 scalp + 1 VEOG
	2. Reference	RawArray	RawArray	Common average reference
	3. Filter	RawArray	RawArray	FIR bandpass 2–20 Hz (zero-phase)
GFP-peak extraction	4. GFP signal	61-ch ts	GFP(t)	Pycrostates: spatial std at each t
	5. Peak detection	GFP(t)	$\{t_i^\star\}$	Local maxima, min spacing ≥ 3 samples
	6. Per-peak topo	61-ch ts	61-D vector	Sample at each $t_i^\star$
Topo map gen.	7. 3D $\rightarrow$ 2D	61×3	61×2	Azimuthal equidistant projection
	8. Interpolation	61×2	40×40	Cubic spline; outside hull = 0
	9. Stack	N peaks	$N \times 40 \times 40$	One map per GFP peak
Norm. & split	10. z-score	$N \times 40 \times 40$	$N \times 40 \times 40$	Train-only $\bar{x}, \sigma$; clip $\pm 5\sigma$
	11. Split	100%	90/10%	Train / Evaluation per participant

Note: N = number of GFP-peak samples per participant after extraction. LEMON eyes-closed recordings comprise 61 scalp electrodes (10–10 montage) sampled at 250 Hz. Topographic maps are extracted at local maxima of the Global Field Power signal. Final topographic maps are 40×40 pixels representing the azimuthally-projected, cubic-interpolated, z-score-normalised scalp potential distributions. Polarity invariance is enforced downstream of preprocessing through sign-flip augmentation, a polarity-invariant reconstruction loss, and an encoder regularisation term (Section 3).

Table E5 Computational Resources and Training Efficiency

Category	Resource	Details
Hardware	GPU	NVIDIA GPU (GPU-accelerated training)
	System RAM	Sufficient for 3.2M+ samples
	Storage	SSD recommended (∼50 GB)
Software	Framework	PyTorch (v2.5)
	Signal Processing	MNE-Python (v1.8)
	Machine Learning	scikit-learn (v1.5)
	Scientific Comp.	SciPy (v1.14)
	Language	Python (v3.12)
Training Efficiency	Single Model	∼12–48 hrs/model (variable)
	Total Time	3654 models (203 participants $\times$ 18 K values)
	Epochs	≤ 100 (median 19; early stopping)

Note: Training times vary based on K-value complexity and convergence speed. GPU specifications omitted due to variations in computational constraints across runs.

References

[1] Müller-Putz, G.R.: Electroencephalography. In: Handbook of Clinical Neurology vol. 168, pp. 249–262. Elsevier B.V., ??? (2020). https://doi.org/10.1016/B978-0-444-63934-9.00018-4

[2] Craik, A., He, Y., Contreras-Vidal, J.L.: Deep learning for electroencephalogram (EEG) classification tasks: a review. Journal of Neural Engineering **16**(3), 031001 (2019) https://doi.org/10.1088/1741-2552/ab0ab5

[3] Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T.H., Faubert, J.: Deep learning-based electroencephalography analysis: a systematic review. Journal of Neural Engineering **16**(5), 051001 (2019) https://doi.org/10.1088/1741-2552/ab260c

[4] Michel, C.M., Koenig, T.: EEG microstates as a tool for studying the temporal dynamics of whole-brain neuronal networks: A review. NeuroImage **180**, 577–593 (2018) https://doi.org/10.1016/j.neuroimage.2017.11.062

[5] Nishida, K., Morishima, Y., Yoshimura, M., Isotani, T., Irisawa, S., Jann, K., Dierks, T., Strik, W., Kinoshita, T., Koenig, T.: EEG microstates associated with salience and frontoparietal networks in frontotemporal dementia, schizophrenia and Alzheimer's disease. Clinical Neurophysiology **124**(6), 1106–1114 (2013) https://doi.org/10.1016/j.clinph.2013.01.005

[6] Jang, J.H., Kim, T.Y., Lim, H.S., Yoon, D.: Unsupervised feature learning for electrocardiogram data using the convolutional variational autoencoder. PLoS ONE **16**(12), 0260612 (2021) https://doi.org/10.1371/journal.pone.0260612

[7] Bethge, D., Hallgarten, P., Grosse-Puppendahl, T., Kari, M., Chuang, L.L., Ozdenizci, O., Schmidt, A.: EEG2Vec: Learning affective EEG representations via variational autoencoders. In: Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, pp. 3150–3157. IEEE, ??? (2022). https://doi.org/10.1109/SMC53654.2022.9945517

[8] Tabejamaat, M., Mohammadzade, H., Negin, F., Bremond, F.: EEG classification with limited data: A deep clustering approach. Pattern Recognition **157**, 110934 (2025) https://doi.org/10.1016/j.patcog.2024.110934

[9] Jiang, Z., Zheng, Y., Tan, H., Tang, B., Zhou, H.: Variational Deep Embedding: An Unsupervised and Generative Approach to Clustering (2017). https://arxiv.org/abs/1611.05148

[10] Wackermann, J., Lehmann, D., Michel, C.M., Strik, W.K.: Adaptive segmentation of spontaneous EEG map series into spatially defined microstates. International Journal of Psychophysiology **14**(3), 269–283 (1993) https://doi.org/10.1016/0167-8760(93)90041-M

[11] Férat, V., Scheltienne, M., Brunet, D., Ros, T., Michel, C.: Pycrostates: a python library to study EEG microstates. Journal of Open Source Software **7**(78), 4564 (2022) https://doi.org/10.21105/joss.04564

[12] Faremi, S., Visentin, A., Longo, L.: Interpretable EEG Microstate Discovery via Variational Deep Embedding: A Systematic Architecture Search with Multi-Quadrant Evaluation (2026). https://arxiv.org/abs/2605.10947

[13] Gómez-Tapia, C., Bozic, B., Longo, L.: On the minimal amount of EEG data required for learning distinctive human features for task-dependent biometric applications. Frontiers in Neuroinformatics **16**, 844667 (2022) https://doi.org/10.3389/fninf.2022.844667

[14] Longo, L.: Modeling cognitive load as a self-supervised brain rate with electroencephalography and deep learning. Brain Sciences **12**(10) (2022) https://doi.org/10.3390/brainsci12101416

[15] Rakhmatulin, I., Dao, M.S., Nassibi, A., Mandic, D.: Exploring convolutional neural network architectures for EEG feature extraction. Sensors **24**(3), 877 (2024) https://doi.org/10.3390/s24030877

[16] Zhang, X., Yao, L., Wang, X., Monaghan, J., McAlpine, D., Zhang, Y.: A survey on deep learning-based non-invasive brain signals: recent advances and new frontiers. Journal of Neural Engineering **18**(3), 031002 (2021) https://doi.org/10.1088/1741-2552/abc902

[17] Zhao, T., Cui, Y., Ji, T., Luo, J., Li, W., Jiang, J., Gao, Z., Hu, W., Yan, Y., Jiang, Y., Hong, B.: VAEEG: Variational auto-encoder for extracting EEG representation. NeuroImage **304**, 120946 (2024) https://doi.org/10.1016/j.neuroimage.2024.120946

[18] Cisotto, G., Zancanaro, A., Zoppis, I.F., Manzoni, S.L.: hvEEGNet: a novel deep learning model for high-fidelity EEG reconstruction. Frontiers in Neuroinformatics **18**, 1459970 (2024) https://doi.org/10.3389/fninf.2024.1459970

[19] Ahmed, T., Longo, L.: Examining the size of the latent space of convolutional variational autoencoders trained with spectral topographic maps of EEG frequency bands. IEEE Access **10**, 107575–107586 (2022) https://doi.org/10.1109/ACCESS.2022.3212777

[20] Chikkankod, A.V., Longo, L.: On the dimensionality and utility of convolutional autoencoder's latent space trained with topology-preserving spectral EEG headmaps. Machine Learning and Knowledge Extraction **4**(4), 1042–1064 (2022) https://doi.org/10.3390/make4040053

[21] Liang, S., Zhang, X., Zhao, H., Dang, Y., Hui, R., Zhang, J.: Double discrete variational autoencoder for epileptic EEG signals classification. IEEE Access **12**,

106567–106578 (2024) https://doi.org/10.1109/ACCESS.2024.3429195

[22] Ellis, C.A., Miller, R.L., Calhoun, V.D.: A convolutional autoencoder-based explainable clustering approach for resting-state EEG analysis. In: 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 1–4 (2023). https://doi.org/10.1109/EMBC40787.2023.10340375

[23] Yue, Y., Deng, J.D., Chakraborti, T., De Ridder, D., Manning, P.: Unsupervised hybrid deep feature encoder for robust feature learning from resting-state EEG data. In: 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 1–5 (2024). https://doi.org/10.1109/EMBC53108.2024.10781741

[24] Bollens, L., Francart, T., Hamme, H.V.: Learning subject-invariant representations from speech-evoked EEG using variational autoencoder. In: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1256–1260. IEEE, ??? (2022). https://doi.org/10.1109/icassp43922.2022.9747297 . http://dx.doi.org/10.1109/ICASSP43922.2022.9747297

[25] Ahmed, T., Longo, L.: Interpreting disentangled representations of person-specific convolutional variational autoencoders of spatially preserving EEG topographic maps via clustering and visual plausibility. Information **14**(9) (2023) https://doi.org/10.3390/info14090489

[26] Criscuolo, S., Apicella, A., Prevete, R., Longo, L.: Interpreting the latent space of a convolutional variational autoencoder for semi-automated eye blink artefact detection in EEG signals. Computer Standards & Interfaces **92**, 103897 (2025) https://doi.org/10.1016/j.csi.2024.103897

[27] Ren, Y., Pu, J., Yang, Z., Xu, J., Li, G., Pu, X., Yu, P.S., He, L.: Deep clustering: A comprehensive survey. IEEE Transactions on Neural Networks and Learning Systems **36**(4), 5858–5878 (2025) https://doi.org/10.1109/TNNLS.2024.3403155

[28] Guo, J., Fan, W., Amayri, M., Bouguila, N.: Deep clustering analysis via variational autoencoder with gamma mixture latent embeddings. Neural Networks **183**, 106979 (2025) https://doi.org/10.1016/j.neunet.2024.106979

[29] Lin, C.-T., Huang, K.-C., Yang, W.-Y., Singh, A.K., Chuang, C.-H., Wang, Y.-K.: Real-time EEG signal enhancement using canonical correlation analysis and Gaussian mixture clustering. Journal of Healthcare Engineering **2018**, 5081258 (2018) https://doi.org/10.1155/2018/5081258

[30] Pan, S., Shen, T., Lian, Y., Shi, L.: A task-related EEG microstate clustering algorithm based on spatial patterns, Riemannian distance, and a deep autoencoder. Brain Sciences **15**(1) (2025) https://doi.org/10.3390/brainsci15010027

[31] Khanna, A., Pascual-Leone, A., Michel, C.M., Farzan, F.: Microstates in resting-state EEG: current status and future directions. Neuroscience and Biobehavioral Reviews **49**, 105–113 (2015) https://doi.org/10.1016/j.neubiorev.2014.12.010

[32] Tarailis, P., Koenig, T., Michel, C.M., Griskova-Bulanova, I.: The functional aspects of resting EEG microstates: A systematic review. Brain Topography **37**(2), 181–217 (2024) https://doi.org/10.1007/s10548-023-00958-9

[33] Seitzman, B.A., Abell, M., Bartley, S.C., Erickson, M.A., Bolbecker, A.R., Hetrick, W.P.: Cognitive manipulation of brain electric microstates. NeuroImage **146**, 533–543 (2017) https://doi.org/10.1016/j.neuroimage.2016.10.002

[34] Dinov, M., Leech, R.: Modeling uncertainties in EEG microstates: Analysis of real and imagined motor movements using probabilistic clustering-driven training of probabilistic neural networks. Frontiers in Human Neuroscience **11**, 534 (2017) https://doi.org/10.3389/fnhum.2017.00534

[35] EskandariNasab, M., Raeisi, Z., Lashaki, R.A., Najafi, H.: A GRU–CNN model for auditory attention detection using microstate and recurrence quantification analysis. Scientific Reports **14**(1), 8861 (2024) https://doi.org/10.1038/s41598-024-58886-y

[36] Brodbeck, V., Kuhn, A., von Wegner, F., Morzelewski, A., Tagliazucchi, E., Borisov, S., Michel, C.M., Laufs, H.: Eeg microstates of wakefulness and NREM sleep. NeuroImage **62**(3), 2129–2139 (2012) https://doi.org/10.1016/j.neuroimage.2012.05.060

[37] Strik, W.K., Dierks, T., Becker, T., Lehmann, D.: Larger topographical variance and decreased duration of brain electric microstates in depression. Journal of Neural Transmission **99**(1-3), 213–222 (1995) https://doi.org/10.1007/BF01271480

[38] Fan, Y., Wen, G., Li, D., Qiu, S., Levine, M.D., Xiao, F.: Video anomaly detection and localization via Gaussian mixture fully convolutional variational autoencoder. Computer Vision and Image Understanding **195**, 102920 (2020) https://doi.org/10.1016/j.cviu.2020.102920

[39] Csore, J., Le Roy, T., Wright, G., Karmonik, C.: Unsupervised classification of multi-contrast magnetic resonance histology of peripheral arterial disease lesions using a convolutional variational autoencoder with a Gaussian mixture model in latent space: A technical feasibility study. Computerized Medical Imaging and Graphics **115**, 102372 (2024) https://doi.org/10.1016/j.compmedimag.2024.102372

[40] Han, J., Gu, X., Yang, G.-Z., Lo, B.: Noise-factorized disentangled representation learning for generalizable motor imagery EEG classification. IEEE Journal of Biomedical and Health Informatics **28**(2), 765–776 (2024) https://doi.org/10.1109/JBHI.2023.3337072

[41] Babayan, A., Erbey, M., Kumral, D., Reinelt, J.D., Reiter, A.M.F., Röbbig, J., Schaare, H.L., Uhlig, M., Anwander, A., Bazin, P.-L., *et al.*: A mind-brain-body dataset of MRI, EEG, cognition, emotion, and peripheral physiology in young and old adults. Scientific Data **6**(1), 180308 (2019) https://doi.org/10.1038/sdata.2018.308

[42] Fu, H., Li, C., Liu, X., Gao, J., Celikyilmaz, A., Carin, L.: Cyclical annealing schedule: A simple approach to mitigating KL vanishing. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), pp. 240–250 (2019). https://doi.org/10.18653/v1/N19-1021

[43] Allen, M., Poggiali, D., Whitaker, K., Marshall, T.R., Langen, J., Kievit, R.A.: Raincloud plots: a multi-platform tool for robust data visualization. Wellcome Open Research **4**, 63 (2019) https://doi.org/10.12688/wellcomeopenres.15191.2