Pattern Recognition Letters 208 (2026) 69–85

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Decision-rule alignment in generative latent spaces: The role of radial compatibility

Víctor Poblete [a,*], José Aillapi [b], Carlos Duarte [b], Luis Veas [b], Felipe Otondo [c]

[a] Data and Audio Mining Laboratory, Institute of Acoustics, Universidad Austral de Chile, Valdivia, 5110566, Chile
[b] Institute of Informatics, Universidad Austral de Chile, Valdivia, 5110566, Chile
[c] Institute of Acoustics, Universidad Austral de Chile, Valdivia, 5110566, Chile

ARTICLE INFO

Edited by: Yonghuai Liu

Keywords:
Variational autoencoders
Posterior mean embeddings
Latent geometry
Radial compatibility
Decision-rule alignment
Classifier ranking
Representation learning

ABSTRACT

Deep generative encoders are frequently reused as fixed feature extractors for downstream classification, but the suitability of a classifier depends on whether its decision assumptions align with the geometry of the learned representation. We study this compatibility problem by using deterministic posterior mean embeddings extracted from variational autoencoders. Within each dataset, the encoder, embeddings, preprocessing, cross-validation partitions, and held-out samples are fixed, while only the downstream decision rule varies. We compare a centroid-based radial rule, a Gaussian maximum a posteriori (MAP) classifier with shrinkage-regularized covariance estimation, logistic regression, and eXtreme Gradient Boosting (XGBoost) across eight heterogeneous acoustic and visual benchmarks. The relative ordering of these decision rules is not invariant: Gaussian MAP performs best on the amphibian vocalization dataset, logistic regression leads on UrbanSound8K and ESC-50, and XGBoost ranks first across the five visual datasets. To interpret this variability, we characterize class-conditional geometry in the posterior mean embedding space and define radial compatibility as the joint presence of within-class isotropy and radial separation from competing classes. This class-centered formulation links the assumptions of centroid-based radial classification to measurable geometric properties of the representation. The cross-dataset geometry–performance analysis is interpreted descriptively because only eight dataset-level observations are available. Overall, the results show that downstream classifier effectiveness cannot be attributed solely to intrinsic model flexibility but should instead be understood as a compatibility problem between decision-rule assumptions and the geometry induced by the fixed generative representation.

1. Introduction

When a deep generative model induces a structured latent representation, representation learning and downstream decision-making become conceptually separable. For a given dataset D, let f_D denote a trained and subsequently frozen encoder that maps an input x to a deterministic posterior mean embedding $\mathrm{h} = f_D(\mathrm{x}) \in \mathbb{R}^{d_z}$, where d_z is the latent dimensionality. Once this representation is fixed, downstream classification depends not only on the information contained in the embeddings but also on how well the assumptions of the selected decision rule align with their class-conditional geometry. This separation raises a fundamental compatibility question: Given a fixed generative representation, which decision paradigm is most appropriate for the geometry induced in its embedding space?

We compare three families of decision strategies applied to identical posterior mean embeddings within each dataset: geometric, probabilistic, and discriminative. These paradigms impose distinct structural assumptions. A centroid-based radial rule represents each class through an approximately isotropic region centered on its prototype. A Gaussian maximum a posteriori (MAP) classifier models class-conditional densities with an explicit covariance structure and therefore accommodates anisotropic dispersion. Discriminative models learn to separate boundaries directly from labeled embeddings, with fewer explicit assumptions about class shape.

The relevance of comparing these paradigms is therefore structural rather than merely empirical. A radial rule may be suitable when class-conditional embeddings form approximately isotropic, prototype-centered regions that remain sufficiently separated from competing classes. When the covariance structure is anisotropic, a Gaussian MAP rule may better represent the orientation and scale of class dispersion.

* Corresponding author.
E-mail address: vpoblete@uach.cl (V. Poblete).

https://doi.org/10.1016/j.patrec.2026.07.024
Received 27 March 2026; Received in revised form 11 July 2026; Accepted 15 July 2026
Available online 16 July 2026

When class regions are irregular, entangled, or strongly overlapping, flexible discriminative boundaries may become advantageous. Under this perspective, performance differences reflect not only model capacity but also compatibility between decision-rule assumptions and the geometry of the fixed representation.

We formalize this idea through the notion of radial compatibility. Radial compatibility denotes the extent to which class-conditional embeddings can be adequately represented by approximately isotropic regions centered on their class prototypes while remaining sufficiently separated from competing radial regions. This definition combines two complementary properties: within-class isotropy and between-class radial separation. It therefore relates directly to the assumptions imposed by centroid-based radial classification rather than to the orientation of embeddings relative to the global latent origin.

To investigate this compatibility problem, we conduct a controlled cross-domain comparison across eight heterogeneous acoustic and visual datasets. Variational autoencoders are used as representative generative models, which can induce structured latent representations through compression and regularization. Within each dataset, the trained encoder, posterior mean embeddings, preprocessing pipeline, cross-validation partitions, and evaluation protocol are fixed, while only the downstream decision rule varies. This design isolates the decision layer from representation learning and enables us to examine whether classifier ordering remains invariant across datasets and whether variations in classifier advantage are associated with measurable properties of class-conditional embedding geometry.

The study makes three principal contributions. First, it provides a controlled comparison of geometric, probabilistic, and discriminative decision paradigms operating on fixed deterministic posterior mean embeddings. Second, it evaluates whether the relative ranking of these paradigms is invariant across heterogeneous datasets. Third, it introduces a centroid-based formulation of radial compatibility that combines within-class isotropy with radial separation from competing classes and uses this formulation together with complementary spectral descriptors to examine geometry–performance associations across domains.

1.1. Formal hypothesis statement

Let $\mathscr{D} = \{D_1, \ldots, D_M\}$ denote the set of datasets considered in the empirical study. For each dataset $D \in D$, let N_D denote its number of samples, let f_D denote its trained and frozen generative encoder, and let

$$h_i^{(D)} = f_D\left(\mathrm{x}_i^{(D)}\right) \in \mathbb{R}^{d_z},\ i = 1, \ldots,\ N_D \tag{1}$$

denote the deterministic posterior mean embedding of the *i-th* sample in the dataset D. Within each dataset, all decision paradigms operate on the same embeddings and identical cross-validation partitions.

Let $\mathscr{P}$ denote the set of decision paradigms, comprising a centroid-based radial rule, a Gaussian MAP classifier, and discriminative models. Let $P_{p,D}$ denote the mean predictive performance obtained by paradigm $p \in P$ on dataset D, and let $R_{p,D}$ denote its corresponding within-dataset rank.

Hypothesis 1. (H1) — Ranking non-invariance

We hypothesize that the relative ranking of decision paradigms varies across datasets. Formally, at least two datasets $D_a, D_b \in \mathscr{D}$ exist for which

$$\left(R_{p,D_a}\right)_{p\in\mathscr{P}} \neq \left(R_{p,D_b}\right)_{p\in\mathscr{P}}. \tag{2}$$

Thus, even when all decision rules are evaluated on an identical fixed representation within each dataset, their relative ordering is expected to change across datasets rather than remain universally constant.

Hypothesis 2. (H2) — Geometry–performance association

For each dataset D, let g_D denote a vector of intrinsic descriptors of its posterior mean embedding geometry. These descriptors characterize properties such as within-class isotropy, radial separation between competing classes, radial compatibility, covariance anisotropy, spectral decay, and effective dimensionality. Importantly, g_D contains only quantities computed from the embeddings and class labels; it does not include classifier performance.

For two decision paradigms p and q, their performance contrast on dataset D is defined as

$$\Delta_{p,q}^{(D)} = P_{p,D} - P_{q,D}. \tag{3}$$

We hypothesize that the cross-dataset variation in at least one performance contrast is systematically associated with at least one intrinsic geometric descriptor. Formally, for some component $g_{D,k}$ of g_D,

$$\rho\left(g_{D,k},\ \Delta_{p,q}^{(D)}\right) \neq 0 \tag{4}$$

where ρ denotes a rank-based association measure computed across datasets.

H1 concerns whether classifier ordering changes across datasets, whereas H2 examines whether the magnitude of these performance differences covaries with measurable properties of the fixed posterior mean geometry. Given the limited number of datasets, the geometry–performance analysis is interpreted primarily through effect sizes, consistency across complementary descriptors, and cautious statistical inference rather than broad population-level generalization.

The conceptual implications of these hypotheses are summarized in Fig. 1. The upper panels illustrate a progression from approximately isotropic and well-separated class regions to anisotropic and eventually irregular or overlapping structures. The lower panel depicts the corresponding conceptual variation in decision-rule suitability. The curves are schematic and do not represent empirical performance values; rather, they illustrate how the assumptions of radial, Gaussian, and discriminative decision rules may become differently aligned with class-conditional geometry.

Under this perspective, classifier superiority is geometry dependent rather than paradigm intrinsic. Downstream classification in generative embedding spaces can therefore be interpreted as a compatibility problem between the structural assumptions of a decision rule and the geometry induced by the fixed representation. This view positions this study at the intersection of generative representation learning, geometric characterization, and decision modeling.

Recent research has begun to clarify how compression and regularization shape generative representations. D'Amato et al. [1], for example, relate β-VAEs to rate–distortion theory and show that compression constraints influence the structure of learned latent representations. Recent self-supervised ecoacoustic models [2] similarly demonstrate that generative encoders can produce reusable latent descriptions of complex acoustic observations. These developments motivate an examination of not only how embeddings are learned but also how their geometry interacts with downstream decision rules.

This study does not assume a universal progression from particular datasets to predefined geometric regimes. Instead, the relevant geometric properties are measured directly from the corrected posterior mean embeddings. Fig. 1 therefore presents only a conceptual framework: Approximately isotropic and separated classes are expected to be more compatible with radial prototype-based decisions, whereas anisotropic covariance, irregular boundaries, and class overlap may favor probabilistic or discriminative alternatives. The validity and strength of these relationships are evaluated empirically across the benchmark suite.

The remainder of this paper is organized as follows. Section 2 reviews related work on generative embedding geometry and decision modeling. Section 3 formalizes the posterior mean representation, the decision paradigms, radial compatibility, and the geometry–performance analysis. Section 4 describes the cross-domain

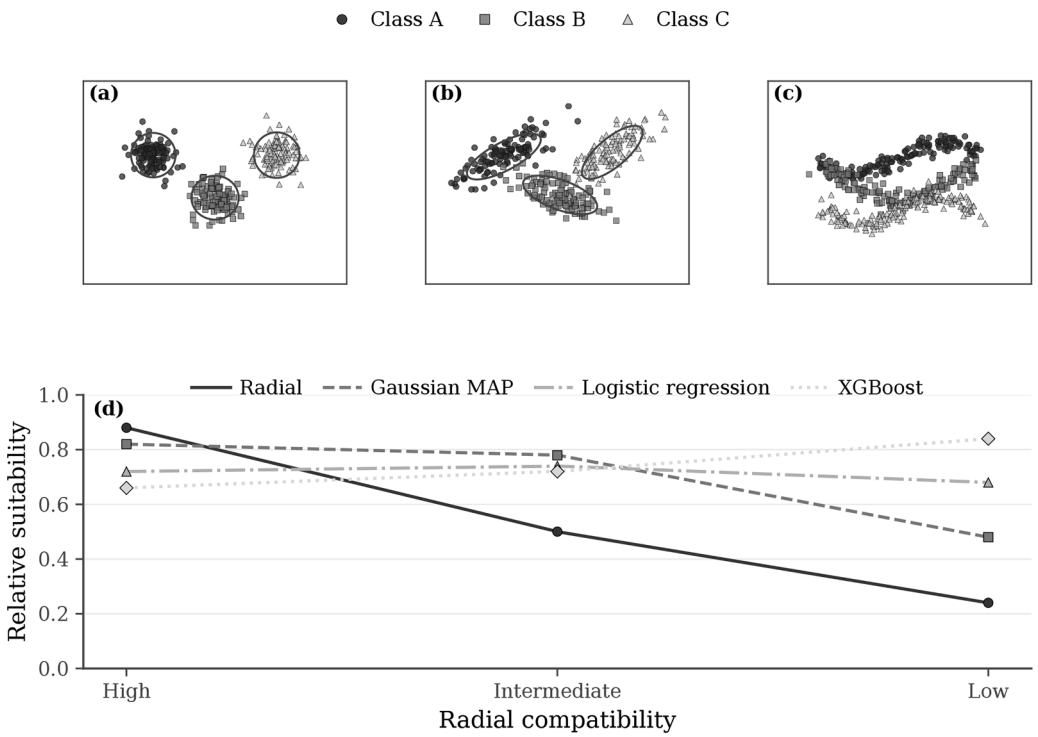

Fig. 1. Conceptual relationship between class geometry, radial compatibility, and decision-rule suitability in generative embedding spaces. The upper panels illustrate three schematic class-conditional organizations in a fixed latent representation: (a) approximately isotropic and well-separated regions, corresponding to high radial compatibility; (b) anisotropic but still separated regions, corresponding to intermediate radial compatibility; and (c) irregular and partially overlapping class regions, corresponding to low radial compatibility. The lower panel summarizes the expected relative suitability of the four decision rules across this progression. Centroid-based radial classification is most compatible with approximately spherical and well-separated class regions, whereas Gaussian MAP and discriminative models become comparatively more suitable as covariance anisotropy, boundary irregularity, and class overlap increase. The curves are schematic and are intended to illustrate the compatibility between decision-rule assumptions and latent geometry rather than empirical performance values.

experimental protocol. Section 5 presents the empirical results, including classifier ranking variability and geometry–performance associations. Section 6 discusses the structural implications, limitations, and directions for future research.

2. Related work

Understanding the compatibility between generative representations and downstream decision rules first requires examining how generative models shape latent geometry. Deep generative models, including variational autoencoders (VAEs) and adversarial autoencoders (AAEs), learn compressed representations whose organization depends on the model architecture, prior distribution, regularization objective, and statistical properties of the training data. Previous studies consequently examined latent spaces in terms of distributional structure, disentanglement, geometric consistency, and stability across generative formulations.

Kaechele et al. [3] compared alternative latent modeling strategies within autoencoder frameworks and showed that modeling choices influence both representation structure and reconstruction behavior. Asperti and Tonelli [4] investigated structural similarities across generative architectures, emphasizing geometric coherence and stability in learned latent spaces. More recently, D'Amato et al. [1] related β-VAEs to rate–distortion theory, demonstrating that compression constraints affect how information and variance are distributed across latent dimensions.

These contributions establish that latent geometry is not an incidental by-product of representation learning; rather, it is a model- and dataset-dependent consequence of generative compression and regularization. Nevertheless, most prior work concentrated on the intrinsic organization of the learned representation rather than on the suitability of alternative downstream decision rules after the encoder has been fixed. In particular, limited attention has been given to whether class-conditional properties such as within-class isotropy, covariance anisotropy, prototype separation, and overlap between class regions systematically favor different decision paradigms.

This study addresses this distinction by treating the posterior mean as the deterministic positional representation used for downstream geometry and classification. The posterior variance remains part of the probabilistic encoder and quantifies uncertainty about latent position, but it is not concatenated with the posterior mean to define Euclidean class geometry. This separation provides a consistent basis for comparing centroid-based radial, Gaussian, and discriminative decision rules while retaining the posterior variance in its proper role as an uncertainty descriptor.

2.1. Geometry of decision regions in learned representations

Beyond the geometry induced by generative modeling itself, several studies have examined how class regions are organized within learned representations. Těťková et al. [5] analyzed the convexity of class regions in deep latent spaces and showed that approximately convex organization can emerge in practice. Ganguli et al. [6] investigated statistically disentangled latent representations and explored how generative factors influence geometric interpretability and class separability.

These studies support the broader view that downstream classification is affected by the structural properties of the learned representation. However, different geometric descriptors capture distinct aspects of class organization and should not be treated as interchangeable. Within-class isotropy characterizes whether dispersion is approximately uniform across latent directions; covariance anisotropy reflects directional

variation in scale; prototype separation measures the relative displacement of class centroids; and overlap describes the extent to which competing class regions intersect. Each of these properties may interact differently with the assumptions imposed by a decision rule.

For centroid-based radial classification, compatibility depends on whether class-conditional embeddings can be adequately represented by approximately spherical regions centered on their prototypes and whether those regions remain sufficiently separated. Gaussian class-conditional models relax the isotropic assumption by incorporating covariance orientation and scale, whereas discriminative classifiers can accommodate decision boundaries that are not directly described by a simple metric or parametric density model.

Although previous work established that representation geometry influences separability and decision-region structure, most studies either characterize geometry independently of downstream classification or analyze only a single classifier family. Controlled comparisons of structurally distinct decision paradigms operating on identical generative embeddings remain limited. Consequently, whether measurable class-conditional geometry can account for changes in classifier ranking and relative advantage across datasets is still unclear.

2.2. Decision strategies in latent spaces

Once representation learning has been completed, latent embeddings are frequently reused as fixed feature spaces for downstream decision tasks. A common strategy is to train discriminative classifiers directly on the embeddings, allowing the decision boundary to be learned from labeled examples without explicitly modeling the class-conditional distribution. Alternatively, probabilistic approaches assume parametric densities in latent space and assign class labels according to posterior probability. Simpler geometric rules, including nearest-centroid and radius-based decisions, are also attractive because of their interpretability and low computational cost.

These alternatives differ fundamentally in how they represent class structure. Centroid-based geometric rules rely primarily on Euclidean proximity to class prototypes and therefore impose strong assumptions regarding isotropy and radial separation. Gaussian MAP classifiers retain a parametric interpretation but introduce class-specific covariance, allowing anisotropic and correlated dispersion to be represented through a Mahalanobis geometry. Discriminative models such as logistic regression and gradient-boosted trees learn to separate functions directly from the embedding coordinates, thereby reducing the need to specify an explicit generative model for each class.

Latent representations have also been used in monitoring, anomaly detection, and interpretability settings. Komorska et al. [7], for example, employed VAE-derived embeddings for anomaly detection in condition monitoring, and Guidotti et al. [8] used latent neighborhoods to construct locally interpretable decision models. These applications illustrate the versatility of learned representations, but they typically adopt a single downstream strategy chosen for a particular task or domain.

As a result, the relative merits of geometric, probabilistic, and discriminative decision rules are rarely examined under controlled representational conditions. In many studies, differences in classifier performance are confounded with variations in feature extraction, representation quality, preprocessing, or model capacity. This work instead evaluates structurally distinct decision paradigms on a common posterior mean representation within each dataset. This controlled setting enables performance differences to be interpreted more directly in terms of the compatibility between decision-rule assumptions and class-conditional embedding geometry.

2.3. Comparative classifier studies

Comparative evaluations of classification algorithms have repeatedly shown that classifier rankings can vary substantially across datasets. Large-scale benchmark studies, such as that of Fernandez-Delgado et al. [9], demonstrate that no single classifier family is uniformly superior across heterogeneous learning problems. Such variability has motivated broad empirical comparisons involving probabilistic, discriminative, tree-based, kernel-based, and ensemble methods.

However, general-purpose classifier benchmarks typically vary several components simultaneously. Feature extraction, preprocessing, representation dimensionality, hyperparameter optimization, and classifier capacity may all differ across methods or datasets. Consequently, any observed performance differences cannot always be attributed specifically to the decision rule because they may also reflect changes in the information or geometry presented to the classifier.

This distinction is especially important when classification is performed on learned embeddings. A classifier may appear superior because it operates on a more favorable representation and not because its structural assumptions are better aligned with the underlying class geometry. Conversely, a simple rule may remain competitive when the representation already organizes classes in a form compatible with that rule. Separating these effects requires an experimental design in which the representation is controlled and only the downstream decision mechanism varies.

This study adopts this controlled perspective. Within each dataset, all classifiers receive the same deterministic posterior mean embeddings and are evaluated using identical cross-validation partitions. The comparison therefore focuses on relative decision-rule suitability rather than unrestricted algorithmic benchmarking. By examining both predictive performance and within-dataset ranking, we assess whether classifier ordering remains stable across domains or changes as the geometry of the fixed representation varies.

2.4. Positioning of this study

Building on the distinctions outlined above, this study explicitly separates representation learning from downstream decision modeling. Within each dataset, the encoder of a trained VAE is frozen and used to extract deterministic posterior mean embeddings, after which geometric, probabilistic, and discriminative decision rules are evaluated on the same representation and under identical cross-validation partitions. This design isolates the decision layer and enables a direct comparison of how different structural assumptions interact with a common embedding geometry.

The study departs from conventional classifier benchmarking in two ways. First, it focuses on the stability of classifier ordering across heterogeneous acoustic and visual datasets rather than on identifying a universally superior algorithm. Second, it complements predictive evaluation with an explicit characterization of class-conditional geometry, including within-class isotropy, covariance anisotropy, prototype separation, radial compatibility, spectral decay, and effective dimensionality.

This framework allows downstream classification in generative embedding spaces to be interpreted as a problem of structural alignment. Radial rules are expected to be most suitable when class regions are approximately isotropic and sufficiently separated; Gaussian MAP can accommodate an anisotropic covariance structure, and discriminative models can represent more flexible decision boundaries. The empirical analysis therefore examines both whether classifier rankings change across datasets and whether the magnitude of performance differences is associated with measurable properties of posterior mean geometry.

The following section formalizes the latent representation, the decision paradigms, the radial compatibility formulation, and the geometry–performance analysis used to evaluate these hypotheses.

3. Methodology

3.1. Posterior mean representation

Let

$$\mathscr{S}_D = \left\{\left(\mathbf{x}_i^{(D)}, \mathbf{y}_i^{(D)}\right)\right\}_{i=1}^{N_D} \tag{5}$$

denote the labeled sample set for dataset D, where $\mathbf{x}_i^{(D)} \in \mathscr{X}_D$, $\mathbf{y}_i^{(D)} \in \{1, \ldots, C_D\}$, N_D is the number of samples, and C_D is the number of classes.

The encoder of a variational autoencoder, parameterized by ϕ, maps each input to the parameters of an approximate posterior distribution

$$q_\phi(\mathbf{z} \mid \mathbf{x}_i^{(D)}) = \mathscr{N}\left(\boldsymbol{\mu}_\phi(\mathbf{x}_i^{(D)}),\ \mathbf{diag}\left[\boldsymbol{\sigma}_\phi^2(\mathbf{x}_i^{(D)})\right]\right), \tag{6}$$

where $\mathbf{z} \in \mathbb{R}^{d_z}$ denotes the latent random variable, and d_z is the latent dimensionality. In the implementation, the encoder outputs the posterior mean and log-variance vectors

$$\left[\boldsymbol{\mu}_\phi(\mathbf{x}_i^{(D)}),\ \mathbf{log}\ \boldsymbol{\sigma}_\phi^2(\mathbf{x}_i^{(D)})\right]. \tag{7}$$

During VAE training, the posterior mean and variance parameterize stochastic sampling and contribute to the variational objective. After training, the encoder is frozen, and each sample is represented deterministically by its posterior mean

$$\mathbf{h}_i^{(D)} = \boldsymbol{\mu}_\phi(\mathbf{x}_i^{(D)}) \in \mathbb{R}^{d_z}. \tag{8}$$

All downstream classifiers and geometric descriptors are computed exclusively from $\mathbf{h}_i^{(D)}$. The posterior log-variance is not concatenated with the posterior mean and is not treated as a spatial coordinate, instead retaining its distinct interpretation as uncertainty about latent position. This convention defines a fixed Euclidean posterior mean embedding space and ensures that distances, centroids, covariance matrices, and geometry descriptors reflect class-conditional positional structure rather than uncertainty magnitude.

3.2. Decision paradigms

Let $\mathbf{h}_i^{(D)} \in \mathbb{R}^{d_z}$ denote the deterministic posterior mean embedding of sample i from dataset D, and let $\mathbf{y}_i^{(D)} \in \{1, \ldots, C_D\}$ denote its class label. Within each dataset, all classifiers operate on the same posterior mean embeddings and are evaluated using identical cross-validation partitions. Model parameters, class centroids, class radii, covariance matrices, and any preprocessing quantities are estimated exclusively from the training portion of each fold and subsequently applied to the corresponding held-out samples.

We evaluate four classifiers grouped into three structurally distinct decision paradigms: (i) a centroid-based radial rule, representing classes through isotropic Euclidean coverage around class prototypes; (ii) a Gaussian MAP classifier, incorporating class-specific covariance through a Mahalanobis geometry; and (iii) two discriminative classifiers—logistic regression and XGBoost—that learn affine or nonlinear separating functions directly from the embeddings.

These paradigms differ in the assumptions they impose on class-conditional organization. The radial rule assumes that Euclidean distance from a class centroid provides an adequate representation of class membership. Gaussian MAP relaxes isotropy by modeling direction-dependent dispersion and correlation. Logistic regression learns affine decision boundaries without explicitly modeling class-conditional densities, whereas XGBoost constructs locally adaptive partitions through ensembles of axis-aligned tree splits. The comparison is intended to evaluate contrasting forms of alignment between decision-rule assumptions and the fixed posterior mean geometry rather than a predetermined hierarchy of classifier flexibility or expected performance.

3.2.1. Centroid-based radial rule

For each class $c \in \{1, \ldots, C_D\}$, let

$$\mathscr{I}_c^{tr} = \left\{i : \mathbf{y}_i^{(D)} = c,\ i \in \mathscr{T}\right\} \tag{9}$$

denote the indices of the training fold samples belonging to class c, where $\mathscr{T}$ denotes the set of training sample indices in the current cross-validation fold, and define $N_c^{\mathrm{tr}} = \left|\mathscr{I}_c^{\mathrm{tr}}\right|$. The class centroid is estimated in the deterministic posterior mean embedding space as

$$\boldsymbol{\mu}_c = \frac{1}{N_c^{(\mathrm{tr})}} \sum_{i \in \mathscr{I}_c^{(\mathrm{tr})}} \mathbf{h}_i^{(D)}. \tag{10}$$

The Euclidean distance of each training embedding from its class centroid is

$$d_{i,c} = \|\mathbf{h}_i^{(D)} - \boldsymbol{\mu}_c\|_2,\ i \in \mathscr{I}_c^{\mathrm{tr}}. \tag{11}$$

A class-specific empirical radius is then defined by the γ-quantile of the within-class centroid distances

$$r_c = Q_\gamma\left(\left\{d_{i,c} : i \in \mathscr{I}_c^{\mathrm{tr}}\right\}\right), \tag{12}$$

where Q_γ denotes the empirical quantile operator. In the experiments, $\gamma = 0.95$.

For a held-out embedding $\mathbf{h}$, the normalized radial score assigned to class c is

$$s_c^{\mathrm{rad}}(\mathbf{h}) = \frac{\|\mathbf{h} - \boldsymbol{\mu}_c\|_2}{r_c + \varepsilon}. \tag{13}$$

where $\varepsilon = 10^{-12}$ is used for numerical stability and to prevent division by zero.

The predicted class is

$$\hat{y}_{rad}(\mathbf{h}) = \underset{c \in \{1, \ldots, C_D\}}{\mathbf{argmin}}\ s_c^{\mathrm{rad}}(\mathbf{h}). \tag{14}$$

Thus, the classifier performs closed-set assignment according to the smallest distance relative to the empirical radial extent of each class. The class radii do not act as rejection thresholds; instead, they normalize Euclidean distances so that classes with different overall dispersion can be compared on a common radial scale.

This rule represents each class through a single centroid and one scalar radius. It therefore assumes that class-conditional dispersion can be adequately summarized by an approximately isotropic region around the class prototype. The rule does not model covariance orientation, correlations among latent dimensions, or irregular boundary structure. Its effectiveness is consequently expected to depend on both within-class isotropy and sufficient separation between competing radial regions, which motivates the radial compatibility formulation introduced in Section 3.4.

3.2.2. Gaussian MAP classifier

The Gaussian MAP classifier models each class-conditional distribution in the deterministic posterior mean embedding space as

$$p(\mathbf{h}|y = c) = \mathscr{N}(\mathbf{h}; \boldsymbol{\mu}_c, \widetilde{\Sigma}_c), \qquad c \in \{1, \ldots, C_D\}, \tag{15}$$

where $\boldsymbol{\mu}_c$ is the training-fold class centroid defined in Eq. (10), and $\widetilde{\Sigma}_c$ is a regularized estimate of the class-specific covariance matrix.

For each class, covariance is estimated exclusively from the corresponding training-fold embeddings using the Ledoit–Wolf shrinkage estimator [10,11]. Let $\widehat{\Sigma}_c^{\mathrm{LW}}$ denote the resulting shrinkage covariance estimate. An additional scale-relative diagonal jitter is applied to ensure numerical stability during matrix inversion and log-determinant computation

$$\widetilde{\Sigma}_c = \widehat{\Sigma}_c^{\mathrm{LW}} + \eta \max\left(\frac{\mathbf{tr}\left(\widehat{\Sigma}_c^{\mathrm{LW}}\right)}{d_z}, \varepsilon_{\mathbf{mach}}\right) \mathbf{I}_{d_z}, \tag{16}$$

where $\eta = 10^{-6}$, $\varepsilon_{\mathbf{mach}}$ denotes the machine epsilon for double-precision floating-point arithmetic, and $\mathbf{I}_{d_z}$ is the $d_z \times d_z$ identity matrix. The added term is proportional to the average marginal variance of the class and therefore avoids introducing a fixed absolute perturbation across datasets with different embedding scales.

For a held-out embedding $\mathbf{h}$, the class-specific log-posterior score, up to an additive constant common to all classes, is

$$g_c^{\mathrm{MAP}}(\mathbf{h}) = -\frac{1}{2}(\mathbf{h} - \mu_c)^{\mathrm{T}} \widetilde{\Sigma}_c^{-1} (\mathbf{h} - \mu_c) - \frac{1}{2}\log\det(\widetilde{\Sigma}_c) + \log\pi_c, \tag{17}$$

where the class prior is estimated from the training fold as

$$\pi_c = \frac{N_c^{\mathrm{tr}}}{N^{\mathrm{tr}}}, \qquad N^{\mathrm{tr}} = \sum_{c=1}^{C_D} N_c^{\mathrm{tr}}. \tag{18}$$

The predicted class is

$$\widehat{y}_{\mathrm{MAP}}(\mathbf{h}) = \underset{c \in \{1,\ldots,C_D\}}{\mathbf{argmax}}\, g_c^{\mathrm{MAP}}(\mathbf{h}). \tag{19}$$

Unlike the centroid-based radial rule, Gaussian MAP accounts for class-specific scale, orientation, and correlation through the covariance matrix. The quadratic term in Eq. (17) induces a Mahalanobis geometry, whereas the log-determinant term accounts for the class-dependent volume of the covariance ellipsoid. Consequently, the resulting decision regions can be anisotropic and are not restricted to the isotropic geometry summarized by a single scalar radius. All class means, covariance estimates, priors, and numerical regularization quantities are computed exclusively from the training portion of each cross-validation fold.

3.2.3. Discriminative classifiers

We evaluate multinomial logistic regression and XGBoost decision trees as representative affine and nonlinear discriminative classifiers, respectively [12,13]. Unlike the radial and Gaussian MAP rules, these models do not specify an explicit class-conditional density or a prototype-based metric. Instead, they learn separating functions directly from the labeled posterior mean embeddings.

3.2.3.1. Logistic regression. Before applying logistic regression, a single isotropic scale is estimated exclusively from the training portion of each cross-validation fold. Let $\mathscr{T}$ denote the corresponding set of training sample indices. The fold-specific scale is defined as

$$a_{\mathscr{T}} = \left[\frac{1}{d_z}\sum_{j=1}^{d_z} \mathbf{Var}_{i \in \mathscr{T}}\left(h_{i,j}^{(D)}\right)\right]^{1/2} \tag{20}$$

where $h_{i,j}^{(D)}$ denotes the j-th coordinate of the embedding $\mathbf{h}_i^{(D)}$, and $\mathbf{Var}_{i \in \mathscr{T}}(\cdot)$ denotes the sample variance computed over the training sample indices of the current cross-validation fold. Training and held-out embeddings are then transformed using the same scalar

$$\overline{\mathbf{h}} = \frac{\mathbf{h}}{a_{\mathscr{T}}}. \tag{21}$$

This operation changes only the global numerical unit of the embedding space. It does not center the data, rotate the representation, or apply dimension-specific reweighting, thereby preserving relative Euclidean geometry.

For the rescaled embedding $\overline{\mathbf{h}}$, multinomial logistic regression models the class probabilities as

$$p_{\mathrm{LR}}(y = c|\overline{\mathbf{h}}) = \frac{\exp\left(\mathbf{w}_c^{\mathrm{T}}\,\overline{\mathbf{h}} + b_c\right)}{\sum_{c'=1}^{C_D}\exp\left(\mathbf{w}_{c'}^{\mathrm{T}}\,\overline{\mathbf{h}} + b_{c'}\right)}, \tag{22}$$

where $\mathbf{w}_c \in \mathbb{R}^{d_z}$ and $b_c \in \mathbb{R}$ are learned from the training fold. The predicted class is

$$\widehat{y}_{\mathrm{LR}}(\mathbf{h}) = \underset{c \in \{1,\ldots,C_D\}}{\mathbf{argmax}}\, p_{\mathrm{LR}}(y = c|\overline{\mathbf{h}}\,). \tag{23}$$

3.2.3.2. XGBoost. XGBoost operates directly on the unscaled deterministic posterior mean embeddings. It constructs a multiclass score by combining an ensemble of gradient-boosted decision trees

$$F_c(\mathbf{h}) = \sum_{m=1}^{M} f_{m,c}(\mathbf{h}), \qquad c \in \{1,\ldots,C_D\}, \tag{24}$$

where $f_{m,c}$ denotes the contribution of the boosted-tree ensemble to the score of class c. The predicted class is

$$\widehat{y}_{\mathrm{XGB}}(\mathbf{h}) = \underset{c \in \{1,\ldots,C_D\}}{\mathbf{argmax}}\, F_c(\mathbf{h}). \tag{25}$$

Logistic regression therefore induces affine separating boundaries in the globally rescaled embedding coordinates, whereas XGBoost constructs nonlinear, locally adaptive partitions through ensembles of axis-aligned tree splits. These classifiers represent two distinct discriminative alternatives rather than successive levels in a predetermined hierarchy. All fold-dependent scaling quantities and model parameters are estimated exclusively from the training portion of each fold, and the same held-out samples and cross-validation partitions are used for every decision rule.

The contrasting structural assumptions of the four classifiers are summarized schematically in Fig. 2. All panels display the same deterministic posterior mean representation, while only the downstream decision mechanism changes. The radial rule describes each class through a centroid and a scalar empirical radius; Gaussian MAP incorporates class-specific covariance; logistic regression learns affine separating boundaries after a training fold-only isotropic rescaling; and XGBoost constructs nonlinear, locally adaptive partitions directly from the unscaled embeddings. The figure is intended to illustrate differences in decision geometry rather than a hierarchy of classifier flexibility or expected performance.

3.3. Evaluation metrics

Classification performance is evaluated using overall accuracy, macro-averaged F1 score, and balanced accuracy. These metrics provide complementary views of global predictive correctness, class-wise precision–recall balance, and class-wise sensitivity.

For a cross-validation fold with held-out index set $\mathscr{V}$, overall accuracy is defined as

$$\mathrm{Acc} = \frac{1}{|\mathscr{V}|}\sum\nolimits_{i \in \mathscr{V}} \mathbb{I}\left[\widehat{y}_i = y_i^{(D)}\right], \tag{26}$$

where $|\mathscr{V}|$ is the number of held-out samples, and $\mathbb{I}[\cdot]$denotes the indicator function, which equals 1 when its argument is true, and 0 otherwise.

For each class c, precision and recall are defined as

$$P_c = \frac{\mathbf{TP}_c}{\mathbf{TP}_c + \mathbf{FP}_c}, \qquad R_c = \frac{\mathbf{TP}_c}{\mathbf{TP}_c + \mathbf{FN}_c}, \tag{27}$$

where $\mathbf{TP}_c$, $\mathbf{FP}_c$, and $\mathbf{FN}_c$ denote the numbers of true-positive, false-positive, and false-negative predictions for class c, respectively. The class-specific F1 score and macro-F1 are then

$$F_{1,c} = \frac{2P_cR_c}{P_c + R_c}, \qquad F_1^{\mathrm{macro}} = \frac{1}{C_D}\sum_{c=1}^{C_D} F_{1,c}. \tag{28}$$

Macro-F1 assigns equal weight to every class and is therefore the primary metric used for cross-dataset classifier comparison and ranking.

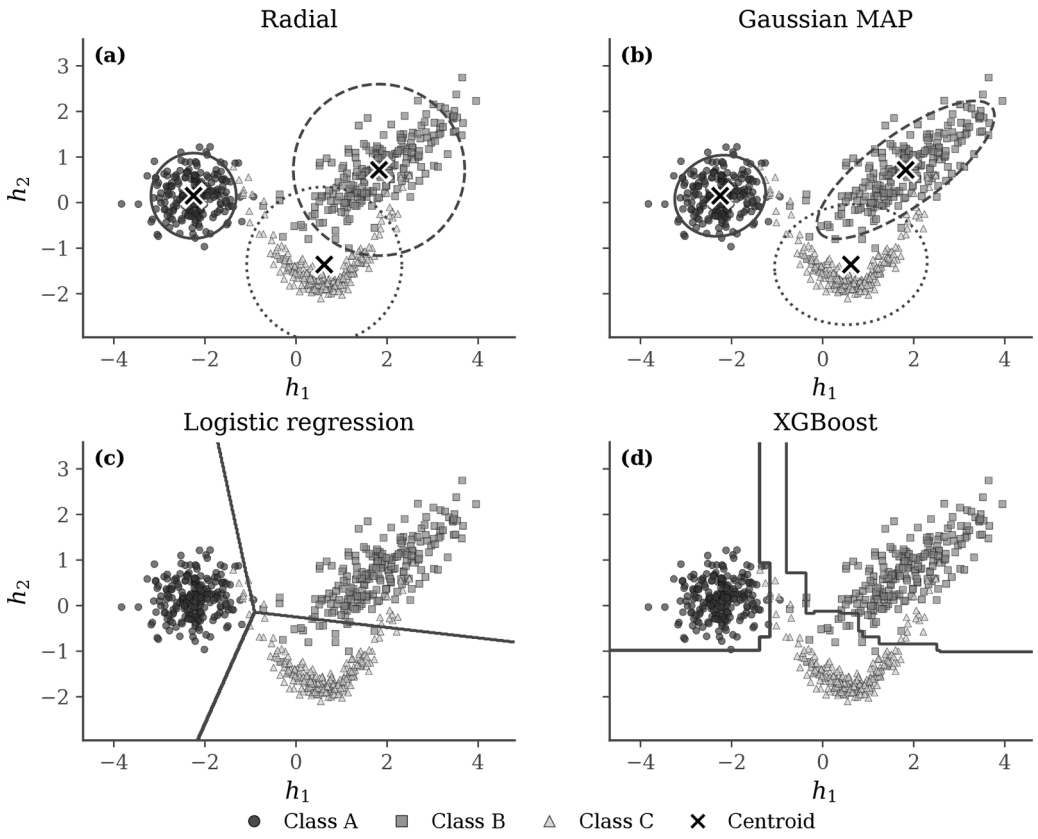

Fig. 2. Contrasting decision assumptions applied to a fixed deterministic posterior mean embedding space. The panels show the same schematic class organization, while only the downstream decision mechanism changes; posterior variance is not used as an embedding coordinate. (a) The centroid-based radial rule represents each class through Euclidean distance to a class prototype normalized by a class-specific empirical radius, producing isotropic radial regions. (b) Gaussian MAP incorporates class-specific covariance, yielding anisotropic Mahalanobis geometry and ellipsoidal class-conditional regions. (c) Logistic regression learns affine separating boundaries after a training fold-only isotropic rescaling of the embedding coordinates. (d) XGBoost operates on the unscaled embeddings and constructs nonlinear, locally adaptive partitions through an ensemble of axis-aligned tree splits. The panels provide a conceptual illustration of contrasting structural assumptions and do not imply a predetermined hierarchy of classifier flexibility or expected performance.

Balanced accuracy is defined as the unweighted mean of class-specific recall values

$$\mathbf{BAcc} = \frac{1}{C_D}\sum_{c=1}^{C_D} R_c. \tag{29}$$

Unlike overall accuracy, balanced accuracy is not dominated by classes with larger sample counts and therefore provides a complementary measure of class-wise sensitivity.

Let $K = 5$ denote the number of cross-validation folds, and let $m_{D,a}^{(k)}$ denote the value of metric m obtained by classifier a on fold k of dataset D. The corresponding cross-validated mean and standard deviation are

$$\overline{m}_{D,a} = \frac{1}{K}\sum_{k=1}^{K} m_{D,a}^{(k)}, \qquad s_{D,a}^{(m)} = \left[\frac{1}{K-1}\sum_{k=1}^{K}\left(m_{D,a}^{(k)} - \overline{m}_{D,a}\right)^2\right]^{1/2}. \tag{30}$$

Performance is reported as mean ± standard deviation across folds when fold-wise variability is shown. Cross-dataset comparisons and classifier rankings are based on $\overline{F}_{1,D,a}^{\,\mathrm{macro}}$, which is the mean macro-F1 score of the classifier a on dataset D.

In the evaluation of H1, the within-dataset rank of classifier a is defined as

$$R_{D,a} = 1 + \sum_{\substack{b\in A\\ b\neq a}} \mathbb{I}\left[\overline{F}_{1,D,b}^{\,\mathrm{macro}} > \overline{F}_{1,D,a}^{\,\mathrm{macro}}\right], \tag{31}$$

where $\mathscr{A} = \{\mathbf{Radial}, \mathbf{MAP}, \mathbf{LR}, \mathbf{XGB}\}$ is the set of evaluated classifiers. A rank of 1 therefore denotes the highest mean macro-F1 score within a dataset. Exact ties, if present, receive the same rank. Variation in $R_{D,a}$ across datasets is used to assess whether classifier ordering remains invariant or changes across domains.

All metrics are computed from held-out predictions generated under identical five-fold cross-validation partitions. Consequently, every classifier is evaluated on the same samples in each fold, and no training observation contributes to the corresponding held-out performance estimate.

3.4. Latent geometry characterization

To examine whether classifier behavior is associated with measurable properties of the fixed representation, we characterize the class-conditional geometry of the deterministic posterior mean embedding space. All descriptors are computed exclusively from the posterior mean embeddings and their class labels; posterior variance and classifier performance are not included.

The centroid-based radial rule evaluates Euclidean displacement relative to each class prototype. Thus, the relevant geometry is defined in class-centered coordinates rather than with respect to the global latent origin. For each class c, let

$$\mathscr{I}_c = \left\{i : y_i^{(D)} = c\right\} \tag{32}$$

denote the set of indices belonging to class c, and let $N_c = |\mathscr{I}_c|$. The class centroid computed from the complete posterior mean representation of dataset D is

$$\boldsymbol{\mu}_c = \frac{1}{\boldsymbol{N}_c} \sum_{i \in \mathscr{I}_c} \mathbf{h}_i^{(D)}. \tag{33}$$

For every sample $\boldsymbol{i} \in \mathscr{I}_c$, we define the class-centered residual vector as

$$\mathbf{u}_{i,c}^{(D)} = \mathbf{h}_i^{(D)} - \boldsymbol{\mu}_c, \tag{34}$$

with corresponding centroid distance

$$\boldsymbol{d}_{i,c} = \|\mathbf{u}_{i,c}^{(D)}\|_2. \tag{35}$$

This class-centered formulation is invariant to a common translation of the embedding space and directly reflects the geometry used by the radial decision rule. Radial compatibility is characterized through two complementary requirements: within-class isotropy, which assesses whether dispersion around each centroid can be adequately summarized by a single radial extent, and between-class radial separation, which assesses whether competing class regions remain sufficiently distinct relative to their empirical radii.

3.4.1. Within-class isotropy

Within-class isotropy describes the extent to which class dispersion is distributed uniformly across latent directions. For each class $\boldsymbol{c}$, we estimate a shrinkage-regularized covariance matrix from the centered residual vectors,

$$\widehat{\Sigma}_c^{\mathrm{LW}} = \mathbf{LW}\left(\left\{\mathbf{u}_{i,c}^{(D)} : \boldsymbol{i} \in \mathscr{I}_c\right\}\right), \tag{36}$$

where $\mathbf{LW}(\cdot)$ denotes the Ledoit–Wolf covariance estimator. Let $\lambda_{c,1} \geq \lambda_{c,2} \geq \cdots \geq \lambda_{c,d_z} \geq 0$ denote its ordered eigenvalues. We normalize the eigenvalue spectrum as

$$\boldsymbol{p}_{c,k} = \frac{\lambda_{c,k}}{\sum_{\ell=1}^{d_z} \lambda_{c,\ell}}, \qquad \boldsymbol{k} = 1, \ldots, \boldsymbol{d}_z. \tag{37}$$

The effective spectral entropy of class $\boldsymbol{c}$ is

$$\boldsymbol{H}_c = -\sum_{k=1}^{d_z} \boldsymbol{p}_{c,k} \log\left(\boldsymbol{p}_{c,k} + \varepsilon\right), \tag{38}$$

where ε is a small numerical constant. A normalized within-class isotropy score is then defined as

$$\boldsymbol{I}_c = \frac{\exp(\boldsymbol{H}_c)}{\boldsymbol{d}_z}. \tag{39}$$

The quantity $\exp(\boldsymbol{H}_c)$ can be interpreted as the effective number of variance-carrying latent directions for class $\boldsymbol{c}$. Dividing by $\boldsymbol{d}_z$ yields a normalized measure of how evenly class variance is distributed across the latent dimensions. Values close to 1 indicate comparatively uniform dispersion across directions, corresponding to an approximately isotropic class region. Smaller values indicate that variance is concentrated along a reduced number of dominant directions, revealing an anisotropic class structure that cannot be adequately summarized by a single scalar radius.

3.4.2. Between-class radial separation

Within-class isotropy alone is insufficient for radial classification; even approximately spherical class regions may be difficult to distinguish when their empirical radial extents overlap. We therefore complement the isotropy score with a measure of separation between competing radial regions.

For each class $\boldsymbol{c}$, the empirical radial extent is computed from the complete class-centered representation as

$$\boldsymbol{r}_c = \boldsymbol{Q}_\gamma\left(\left\{\boldsymbol{d}_{i,c} : \boldsymbol{i} \in \mathscr{I}_c\right\}\right), \qquad \gamma = 0.95, \tag{40}$$

where $\boldsymbol{Q}_\gamma$ denotes the empirical quantile operator, and $\boldsymbol{d}_{i,c}$ is defined in Eq. (35). The same quantile level used by the radial classifier is retained here so that the geometric descriptor reflects the operational class extent assumed by the decision rule.

The centroid separation for two distinct classes $\boldsymbol{c}$ and $\boldsymbol{j}$ is

$$\boldsymbol{\delta}_{c,j} = \|\boldsymbol{\mu}_c - \boldsymbol{\mu}_j\|_2. \tag{41}$$

We normalize this distance by the sum of the corresponding empirical radii

$$\boldsymbol{\rho}_{c,j} = \frac{\boldsymbol{\delta}_{c,j}}{\boldsymbol{r}_c + \boldsymbol{r}_j + \varepsilon}, \qquad \boldsymbol{c} \neq \boldsymbol{j}, \tag{42}$$

where ε is a small positive constant introduced for numerical stability. The ratio $\boldsymbol{\rho}_{c,j}$ compares the distance between class prototypes with the combined empirical radial extent of the two classes. Values below 1 indicate overlap between the corresponding radial regions, a value of 1 corresponds to approximate tangency, and values above 1 indicate radial separation.

Radial classification is most vulnerable to the least radially separated competing class. Therefore, the class-level separation score is defined as

$$\boldsymbol{S}_c = \min_{j \neq c}\left\{\min\left(1, \boldsymbol{\rho}_{c,j}\right)\right\}. \tag{43}$$

Thus, $0 \leq \boldsymbol{S}_c \leq 1$. Values close to 1 indicate that the empirical radial region of class $\boldsymbol{c}$ does not overlap with that of its least radially separated competitor, whereas smaller values indicate increasing radial overlap. The truncation at 1 reflects the compatibility requirement of sufficient separation: Once the empirical radial region of class $\boldsymbol{c}$ ceases to overlap with that of its least radially separated competitor, additional centroid distance does not further increase this bounded separation score.

3.4.3. Radial compatibility

Radial compatibility requires two conditions to hold simultaneously: The dispersion of a class should be approximately isotropic around its centroid, and its empirical radial region should remain sufficiently separated from competing classes. We therefore combine the within-class isotropy score $\boldsymbol{I}_c$ and the between-class radial separation score $\boldsymbol{S}_c$ through their geometric mean

$$\mathbf{RC}_c = \sqrt{\boldsymbol{I}_c \boldsymbol{S}_c}, \tag{44}$$

where $0 \leq \mathbf{RC}_c \leq 1$. The geometric mean assigns equal importance to both components and limits compensation between them. Consequently, a high value of one component cannot fully offset a low value of the other, and $\mathbf{RC}_c = 0$ whenever either isotropy or radial separation is zero.

Values of $\mathbf{RC}_c$ close to 1 indicate that class $\boldsymbol{c}$ forms an approximately isotropic region around its centroid and remains radially separated from its least radially separated competitor. Smaller values indicate an anisotropic within-class structure, an overlap between competing empirical radial regions, or both. The score therefore measures compatibility with the structural assumptions of the centroid-based radial rule rather than general class separability or classifier performance.

A dataset-level radial compatibility score is obtained by averaging equally across its $\boldsymbol{C}_D$ classes

$$\mathbf{RC}_D = \frac{1}{\boldsymbol{C}_D} \sum_{c=1}^{\boldsymbol{C}_D} \mathbf{RC}_c. \tag{45}$$

This aggregation is consistent with the class-balanced perspective of macro-F1 and balanced accuracy. It ensures that all classes contribute equally to the dataset-level descriptor, irrespective of their class frequencies. Both $\mathbf{RC}_c$ and $\mathbf{RC}_D$ are computed exclusively from posterior mean embeddings and class labels and do not incorporate classifier predictions or performance values.

3.4.4. Spectral geometry descriptors

In addition to radial compatibility, we characterize the covariance structure of each class by using complementary spectral descriptors. These quantities provide information about anisotropy, variance concentration, and effective dimensionality that is not fully summarized by $\mathbf{RC}_c$.

For class c, let $\lambda_{c,1} \geq \lambda_{c,2} \geq \cdots \geq \lambda_{c,d_z} \geq 0$ denote the ordered eigenvalues of the Ledoit–Wolf covariance estimate $\widehat{\Sigma}_c^{\mathrm{LW}}$ introduced in Eq. (36). Here, c indexes the class, and k indexes the ordered principal covariance directions.

The spectral condition number is defined as

$$\kappa_c = \frac{\lambda_{c,1}}{\lambda_{c,d_z} + \varepsilon}, \tag{46}$$

where ε is a small positive constant introduced for numerical stability. Values of κ_c close to 1 indicate comparatively uniform variance across latent directions, whereas larger values indicate stronger covariance anisotropy and greater disparity between dominant and weak-variance directions.

To characterize the rate at which variance decreases across ordered latent directions, we fit a linear model to the leading portion of the eigenvalue spectrum in log–log coordinates

$$\log(\lambda_{c,k} + \varepsilon) = \alpha_c + \beta_c \log k + e_{c,k}, \qquad k = 1, \ldots, m_c, \tag{47}$$

where $m_c = \min(m_{\max}, d_z)$, α_c is an intercept, $\beta_c$is the class-specific spectral slope, and $e_{c,k}$ denotes the residual term. The truncation level was fixed a priori at $m_{\max} = 20$ for all classes and datasets. This restriction focuses the estimate on the dominant portion of the covariance spectrum and reduces the influence of very small trailing eigenvalues, which are more sensitive to numerical regularization and finite-sample variability. More negative values of β_c indicate faster eigenvalue decay and stronger concentration of variance along a limited number of dominant directions, whereas values closer to zero indicate a flatter spectrum and a more even distribution of variance.

The effective latent dimensionality at cumulative variance level τ is defined as

$$k_{\mathrm{eff},\tau}^{(c)} = \min\left\{ q \in \{1, \ldots, d_z\} : \frac{\sum_{k=1}^{q} \lambda_{c,k}}{\sum_{k=1}^{d_z} \lambda_{c,k}} \geq \tau \right\}. \tag{48}$$

In the main analysis, $\tau = 0.90$, yielding $k_{\mathrm{eff},90}^{(c)}$, the smallest number of leading principal covariance directions required to explain at least 90% of the within-class variance. Lower values indicate stronger spectral concentration, whereas higher values indicate that variance is distributed across a larger number of principal directions.

Dataset-level descriptors are obtained by aggregating across classes without weighting by class frequency

$$\kappa_D = \mathbf{median}_c(\kappa_c), \qquad \beta_D = \frac{1}{C_D} \sum_{c=1}^{C_D} \beta_c, \qquad k_{\mathrm{eff},90}^{(D)} = \frac{1}{C_D} \sum_{c=1}^{C_D} k_{\mathrm{eff},90}^{(c)}. \tag{49}$$

The median is used for the condition number because this descriptor can be strongly affected by classes with extremely small trailing eigenvalues. Spectral slope and effective dimensionality are averaged across classes to provide class-balanced dataset-level summaries. Together with $\mathbf{RC}_D$, these descriptors form an intrinsic characterization of the posterior mean geometry and are computed without using classifier predictions or performance values.

3.5. Geometry–performance association

Classifier rankings describe the relative ordering of decision rules within each dataset, but they do not quantify the magnitude of the corresponding performance differences. To evaluate H2, we therefore relate intrinsic dataset-level geometry descriptors to pairwise contrasts in mean macro-F1 performance.

For classifiers $a, b \in \mathscr{A}$, the performance contrast on dataset D is defined as

$$\Delta_{a-b}^{(D)} = \overline{F}_{1,D,a}^{\,\mathrm{macro}} - \overline{F}_{1,D,b}^{\,\mathrm{macro}}, \tag{50}$$

where $\overline{F}_{1,D,a}^{\,\mathrm{macro}}$ denotes the mean cross-validated macro-F1 score of classifier a on dataset D. Positive values indicate that classifier a has an advantage over classifier b, whereas negative values indicate that classifier b performs better.

Radial compatibility is defined with respect to the structural assumptions of the centroid-based radial rule. Hence, the primary contrasts compare each non-radial classifier with the radial baseline

$$\Delta_{\mathrm{MAP-Radial}}^{(D)}, \qquad \Delta_{\mathrm{LR-Radial}}^{(D)}, \qquad \Delta_{\mathrm{XGB-Radial}}^{(D)}. \tag{51}$$

These contrasts quantify the extent to which covariance-aware or discriminative decision rules improve upon the radial classifier while the posterior mean representation remains fixed.

For each dataset, the intrinsic geometry descriptor vector is

$$\mathbf{g}_D = \left[\mathbf{RC}_D,\ \kappa_D,\ \beta_D,\ k_{\mathrm{eff},90}^{(D)} \right]^{\mathrm{T}}. \tag{52}$$

The vector $\mathbf{g}_D$ contains only quantities computed from posterior mean embeddings and class labels. It does not contain classifier predictions, performance values, or quantities derived from them.

For each performance contrast $\Delta_{a-b}^{(D)}$ and each geometric descriptor $g_{D,q}$, the cross-dataset monotonic association is quantified using Spearman's rank correlation coefficient

$$\rho_q^{(a-b)} = \rho_S\left(\left\{ g_{D,q} \right\}_{D \in \mathscr{D}}, \left\{ \Delta_{a-b}^{(D)} \right\}_{D \in \mathscr{D}} \right), \tag{53}$$

where q indexes the components of $\mathbf{g}_D$. Spearman correlation is used because it evaluates monotonic association without requiring linearity or Gaussian distributions and, by operating on ranks, is less sensitive to extreme observations than Pearson correlation.

The expected direction of association depends on the descriptor. Higher radial compatibility, $\mathbf{RC}_D$, is expected to correspond to a smaller advantage of non-radial rules over the radial classifier. By contrast, larger condition numbers may be associated with larger non-radial advantages. Because faster spectral decay is represented by more negative values of β_D and stronger variance concentration is represented by lower values of $k_{\mathrm{eff},90}^{(D)}$, negative associations with these two descriptors would be consistent with increasing advantages for covariance-aware or discriminative classifiers. These directional expectations are treated as structural hypotheses rather than deterministic relationships.

Because the number of benchmark datasets is limited, the geometry–performance analysis emphasizes correlation magnitude, direction, consistency across complementary descriptors, and agreement with the observed classifier rankings. Statistical significance is treated cautiously and is not used as the sole criterion for supporting or rejecting H2. Scatter plots and fitted linear trends are used for descriptive visualization, whereas Spearman's rank correlation is still the primary association measure.

All geometry descriptors are computed independently of classifier performance. Consequently, the analysis evaluates whether intrinsic properties of the posterior mean representation covary with decision-rule advantage without using performance information to construct the explanatory variables.

4. Experimental setup

The experimental setup was designed to separate representation learning from downstream decision modeling. For each dataset, a variational autoencoder was trained and subsequently frozen, and its

deterministic posterior mean embeddings were used as the common representation for all downstream analyses. Within each dataset, the four classifiers were evaluated using identical cross-validation partitions, while all fold-dependent quantities were estimated exclusively from the corresponding training portion.

The benchmark suite comprises eight datasets spanning acoustic and visual domains. This heterogeneity was introduced to obtain a broad range of class-conditional embedding geometries and classifier behaviors rather than to impose a predefined ordering of dataset complexity. Geometric descriptors were measured directly from the corrected posterior mean embeddings and were not assigned from domain labels or qualitative expectations.

The experimental analysis addresses two complementary questions. First, within each dataset, classifier performance and ranking are compared under a fixed representation to evaluate whether decision-rule ordering remains invariant across domains. Second, each dataset is treated as one observation in the cross-dataset geometry–performance analysis, where intrinsic geometric descriptors are related to pairwise classifier performance contrasts.

4.1. Cross-dataset evaluation design

For each dataset $\boldsymbol{D} \in \mathscr{D}$, the experimental pipeline produces three outputs: $\{\mathbf{H}_{\boldsymbol{D}},\ \mathbf{m}_{\boldsymbol{D}},\ \mathbf{g}_{\boldsymbol{D}}\}$, where

$$\mathbf{H}_{\boldsymbol{D}} = \left[\mathbf{h}_1^{(\boldsymbol{D})}, \ldots, \mathbf{h}_{N_{\boldsymbol{D}}}^{(\boldsymbol{D})}\right]^{\mathrm{T}} \in \mathbb{R}^{N_{\boldsymbol{D}} \times d_z} \tag{54}$$

is the deterministic posterior mean embedding matrix,

$$\mathbf{m}_{\boldsymbol{D}} = \left[\ \overline{F}_{1,\boldsymbol{D},\mathrm{Radial}}^{\ \mathrm{macro}},\ \overline{F}_{1,\boldsymbol{D},\mathrm{MAP}}^{\ \mathrm{macro}},\ \overline{F}_{1,\boldsymbol{D},\mathrm{LR}}^{\ \mathrm{macro}},\ \overline{F}_{1,\boldsymbol{D},\mathrm{XGB}}^{\ \mathrm{macro}}\ \right]^{\mathrm{T}} \tag{55}$$

contains the mean cross-validated macro-F1 scores, and $\mathbf{g}_{\boldsymbol{D}}$ is the intrinsic geometry descriptor vector defined in Eq. (52).

The within-dataset component of the analysis uses $\mathbf{m}_{\boldsymbol{D}}$ and the corresponding classifier ranks to evaluate H1. Because all decision rules operate on the same $\mathbf{H}_{\boldsymbol{D}}$ and identical held-out samples, within-dataset performance differences can be interpreted primarily in terms of the downstream decision mechanism and its interaction with the fixed representation.

For each classifier contrast $\boldsymbol{a} - \boldsymbol{b}$, the cross-dataset component uses the eight pairs

$$\left\{\left(\mathbf{g}_{\boldsymbol{D}}, \boldsymbol{\Delta}_{\boldsymbol{a}-\boldsymbol{b}}^{(\boldsymbol{D})}\right)\right\}_{\boldsymbol{D}\in\mathscr{D}} \tag{56}$$

to evaluate H2. Each dataset therefore contributes one geometry vector and one value for each classifier contrast. Domain labels are used only for descriptive grouping and visualization; the association analysis is based on the measured geometric descriptors rather than on an assumed acoustic–visual distinction.

This design does not require the benchmark datasets to follow a predetermined progression from simple to complex geometry. Instead, variability in isotropy, radial separation, covariance anisotropy, spectral decay, and effective dimensionality is quantified empirically from the posterior mean embeddings. The benchmark suite is therefore used to test whether changes in decision-rule advantage covary with the observed geometric structure across heterogeneous domains.

4.2. Datasets

The benchmark suite comprises eight multiclass datasets spanning acoustic and visual domains. Dataset inclusion was guided by three practical considerations: availability of labeled samples, compatibility with the corresponding generative embedding pipeline, and sufficient class diversity to support class-conditional geometric characterization. The datasets were not assigned a priori to predefined geometric regimes; isotropy, radial separation, covariance anisotropy, spectral decay, and effective dimensionality were measured directly from their posterior mean embeddings.

Table 1 summarizes the domain, number of classes, and number of samples used in the downstream classification and geometric analyses. The acoustic group includes an ecoacoustic amphibian-vocalization dataset, UrbanSound8K, and ESC-50. The visual group includes MNIST, Fashion-MNIST, CIFAR-10, SVHN, and CIFAR-100.

4.2.1. Acoustic datasets

4.2.1.1. Ecoacoustic amphibian vocalizations. The amphibian dataset contains 400 audio segments distributed equally across four species: *Batrachyla leptopus, Batrachyla taeniata, Calyptocephalella gayi*, and *Pleurodema thaul*. The recordings were obtained through passive acoustic monitoring in natural–urban wetland environments. One hundred labeled segments per species were retained for the analysis.

4.2.1.2. UrbanSound8K. UrbanSound8K is a 10-class benchmark of urban environmental sounds. The analysis includes 7,738 labeled audio clips represented by corrected 128-dimensional deterministic posterior mean embeddings. The original class definitions and labels provided by the benchmark were retained throughout the downstream classification and geometric analyses.

4.2.1.3. ESC-50. ESC-50 contains 2,000 environmental audio recordings distributed uniformly across 50 classes. All recordings represented in the corrected posterior mean embedding table were included in the downstream classification and geometric analyses.

4.2.2. Visual datasets

4.2.2.1. MNIST. MNIST contains 70,000 grayscale images of handwritten digits from 10 classes. The full set of corrected posterior mean embeddings was used.

4.2.2.2. Fashion-MNIST. Fashion-MNIST contains 70,000 grayscale images representing 10 categories of clothing and footwear. It shares the image dimensions and class count of MNIST while describing a distinct visual recognition task.

4.2.2.3. CIFAR-10. CIFAR-10 contains 60,000 color images distributed across 10 object classes. All corrected posterior mean embeddings were included.

4.2.2.4. SVHN. The Street View House Numbers (SVHN) dataset contains cropped color images of digits extracted from natural street scenes. The analysis uses 99,289 labeled samples across 10 classes.

4.2.2.5. CIFAR-100. CIFAR-100 contains 60,000 color images distributed across 100 fine-grained object classes. Its larger class count

Table 1
Dataset composition used in the downstream classification and geometric analyses.

Dataset	Domain	Samples	Classes	Input modality
Ecoacoustic amphibian vocalizations	Acoustic	400	4	Audio segments
UrbanSound8K	Acoustic	7,738	10	Environmental audio
ESC-50	Acoustic	2,000	50	Environmental audio
MNIST	Visual	70,000	10	Grayscale images
Fashion-MNIST	Visual	70,000	10	Grayscale images
CIFAR-10	Visual	60,000	10	Color images
SVHN	Visual	99,289	10	Color digit images
CIFAR-100	Visual	60,000	100	Color images

provides a complementary multiclass setting for evaluating decision-rule ranking and class-conditional geometry.

All downstream analyses used the deterministic 128-dimensional posterior mean embeddings associated with these samples. The public benchmark datasets retained their original labels. The amphibian dataset was collected by the authors and can be made available upon reasonable request, subject to the corresponding data management and sharing conditions.

4.3. Generative embedding protocol

For each dataset, the encoder of a trained variational autoencoder was frozen and used to extract the deterministic 128-dimensional posterior mean embeddings defined in Eq. (8). The posterior log-variance remained part of the probabilistic VAE formulation during representation learning, but it was not concatenated with the posterior mean and was not used as an embedding coordinate in any downstream classifier or geometry descriptor.

After encoder training, posterior mean embeddings were extracted once for all labeled samples included in the downstream analysis and then held fixed. Radial classification, Gaussian MAP, logistic regression, and XGBoost therefore operated on the same deterministic representation within each dataset, and all intrinsic geometry descriptors were computed from that representation.

No encoder fine-tuning, domain adaptation, or classifier-dependent modification of the latent representation was performed during downstream evaluation. Any fold-dependent preprocessing or classifier-specific transformation was estimated exclusively from the training portion of the corresponding cross-validation fold and then applied to the held-out samples. In particular, logistic regression used the training-fold-only isotropic rescaling defined in Eqs. (20)–(21), whereas the radial rule, Gaussian MAP, XGBoost, and the geometry descriptors operated on the unscaled posterior mean coordinates.

This protocol isolates the downstream decision stage within each dataset. Cross-dataset differences are not interpreted as arising from identical representation-learning dynamics but are instead related to the measured geometry of each resulting posterior mean space and the corresponding classifier behavior.

4.4. Controlled evaluation protocol

The downstream evaluation protocol ensured that all four decision rules were compared under matched within-dataset conditions. The posterior mean embedding table, class labels, cross-validation partitions, and held-out samples were identical across classifiers. Thus, each decision rule was evaluated on the same observations under the same fold structure.

For each fold, all preprocessing quantities and model parameters were estimated exclusively from the training subset and then applied to the held-out subset. These quantities included class centroids and empirical radii for the radial rule, class priors and Ledoit–Wolf covariance estimates for Gaussian MAP, the isotropic scale used by logistic regression, and the fitted parameters of the discriminative models. No held-out information was used during model fitting, preprocessing estimation, or fold-dependent computation.

The radial rule, Gaussian MAP, logistic regression, and XGBoost were evaluated using fixed configurations. Logistic regression was fitted with L_2 regularization, inverse regularization strength $C = 1$, and the Newton conjugate-gradient solver. The primary convergence tolerance was 10^{-8}, with a maximum of 5,000 iterations. If, and only if, the primary fit of a fold generated a convergence warning, that fold was deterministically refitted using a tolerance of 10^{-6}. The accepted fit provided the final held-out predictions for that fold.

XGBoost was configured with 400 boosting rounds, a maximum tree depth of 6, a learning rate of 0.05, row subsampling of 0.9, feature subsampling of 0.9, and an L_2 regularization parameter of 1. The multiclass probabilistic objective and multiclass log-loss evaluation criterion were used throughout.

The deterministic convergence fallback used for logistic regression changed only the numerical tolerance of the optimizer. It did not change the model family, regularization strength, training data, held-out samples, or evaluation partition. Randomness was controlled through fixed seeds wherever stochastic fitting or partitioning required initialization.

Geometry descriptors were computed from the complete deterministic posterior mean representation of each dataset and its class labels independently of classifier predictions and performance values. They were not estimated fold by fold and were not used to tune, select, or weight any classifier.

4.5. Cross-dataset geometry–performance analysis

The cross-dataset analysis treated the eight benchmark datasets as dataset-level observations. For each dataset D, the intrinsic geometry descriptor vector $\boldsymbol{g}_D$ defined in Eq. (52) was paired with the three performance contrasts relative to the radial classifier: $\Delta_{MAP-Radial}$, $\Delta_{LR-Radial}$, and $\Delta_{XGB-Radial}$.

For each classifier contrast, Spearman's rank correlation coefficient was computed separately against $\mathbf{RC}_D$, κ_D, β_D, and $k_{eff,90}^{(D)}$. These correlations describe whether monotonic changes in intrinsic posterior mean geometry were accompanied by changes in the advantage of each non-radial decision rule over the radial baseline.

Because only eight datasets were available, the analysis emphasized correlation magnitude, direction, and consistency across complementary descriptors. Statistical significance was interpreted cautiously and was not treated as the sole criterion for supporting or rejecting H2. Scatter plots were used for descriptive visualization, with the dataset identity shown explicitly and the acoustic/visual domain indicated only as a descriptive grouping. Fitted linear trends were included as visual guides, whereas Spearman's rank correlation remained the primary association measure.

Classifier ranking results from H1 were interpreted alongside the geometry–performance correlations. This joint interpretation allowed rank changes to be compared with pairwise performance contrasts and with the measured structure of the corresponding posterior mean spaces. All geometry descriptors were computed independently of classifier performance, and no classifier result was used to construct, normalize, select, or weight the explanatory variables.

5. Empirical results across datasets

This section evaluates decision-rule behavior in deterministic posterior mean embedding spaces. All classification results were obtained using identical five-fold cross-validation partitions for the four decision rules within each dataset. The radial rule, Gaussian MAP, and XGBoost operated directly on the unscaled 128-dimensional posterior mean embeddings, whereas logistic regression used the training fold-only isotropic rescaling described in Section 3.2.3.

The analysis proceeds in five stages. First, the amphibian vocalization dataset illustrates how distinct decision assumptions interact with a shared representation. Second, predictive performance is compared across the eight datasets. Third, within-dataset classifier ranks are examined to evaluate H1. Fourth, performance contrasts relative to the radial classifier are related to intrinsic geometric descriptors to evaluate H2. The analysis then integrates classifier behavior and embedding geometry into a cross-dataset structural summary.

All geometric descriptors were computed exclusively from posterior mean embeddings and class labels; posterior log-variance, classifier predictions, and performance values were not used. Thus, the geometry–performance analysis relates independently computed representation properties to held-out classification performance.

5.1. Within-dataset illustration

We first examine the ecoacoustic amphibian vocalization dataset as a within-dataset illustration of decision-rule behavior on a common deterministic representation. The dataset comprises 400 audio segments equally distributed across four species: *B. leptopus, B. taeniata, C. gayi*, and *P. thaul*. All classifiers were evaluated using the same 128-dimensional posterior mean embeddings and five-fold cross-validation partitions.

Fig. 3 shows a two-dimensional t-SNE projection of the posterior mean embedding space. The projection reveals species-dependent organization: *C. gayi* and *P. thaul* form comparatively compact and distinguishable groups, whereas *B. leptopus* and *B. taeniata* overlap more strongly. This visualization is qualitative only; all classifiers and geometric descriptors were computed in the original 128-dimensional posterior mean space.

Table 2 reports five-fold cross-validated accuracy, macro-F1, and balanced accuracy. Gaussian MAP achieved the highest macro-F1 score (0.884), closely followed by logistic regression (0.880) and XGBoost (0.871), whereas the radial classifier obtained a substantially lower score (0.665). Thus, the macro-F1 advantage of Gaussian MAP over the radial rule was approximately 0.219.

Because the species were equally represented in each held-out fold, balanced accuracy coincided with overall accuracy for all classifiers. The observed differences therefore reflect decision-rule behavior rather than class-frequency imbalance.

The out-of-fold confusion matrices in Fig. 4 clarify the class-specific origin of these differences. The radial rule showed pronounced asymmetric confusion between the two *Batrachyla* species, classifying 69% of *B. leptopus* segments as *B. taeniata*. This error decreased to 16%, 18%, and 14% under Gaussian MAP, logistic regression, and XGBoost, respectively. The reverse confusion was more persistent, with approximately 24%–25% of *B. taeniata* segments assigned to *B. leptopus* by the three non-radial classifiers. For *C. gayi* and *P. thaul*, the non-radial classifiers achieved nearly perfect recall, whereas the radial rule also confused 13% of *C. gayi* with *P. thaul* and 18% of *P. thaul* with *B. taeniata*. These patterns indicate that the radial deficit reflects both the strong interaction between *Batrachyla* classes and additional class-specific errors consistent with isotropic centroid-based limitations.

Overall, the amphibian dataset illustrates that the fixed posterior mean representation contains discriminative information not fully accessible to isotropic radial normalization. Gaussian MAP benefits from class-specific covariance structure, while logistic regression and XGBoost also exploit information available through affine and nonlinear decision boundaries. The small differences among these three non-radial classifiers indicate that no single alternative mechanism is uniquely favored in this dataset; rather, several structurally distinct rules can exploit class organization that the radial rule does not capture.

Table 2
Five-fold cross-validated classification performance on the ecoacoustic amphibian vocalization dataset using the corrected 128-dimensional deterministic posterior mean embeddings. The radial rule, Gaussian MAP, and XGBoost operate on the unscaled posterior mean representation, whereas logistic regression uses a training fold-only isotropic rescaling. Overall accuracy, macro-F1, and balanced accuracy are reported as means across folds. Balanced accuracy assigns equal weight to the recall of each species.

Model	Accuracy	Macro-F1	Balanced accuracy
Radial	0.680	0.665	0.680
Gaussian MAP	0.885	0.884	0.885
Logistic regression	0.880	0.880	0.880
XGBoost	0.873	0.871	0.873

5.2. Cross-dataset classification performance

Table 3 reports the mean five-fold cross-validated macro-F1 scores obtained by the four classifiers across the eight benchmark datasets. Absolute performance varied substantially across datasets, reflecting differences in class count, input domain, and representation difficulty. The relevant comparison is therefore within each dataset, where all decision rules operated on the same posterior mean embeddings and identical held-out samples.

No classifier achieved the highest macro-F1 score across all datasets. Gaussian MAP performed best on the ecoacoustic amphibian dataset, logistic regression achieved the highest scores on UrbanSound8K and ESC-50, and XGBoost ranked first on all five visual datasets. The radial classifier did not rank first on any benchmark.

The magnitude of the best-classifier advantage over the radial rule also varied substantially: approximately 0.219 for the amphibian dataset, 0.337 for UrbanSound8K, 0.354 for ESC-50, 0.124 for MNIST, 0.192 for Fashion-MNIST, 0.207 for CIFAR-10, 0.338 for SVHN, and 0.150 for CIFAR-100. Thus, even when the radial rule was not the top-performing classifier, the extent of its performance deficit depended strongly on the dataset.

Gaussian MAP was particularly competitive on the amphibian dataset and MNIST, but its relative performance declined on several more challenging visual datasets. On CIFAR-100, its macro-F1 score was lower than that of the radial rule, although both were substantially

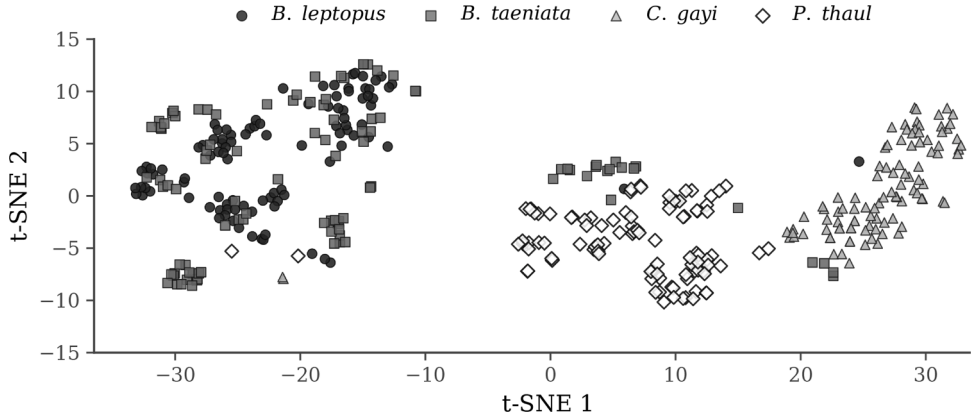

Fig. 3. t-SNE visualization of posterior mean embeddings for amphibian vocalizations. Each point represents an audio segment projected from the learned 128-dimensional deterministic posterior mean representation into two dimensions. Species are distinguished by marker shape and grayscale level (*B. leptopus, B. taeniata, C. gayi*, and *P. thaul*). The visualization provides a qualitative illustration of species-dependent organization in the embedding space, highlighting cluster separation and overlap patterns; distances in the two-dimensional projection should not be interpreted quantitatively as distances in the original 128-dimensional space.

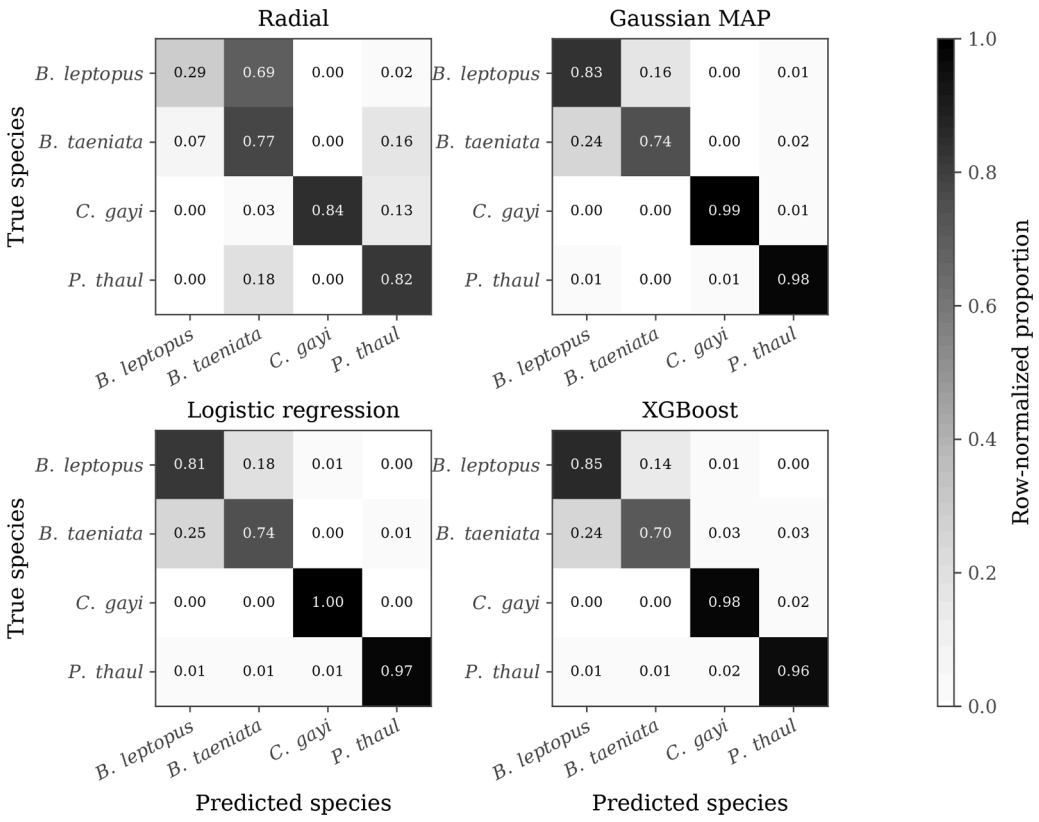

Fig. 4. Out-of-fold confusion matrices for the four decision rules on the ecoacoustic amphibian vocalization dataset. The matrices aggregate held-out predictions from the five cross-validation folds and are normalized by true class, so each row represents the distribution of predictions for one species. All classifiers were evaluated using the same deterministic 128-dimensional posterior mean embeddings and identical held-out samples. Species order, row normalization, and grayscale limits are held constant across panels to permit direct comparison of class-specific error patterns.

Table 3
Mean five-fold cross-validated macro-F1 scores for the four decision rules across the eight benchmark datasets using the corrected 128-dimensional deterministic posterior mean embeddings. Within each dataset, all classifiers were evaluated on identical cross-validation partitions and held-out samples. The highest macro-F1 score in each row is shown in bold. Absolute values should be interpreted within datasets because task difficulty, class count, and input domain differ across benchmarks.

Dataset	Radial	Gaussian MAP	Logistic regression	XGBoost
Ecoacoustic amphibian vocalizations	0.665	0.884	0.880	0.871
UrbanSound8K	0.247	0.529	0.585	0.563
ESC-50	0.099	0.397	0.453	0.365
MNIST	0.854	0.971	0.975	0.978
Fashion-MNIST	0.676	0.777	0.856	0.868
CIFAR-10	0.254	0.296	0.416	0.461
SVHN	0.056	0.229	0.293	0.394
CIFAR-100	0.070	0.045	0.179	0.220

outperformed by logistic regression and XGBoost. This finding indicates that class-specific covariance modeling is not universally sufficient when the class structure is highly complex, irregular, or difficult to estimate reliably. Conversely, logistic regression and XGBoost generally improved upon the radial classifier, with XGBoost achieving the strongest visual dataset performance and logistic regression leading on the two public environmental sound benchmarks.

Taken together, these results show that classifier performance depends on the interaction between the fixed posterior mean geometry and the assumptions of the downstream decision rule. The absence of a universally superior classifier provides the first empirical indication in support of H1, which is examined through within-dataset classifier rankings in Section 5.3.

5.3. *Classifier ranking variability*

To evaluate H1, the four decision rules were ranked within each dataset according to their mean cross-validated macro-F1 scores, with rank 1 assigned to the best-performing classifier. Table 4 reports the resulting within-dataset ranks and their averages across the eight benchmarks.

Classifier ordering varied across datasets. Gaussian MAP ranked first on the amphibian dataset, logistic regression ranked first on UrbanSound8K and ESC-50, and XGBoost ranked first on the five visual datasets. Thus, the best-performing decision rule changed across the

Table 4
Within-dataset classifier ranks based on mean five-fold cross-validated macro-F1. A rank of 1 denotes the highest score within a dataset. The final row reports the mean rank across the eight benchmarks. Lower mean ranks indicate better average relative performance.

Dataset	Radial	Gaussian MAP	Logistic regression	XGBoost
Ecoacoustic amphibian vocalizations	4	1	2	3
UrbanSound8K	4	3	1	2
ESC-50	4	2	1	3
MNIST	4	3	2	1
Fashion-MNIST	4	3	2	1
CIFAR-10	4	3	2	1
SVHN	4	3	2	1
CIFAR-100	3	4	2	1
Mean rank	3.875	2.750	1.750	1.625

benchmark suite, directly supporting H1.

The relative positions of the remaining classifiers also varied. Gaussian MAP ranked second on ESC-50, third on UrbanSound8K and four visual datasets, and fourth on CIFAR-100. Logistic regression was more stable, ranking first on two acoustic datasets and second on the remaining six benchmarks. XGBoost achieved the best mean rank overall, but it ranked third on the amphibian and ESC-50 datasets, showing that it did not have a universal advantage. The radial classifier ranked fourth on seven datasets and third on CIFAR-100, indicating that isotropic centroid-based normalization was generally less effective than covariance-aware or discriminative alternatives in the examined posterior mean spaces.

Average ranks were 1.625 for XGBoost, 1.750 for logistic regression, 2.750 for Gaussian MAP, and 3.875 for the radial rule. These averages summarize broad tendencies but do not replace the dataset-specific rankings: XGBoost had the best average rank despite not ranking first on any acoustic dataset, while Gaussian MAP achieved the best score on the amphibian dataset despite a weaker overall mean rank.

Overall, the observed rank reversals support H1. Decision-rule ordering varies across datasets rather than remaining fixed across domains. This finding motivates an examination of whether these changes are associated with measured properties of the corresponding posterior mean geometries, as addressed in Section 5.4.

5.4. Geometry–performance associations

We next examined whether cross-dataset variation in decision-rule advantage was associated with intrinsic properties of the corrected posterior mean embedding spaces. For each of the eight datasets, radial compatibility and complementary spectral descriptors were paired with the macro-F1 contrasts of Gaussian MAP, logistic regression, and XGBoost relative to the radial classifier. Spearman's rank correlation was used as the primary measure of monotonic association, following Section 3.5.

Table 5 reports the resulting coefficients. The strongest observed association was between radial compatibility and the Gaussian MAP advantage, with $\rho_s = 0.643$ and a two-sided $p = 0.086$. This association was positive, indicating that datasets with higher radial compatibility tended, within this benchmark suite, to exhibit a larger Gaussian MAP advantage over the radial classifier, in contrast to the original structural expectation.

Notably, β_D and $k_{eff,90}^{(D)}$ yielded identical Spearman coefficients for all three classifier contrasts because they induced the same rank ordering of the eight datasets in this benchmark suite. This should not be interpreted as mathematical equivalence between the descriptors because their numerical values and geometric meanings remain distinct.

To visualize the strongest observed association in Table 5, Fig. 5 shows the relationship between dataset-level radial compatibility and the mean macro-F1 advantage of Gaussian MAP over the radial classifier. The plot is included as a descriptive aid rather than as an inferential model, and it highlights both the positive direction of the association and the dataset-level variability around this trend.

The Gaussian MAP contrast showed moderate positive associations with spectral slope and effective dimensionality ($\rho_s = 0.548$ in both cases) and a moderate negative association with condition number ($\rho_s = -0.381$). Logistic regression advantages exhibited weaker associations, whereas associations involving the XGBoost contrast were uniformly weak, with absolute coefficients not exceeding 0.214.

These results do not provide consistent support for the directional predictions of H2. Higher radial compatibility was not systematically associated with smaller non-radial advantages, and stronger covariance anisotropy or variance concentration was not consistently associated with larger gains for covariance-aware or discriminative classifiers. Because the analysis contains only eight dataset-level observations, the coefficients should be interpreted descriptively; none of the 12 associations reached the conventional 0.05 significance threshold.

Accordingly, H2 is not supported in its general directional form. While the measured posterior mean geometry remains useful for characterizing individual datasets, its relationship with classifier advantage is more heterogeneous than can be represented by a single monotonic cross-dataset association. This situation calls for a joint structural interpretation of descriptors and classifier rankings rather than a one-dimensional predictive account.

Table 5
Spearman rank correlations between intrinsic dataset-level geometry descriptors and mean macro-F1 contrasts relative to the radial classifier across the eight benchmark datasets. Values in parentheses are two-sided p-values. Positive contrasts indicate an advantage of the corresponding non-radial classifier over the radial baseline.

Descriptor	$\Delta_{\text{MAP}-\text{Radial}}$	$\Delta_{\text{LR}-\text{Radial}}$	$\Delta_{\text{XGB}-\text{Radial}}$
$\mathbf{RC}_D$	0.643 (0.086)	0.357 (0.385)	−0.214 (0.610)
κ_D	−0.381 (0.352)	−0.238 (0.570)	0.048 (0.911)
β_D	0.548 (0.160)	0.429 (0.289)	0.143 (0.736)
$k_{\text{eff},90}^{(D)}$	0.548 (0.160)	0.429 (0.289)	0.143 (0.736)

5.5. Cross-dataset structural summary

The joint pattern of classifier rankings and intrinsic geometry descriptors reveals substantial heterogeneity across the eight posterior mean representation spaces. No simple progression from acoustic to visual datasets, or from lower to higher class count, accounts for the observed decision-rule behavior. Instead, each dataset exhibits a structural profile defined by its combination of radial compatibility, covariance anisotropy, spectral decay, effective dimensionality, and class-specific organization.

SVHN exhibited the lowest dataset-level radial compatibility ($\mathbf{RC}_D = 0.016$), together with the largest condition number, steepest spectral decay, and lowest effective dimensionality. This describes a strongly anisotropic and variance-concentrated representation in which the radial rule performed poorly and XGBoost achieved the highest macro-F1 score. CIFAR-10 also showed relatively low radial compatibility, but with a less extreme covariance spectrum and higher effective dimensionality; XGBoost nevertheless remained the best-performing rule.

The acoustic datasets occupied different geometric regimes despite sharing the same broad input modality. Amphibian vocalizations and ESC-50 had the highest radial compatibility scores, whereas UrbanSound8K showed a lower value. Their classifier rankings nevertheless differed: Gaussian MAP ranked first for amphibian vocalizations, while logistic regression ranked first for UrbanSound8K and ESC-50. Thus, similar values of a single descriptor do not imply identical decision-rule ordering.

Among the visual datasets, MNIST combined comparatively high radial compatibility with a strongly anisotropic covariance spectrum and low effective dimensionality, and all three non-radial classifiers performed similarly well. Fashion-MNIST showed lower radial compatibility, larger condition number, and stronger spectral concentration, with clearer advantages for XGBoost and logistic regression. CIFAR-100 presented a different case: Gaussian MAP performed below the radial classifier despite having higher radial compatibility than SVHN and CIFAR-10 and a condition number lower than those of several other visual datasets. Its large number of classes and fine-grained structure may make class-specific covariance estimates insufficient for capturing the required decision boundaries.

Overall, the structural summary reinforces two conclusions. First, classifier ordering is not invariant across datasets, supporting H1. Second, geometry–performance relationships are heterogeneous and cannot be reduced to a single monotonic descriptor–advantage relation, so H2 is not supported in its general directional form. The descriptors remain informative for characterizing how posterior mean spaces differ, but classifier behavior appears to reflect the joint influence of radial

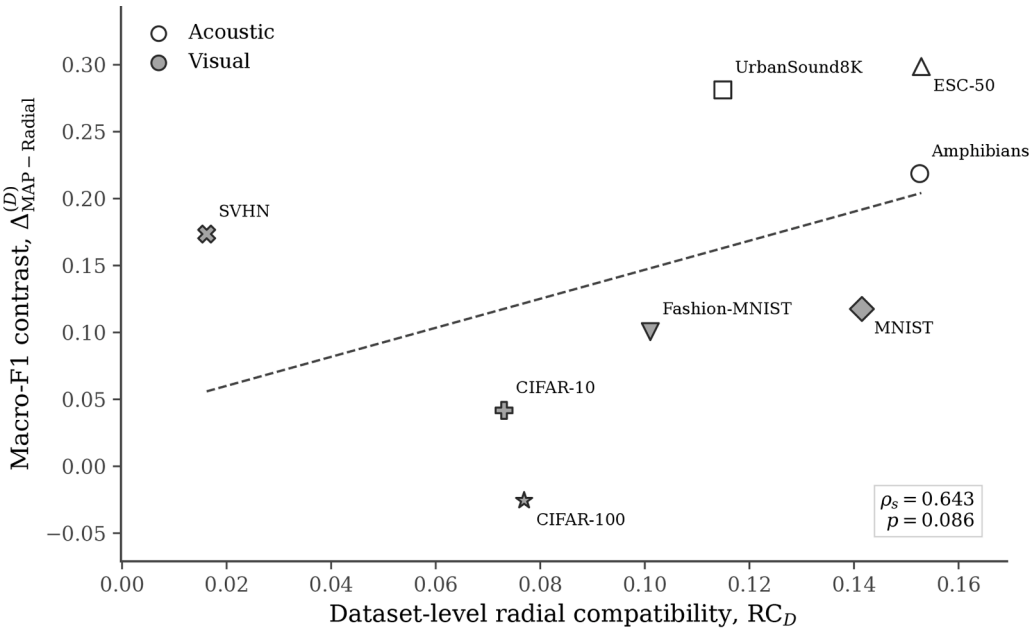

Fig. 5. Association between dataset-level radial compatibility and the mean macro-F1 advantage of Gaussian MAP over the radial classifier across the eight benchmark datasets. Each point represents one dataset; open markers denote acoustic datasets, and filled gray markers denote visual datasets. The dashed line is a least-squares trend included only as a descriptive visual guide. Spearman's rank correlation was positive ($\rho_s = 0.643$, two-sided $p = 0.086$), opposite to the original directional expectation that higher radial compatibility would correspond to a smaller Gaussian MAP advantage.

compatibility, anisotropy, spectral concentration, class count, and class-specific organization rather than any individual descriptor in isolation.

The results therefore support a conditional view of decision-rule alignment. Radial, covariance-aware, affine, and locally adaptive classifiers impose different structural assumptions on a fixed representation, and their relative effectiveness depends on how those assumptions interact with the geometry of each dataset. Intrinsic geometry can help diagnose this interaction, but it should be interpreted as a multidimensional structural profile rather than as a universal scalar predictor of classifier performance.

6. Discussion

The results show that downstream classifier behavior cannot be understood independently of the geometry of the representation on which the decision rule operates. Within each dataset, all four classifiers received the same deterministic 128-dimensional posterior mean embeddings and were evaluated using identical held-out samples. Thus, performance differences reflect how distinct decision mechanisms interact with a fixed representation rather than differences in representation learning, data partitioning, or test-set composition.

The classifier ranking reversals observed across the benchmark suite support H1. Gaussian MAP achieved the highest performance on the ecoacoustic amphibian dataset, logistic regression ranked first on UrbanSound8K and ESC-50, and XGBoost ranked first on all five visual datasets. This variation shows that decision-rule ordering is not invariant across datasets and cannot be reduced to a simple acoustic–visual distinction. The three acoustic datasets produced different winning classifiers and geometric profiles, while the visual datasets differed in the magnitude of discriminative advantage. The relevant interaction is therefore dataset-specific rather than modality-specific.

The weaker performance of the radial classifier indicates that centroid-based isotropic normalization is often too restrictive for the posterior mean spaces examined here. The radial rule represents each class by one centroid and one scalar radius, discarding covariance orientation, directional variation, multimodal organization, and irregular boundaries. The amphibian confusion analysis illustrates this limitation: The radial rule produced pronounced asymmetric confusion between the two *Batrachyla* species and additional errors involving otherwise well-separated classes, whereas Gaussian MAP, logistic regression, and XGBoost substantially reduced these errors on the same embeddings. The fixed representation therefore contained discriminative information that was not fully accessible to the radial decision mechanism.

However, Gaussian MAP was not universally superior. Its strong performance on amphibian vocalizations and MNIST suggests that class-specific covariance can be useful when covariance estimates capture stable directional organization. Conversely, its weaker results on several visual datasets and its performance below the radial classifier on CIFAR-100 indicate that covariance awareness alone is insufficient when class structure is highly irregular, class count is large, or a single Gaussian approximation does not adequately represent decision regions. Under such conditions, affine or locally adaptive discriminative boundaries may exploit the representation more effectively.

The results for H2 were more nuanced. The strongest geometry–performance association was the positive correlation between dataset-level radial compatibility and the Gaussian MAP advantage over the radial classifier, which contradicts the original expectation that higher radial compatibility would reduce the benefit of covariance-aware modeling. This indicates that radial compatibility should not be interpreted as a direct predictor of radial classifier superiority. A dataset may exhibit favorable isotropy and radial separation while still containing the directional covariance structure that Gaussian MAP can exploit. Conversely, low radial compatibility does not determine which alternative classifier will achieve the largest gain.

The spectral descriptors also did not yield a consistent directional account of classifier advantage across all decision rules. Condition number, spectral slope, and effective dimensionality showed moderate associations in some comparisons and weak associations in others. The Spearman coefficients for spectral slope and effective dimensionality were identical because these descriptors induced the same dataset ordering in the benchmark suite, not because they are mathematically equivalent.

Taken together, these findings suggest that posterior mean geometry

is informative primarily as a multidimensional structural profile. Radial compatibility, covariance anisotropy, spectral concentration, effective dimensionality, class count, and class-specific organization characterize complementary aspects of the embedding space, but none of them alone determines the optimal downstream decision rule. Decision rule alignment should therefore be viewed as conditional and configurational rather than as a one-dimensional correspondence between a scalar geometric descriptor and predictive performance.

This interpretation clarifies the role of the proposed geometric analysis. The descriptors are not intended to replace cross-validated evaluation or provide a universal classifier selection rule from eight datasets. Instead, they offer diagnostic information about why classifiers with different structural assumptions may behave differently on a shared representation. The corrected posterior mean framework makes this comparison well defined by ensuring that all geometric quantities and downstream decisions refer to the same deterministic embedding space.

6.1. Limitations

Several limitations should be considered when interpreting the results. First, the cross-dataset geometry–performance analysis is based on only eight dataset-level observations. This limits statistical power, increases the uncertainty of the rank correlation estimates, and prevents broad population-level generalization. The reported Spearman coefficients should therefore be interpreted primarily as descriptive effect sizes within the benchmark suite rather than as definitive evidence of universal cross-domain relationships.

Second, the benchmark datasets differ in domain, sample size, class count, data complexity, encoder architecture, training dynamics, and representation quality. The experimental design controls the representation within each dataset, allowing matched comparisons of decision rules on identical embeddings and held-out samples, but it does not isolate the causal contribution of individual geometric properties across datasets. Because all representations analyzed here were derived from VAEs, the conclusions should also be interpreted within this representation-learning setting; extensions to alternative encoder families are discussed in Section 6.2. Observed associations may reflect combinations of geometry, class granularity, data availability, task difficulty, and representation-learning dynamics.

Third, the geometric descriptors summarize class-conditional structure through dataset-level aggregations. Radial compatibility is averaged across classes, the condition number is summarized by the class median, and spectral slope and effective dimensionality are summarized by class means. These summaries are class-balanced and robust, but they may conceal within-dataset heterogeneity. In addition, radial compatibility captures only two properties directly related to the centroid-based radial rule: within-class isotropy and separation between empirical radial regions. It does not explicitly represent multimodality, nonconvexity, local manifold structure, class-conditional skewness, or higher-order interactions among multiple competing classes.

Fourth, the study evaluates four specific classifiers under fixed configurations. These models represent distinct geometric, probabilistic, affine, and nonlinear decision paradigms, but they do not exhaust the range of downstream methods applicable to generative embeddings. Alternative covariance estimators, kernel methods, neural classifiers, metric-learning approaches, mixture-based probabilistic models, or more flexible prototype-based rules could produce different rankings. Likewise, the relative performance of logistic regression and XGBoost may depend on regularization, optimization, and hyperparameter choices.

Another limitation is the dependence of the spectral descriptors on covariance estimation and numerical regularization in high-dimensional spaces. Ledoit–Wolf shrinkage improves stability, particularly for classes with limited sample size, but trailing eigenvalues and derived quantities such as condition numbers can remain sensitive to sample size and covariance structure. The leading-spectrum restriction used for spectral slope and the 90% variance threshold used for effective dimensionality were fixed a priori, but alternative choices could affect the numerical values of these descriptors.

Overall, the analysis is observational at the dataset level. The geometry descriptors were computed independently of classifier performance, avoiding circular construction of the explanatory variables, but the resulting associations do not establish causality. The study shows that classifier rankings vary across fixed posterior mean representations and that some geometric descriptors covary with performance contrasts, but it cannot determine whether changing a particular descriptor would directly produce a predictable change in classifier advantage.

6.2. Future work

Future work should extend the analysis to a substantially larger and more diverse collection of datasets. Increasing the number of dataset-level observations would provide more stable estimates of geometry–performance associations, enable stronger uncertainty quantification, and support comparisons across domains without relying on a small heterogeneous benchmark suite. Additional acoustic, visual, temporal, and multimodal tasks should be considered, together with systematic variation in sample size, class count, class imbalance, and representation dimensionality.

A second direction is to evaluate whether the observed relationships remain stable across alternative representation-learning procedures. This study uses deterministic posterior mean embeddings from variational autoencoders, but the same controlled decision-layer analysis could be applied to β-VAEs, AAEs, vector-quantized models, normalizing flow-based encoders, diffusion-derived representations, and self-supervised or foundation model embeddings. Repeating the analysis across architectures, latent dimensions, regularization strengths, and random seeds would help distinguish dataset-dependent geometric effects from properties induced by a particular encoder family.

Controlled synthetic experiments would provide a complementary route for testing causal relationships. Synthetic latent spaces could be constructed so that within-class isotropy, covariance anisotropy, centroid separation, overlap, multimodality, non-convexity, and effective dimensionality are varied independently. Such experiments would make it possible to isolate which geometric factors are responsible for changes in radial, covariance-aware, affine, or nonlinear decision-rule advantage.

The geometric characterization could also be expanded beyond dataset-level scalar summaries. Future descriptors may incorporate local density, manifold curvature, multimodality, class-region convexity, neighborhood purity, higher-order moments, and multiclass overlap. Class-level or hierarchical models of geometry may be especially useful for explaining within-dataset heterogeneity and for identifying classes whose structure violates the assumptions of a particular decision rule.

Another important direction is to develop multivariate geometry–performance models. The results suggest that classifier suitability is configurational rather than determined by a single descriptor. With a larger benchmark collection, combinations of radial compatibility, covariance anisotropy, spectral concentration, effective dimensionality, class count, and sample availability could be examined using regularized regression, hierarchical models, or other interpretable multivariate approaches. Such models should be validated on held-out datasets to avoid replacing descriptive association with an overfitted classifier selection rule.

A particularly important direction is to examine whether geometry can guide decision-rule selection prospectively. In this setting, descriptors would be estimated from training data and used to recommend, initialize, or adapt a downstream classifier before held-out evaluation. Nested validation across independent datasets would be essential to determine whether geometric diagnostics can improve classifier selection without information leakage. Doing so would extend this work from

post hoc structural interpretation toward geometry-aware decision modeling in generative embedding spaces.

7. Conclusion

This study examined the interaction between downstream decision rules and the geometry of fixed deterministic posterior mean representations learned by variational autoencoders. By holding the representation, cross-validation partitions, and held-out samples constant within each dataset, the analysis isolated the decision layer and enabled a controlled comparison of a centroid-based radial rule, Gaussian MAP, logistic regression, and XGBoost across eight heterogeneous acoustic and visual benchmarks.

The empirical results support H1, which stated that classifier ranking varies across datasets. Gaussian MAP, logistic regression, and XGBoost each achieved the highest performance on different subsets of the benchmark suite, demonstrating that no single decision paradigm was universally optimal. By contrast, H2 was not supported in its general directional form. The strongest observed geometry–performance association was positive and opposite to the original expectation, while the remaining associations did not reveal a consistent monotonic pattern across classifiers and descriptors.

These findings indicate that downstream classifier effectiveness should be understood as a conditional alignment problem between decision-rule assumptions and the geometry of the fixed representation. Radial compatibility, covariance anisotropy, spectral concentration, and effective dimensionality provide useful structural diagnostics, but none of the individual descriptors considered here functions as a universal predictor of classifier advantage. The relevant geometry appears to be multidimensional and dataset-dependent.

Therefore, instead of identifying a universally superior classifier, this work mainly demonstrates that decision-rule suitability varies systematically across fixed generative representations and should be evaluated in relation to their class-conditional structure. This perspective offers a principled basis for future geometry-aware analysis and classifier selection in generative embedding spaces.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the authors used OpenAI's ChatGPT to support language editing, consistency checks, manuscript organization, and the preparation of responses to reviewer comments. The authors reviewed and edited all AI-assisted output and take full responsibility for the content of the submitted article.

Funding

This research was supported by the Chilean National Agency for Research and Development through FONDECYT grant no. 1250388. The funding source had no involvement in study design, data analysis, the interpretation of results, or manuscript preparation.

CRediT authorship contribution statement

Víctor Poblete: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Conceptualization. **José Aillapi:** Writing – review & editing, Validation, Software, Data curation. **Carlos Duarte:** Writing – review & editing, Validation, Software. **Luis Veas:** Writing – review & editing, Software, Resources. **Felipe Otondo:** Writing – review & editing, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work used the high-performance computing resources of the Patagón supercomputer at the Universidad Austral de Chile (FONDEQUIP EQM180042). The authors thank the Patagón technical team for their support and for granting access to computational resources.

Data availability

The datasets used and/or analyzed for this study are publicly available for standard benchmarks. The ecoacoustic amphibian dataset is available from the corresponding author upon reasonable request.

References

[1] L. D'Amato, G. Lancia, G. Pezzulo, The geometry of efficient codes: how rate-distortion trade-offs distort the latent representations of generative models, PLoS Comput. Biol. 21 (5) (2025) 1–30, https://doi.org/10.1371/journal.pcbi.1012952.
[2] K. Gibb, A. Eldridge, C. Sandom, I. Simpson, Towards interpretable learned representations for ecoacoustics using variational auto-encoding, Ecol. Inform. 80 (2024) 1–17, https://doi.org/10.1016/j.ecoinf.2023.102449.
[3] F. Kaechele, M. Coblenz, O. Grothe, A comparison of latent space modeling techniques in a plain-vanilla autoencoder setting, Mach. Learn. 114 (7) (2025) 1–35, https://doi.org/10.1007/s10994-025-06784-3.
[4] A. Asperti, V. Tonelli, Comparing the latent space of generative models, Neural Comput. Appl. 35 (4 SI) (2023) 3155–3172, https://doi.org/10.1007/s00521-022-07890-2.
[5] L. Tětková, T. Brüsch, T. Dorszewski, F. Mager, R. Aagaard, J. Foldager, T. Alstrom, L. Hansen, On convex decision regions in deep network representations, Nat. Commun. 16 (1) (2024) 1–12, https://doi.org/10.1038/s41467-025-60809-y.
[6] A. Ganguli, N. Ramachandra, J. Bessac, E. Constantinescu, Enhancing interpretability in generative modeling: statistically disentangled latent spaces guided by generative factors in scientific datasets, Mach. Learn. 114 (9) (2025) 1–17, https://doi.org/10.1007/s10994-025-06816-y.
[7] I. Komorska, A. Puchalski, Condition monitoring using a latent space of variational autoencoder trained only on a healthy machine, Sensors 24 (21) (2024) 1–14, https://doi.org/10.3390/s24216825.
[8] R. Guidotti, A. Monreale, S. Matwin, D. Pedreschi, Black box explanation by learning image exemplars in the latent feature space, Eds., in: U. Brefeld, E. Fromont, A. Hotho, A. Knobbe, M. Maathuis, C. Robardet (Eds.), Proceedings of the Lecture Notes in Artificial Intelligence, European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Wurzburg, Germany, 2019, pp. 189–205, https://doi.org/10.1007/978-3-030-46150-8_12.
[9] M. Fernandez-Delgado, E. Cernadas, S. Barro, D. Amorim, Do we need hundreds of classifiers to solve real world classification problems? J. Mach. Learn. Res. 15 (2014) 3133–3181.
[10] O. Ledoit, M. Wolf, A well-conditioned estimator for large-dimensional covariance matrices, J. Multivar. Anal. 88 (2) (2004) 365–411, https://doi.org/10.1016/S0047-259X(03)00096-4.
[11] J. Schäfer, K. Strimmer, A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics, Stat. Appl. Genet. Mol. Biol. 4 (1) (2005), https://doi.org/10.2202/1544-6115.1175.
[12] C. Bishop, Pattern Recognition and Machine Learning, Springer, New York, 2006.
[13] T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794, https://doi.org/10.1145/2939672.2939785.