Interpretable Modeling and Visualization of Biomedical Data

S.Ghosh*[a,e], E.S. Baranowski[b,c], M. Biehl[a,f], W. Arlt[b,c,g,h], P.Tino[d], K.Bunte[a]

[a] *Bernoulli Institute of Mathematics, Computer Science and Artificial Intelligence, University of Groningen, Nijenborg 9, Groningen, 9747AG, NL*
[b] *Institute of Metabolism and Systems Research, The University of Birmingham, Birmingham, B15 2TT, UK*
[c] *Birmingham Children's Hospital, Birmingham Women's and Children's NHS Foundation Trust, Birmingham, B4 6NH, UK*
[d] *School of Computer Science, The University of Birmingham, Birmingham, B15 2TT, UK*
[e] *Department of Mathematics and Computer Science, Technical University of Eindhoven, Eindhoven, 5612 AZ, NL*
[f] *Center for Systems Modelling and Quantitative Biomedicine, Institute of Metabolism and Systems Research, The University of Birmingham, Birmingham, B15 2TT, UK*
[g] *Medical Research Council London, Institute of Medical Sciences, London, W12 0NN, UK*
[h] *Faculty of Medicine, Imperial College, London, SW7 2AZ, UK*

Abstract

Applications of interpretable machine learning (ML) techniques on medical datasets facilitate early and fast diagnoses, along with getting deeper insight into the data. Furthermore, the transparency of these models increase trust among application domain experts. Medical datasets face common issues such as heterogeneous measurements, imbalanced classes with limited sample size, and missing data, which hinder the straightforward application of ML techniques. In this paper we present a family of prototype-based (PB) interpretable models which are capable of handling these issues. Moreover we propose a strategy of harnessing the power of ensembles while maintaining the intrinsic interpretability of the PB models, by averaging over the model parameter manifolds. All the models were evaluated on a synthetic (publicly available dataset) in addition to detailed analyses of two real-world medical datasets (one publicly available). The models and strategies we introduce address the challenges of real-world medical data, while remaining computationally inexpensive and transparent. Moreover, they exhibit similar or superior in performance compared to alternative techniques.

Keywords: imbalanced classification, missing data, learning vector quantization, dissimilarity learning, dimensionality reduction, visualization

1. Introduction

For Machine Learning (ML) techniques to be applied on anthropocentric sectors, such as healthcare, judiciary, finance, it is crucial that their decision-making mechanism are understandable and explainable by human experts [1–3]. In such domains, performance metrics of a ML model are typically not enough to ensure its accountability and trustworthiness [3; 4]. Additionally, this ensures improved fairness and prevents biased learning [3–5]. Explainable AI (XAI) techniques applied *pre-training* (e.g., Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding (tSNE) [6]) or *post-training* (e.g., Local interpretable model-agnostic explanations (LIME) [4], DeepView [7], Feature Relevance Information (FRI) [8] and SHapley Additive exPLanations (SHAP)[9]) of a model can approximate local explanations of the latter (trained) model's decisions, using simpler surrogate models. Being model-agnostic by definition, these XAI techniques cannot access the actual working-logic of the trained models, and can therefore be fooled by adversarial classifiers, as demonstrated in [10]. On the contrary, intrinsically interpretable models (e.g., decision trees (DTs), linear/logistic regression, K-nearest neighbours (KNNs), and nearest prototype based classifiers (NPCs) [5], among others) can illustrate their respective decision-making or working logic straightforwardly, and globally. We follow these three proposed criteria for evaluating intrinsic interpretability/explainability [5; 11; 12] (referred to henceforth as IICs) of different classifiers: namely the model's intrinsic ability to (IIC-1) select features from the input pattern ([3] describes as *saliency*); (IIC-2) assign class-specific representatives; and (IIC-3) provide direct information about the decision-boundary.

Particularly medical classification problems often exhibit complications, such as heterogeneous measurements, high class imbalance and large amounts of missing data, which are difficult for ML in general. Heterogeneity arises due to different data collection techniques for different groups within the same medical cohort, such as, babies and children versus adolescents and adults. Moreover, even the normal range of a measured physiological feature often varies greatly with subject's age, sex, or BMI, etc. Class imbalance is common in astronomy (to find galaxies [13]); telecommunications management (to detect fraudulent calls); geo-spatial image analysis (rubble and oil-spills detection) [14]; and in healthcare (identifying rare conditions). Unaddressed, it can lead to biased and subsequent poor classification performance, due to the minority classes being under-represented during training and the overall accuracy often failing to reflect the true performance. While Bayesian methods, employing class priors, handle this challenge in an embedded manner [15], prominent model-agnostic strategies include bagging, boosting, and sampling (including oversampling, undersampling [16; 17] and SMOTE [14]). For missing values broadly three categories are outlined [1; 18–20]: (i) *missing completely at random (MCAR)*, if the missingness is

neither dependent on the observed nor on the missing values of the dataset; (ii) *missing at random (MAR)*, if the missingness is independent of the missing values but likely to be dependent on the observed values; and (iii) *missing not at random (MNAR)*, if the missingness is dependent on the missing values themselves. The cause for this can be systematic, such as the instrument failing to record a parameter due to censoring, or due to dataset being compounded from different studies or labs, which were not measuring the same variables. Imputation is a commonly used model-agnostic, strategy to fill missing data entries during preprocessing, such that subsequently any classifier can be used on the imputed set. Non-linear strategies, such as *not missing at random importance weighted autoencoder, (not-MIWAE)* [21], backpropagate through the samples with observed data to find unbiased estimates of gradients of the autoencoder bounds. However, non-linear techniques typically require large amount of training instances and hence are often not applicable if the size is very limited, as often encountered in the medical domain. For the latter multiple imputation (MI) is generally effective against MCAR and MAR [22]. An example is the openly accessible regression-based (linear) technique called Multivariate Imputation by Chained Equations (MICE) [23; 24] that is often used, especially using the predictive mean matching (PMM) strategy proposed by [25]. Alternatively ML strategies which can handle partially observed data, were introduced, such as (i) generative modelling, and (ii) Partial Distance strategy (PDS)) [26–28]. Examples of latter include distance-based ML, that incorporate a scaling factor for computing distances dependent on the available dimensions, such as KNN, NaN-Learning Vector Quantization (NaN-LVQ) [29], and Angle LVQ [30; 31]. Prominent generative models such as Linear Discriminant Analysis (LDA), Probabilistic PCA (PPCA), and Factor Analysis (FA) show promising results for MCAR and MAR (ignorable missingness) but cannot necessarily be assumed to work well on MNAR [1; 32–34].

Prototype-based machine learning techniques [35; 36] incorporating adaptive dissimilarities have shown promisingly robust with regard to heterogeneous measurements, such as different lighting conditions, and small training set sizes [37; 38]. They have furthermore been incorporated as layers in Neural Network architectures [39–42]. Moreover, they can handle missing data and are intrinsically interpretable, in terms of feature importance information, class-specific representatives and information about the decision boundary [29–31]. Since healthcare demands interpretability, while presenting the aforementioned challenges in data quality and quantity this paper extends our conference contributions [30; 31] twofold. Firstly, the original cost-function is extended to enable (i) learning from probabilistic (or uncertain) labels and (ii) returning class probabilities as classification result. Secondly, we introduce a geodesic and rank-preserving model-averaging and clustering strategy for ML models based on low-rank quadratic form dissimilarities, such as Large Margin Nearest Neighbour (LMNN) [43], Near-

est Component Analysis (NCA) [44], and Learning vector Quantization (LVQ) [35; 36]. This proposed strategy enables identification and interpretation of local optima, and generally shows a more robust performance than that of a single classifier, while retaining the interpretability of an individual model (unlike any ensembling approach). Thereafter we systematically investigate the influence of (i) missing values of types MCAR and MNAR, (ii) the amount of missingness and (iii) the training set size to compare the classification performance of several common strategies to deal with such problems. We furthermore demonstrate with real-world medical datasets, that our proposed approaches are not only *competitive* in terms of performance, but are also easily and intuitively interpretable.

In Sec. 2 we present the works related to our proposed novel contributions, followed by the aforementioned contributions themselves (Sec. 3, 4 and 5). Next we illustrate the role of relevant hyperparameters and compare the performance of our proposed methods to that of the state-of-the-art transparent ML techniques, on publicly available synthetic datasets in Sec. 6. Thereafter we compare the performance and interpretability of two real-world medical datasets in Sec. 7 and 8, respectively. Finally Sec. 9 highlights the strengths of our novel contributions against relevant existing interpretable ML techniques, in the light of some of the demands and challenges posed by healthcare applications.

2. Related work

2.1. Prototype based classifiers

A family of prototype based classifiers (PBCs) is based on the concept of LVQ, which follows the Nearest Prototype Classification (NPC) scheme. This means a new vector is assigned the class label of the prototype to which it is closest, according to a chosen dissimilarity measure [35]. Techniques implementing this concept are computationally efficient and often allow interpretation of the prototypes as representatives of classes ensuring transparency with regards to IIC-2. Assume the data consist of N instances $\boldsymbol{x}_i \in \mathbb{R}^{\mathrm{D}}$ accompanied by labels y_i denoting one of C classes and let $\boldsymbol{w}^j \in \mathbb{R}^{\mathrm{D}}$ denote one of C prototypes with labels $c(\boldsymbol{w}^j)$. Generalized LVQ (GLVQ) [36] performs a supervised training procedure aimed at minimizing the cost-function:

$$E = \sum_{i=1}^{N} f\left(\lambda_i\right), \text{ where } \lambda_i = \frac{d_i^J - d_i^K}{d_i^J + d_i^K} \ , \tag{1}$$

which exhibits a large margin principle, proven by [45]. Here, the dissimilarity of each data sample $\boldsymbol{x}_i$ to its nearest correct prototype with $y_i = c(\boldsymbol{w}^J)$ is defined by d_i^J, and by d_i^K for the nearest wrong prototype ($y_i \neq c(\boldsymbol{w}^K)$). f is a monotonic function and we use the identity ($f(a) = a$) in this contribution. The cost-function Eq. (1) is non-convex and can be optimized using gradient methods, such

as the Fast Limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm (L-BFGS) [46–50] used in this contribution. Extensions to GLVQ named Generalized Matrix LVQ (GMLVQ) [51–53] introduced parameterized adaptive dissimilarity measures, such as the quadratic form:

$$d_i^L = (\boldsymbol{x}_i - \boldsymbol{w}^L)^\top \Lambda (\boldsymbol{x}_i - \boldsymbol{w}^L) \quad \text{with} \sum_i \Lambda_{ii} = 1 \ , \tag{2}$$

with $L \in \{J, K\}$ and a positive semi-definite (PSD) matrix $\Lambda \in \mathbb{R}^{\mathrm{D}\times\mathrm{D}}$. The latter contains additional parameters for optimization, that allows feature relevance detection for IIC-1. The complexity of the metric tensor can be increased, from locally linear decision-boundaries to non-linear ones, through local or classwise matrices $\Omega_{L|c}$ attached to the prototypes [54]. This family of interpretable classifiers can intuitively deal with missing data as illustrated in [30], where the authors compared two variants of these PBCs, (i) applying PDS on Euclidean distance (NaNLVQ), and the other using a parametrized cosine-dissimilarity (Angle LVQ).

2.2. Angle LVQ

In Angle LVQ (ALVQ) the quadratic form $d_i^{\{J,K\}}$ in Eq. (2) is replaced by a parameterized angle-based dissimilarity:

$$\begin{aligned} d_i^L =& g_\beta\left(b\right), \text{ where } b = b_\Lambda(\boldsymbol{x_i}, \boldsymbol{w}^L) = \frac{\boldsymbol{x}_i^\top \Lambda \boldsymbol{w}^L}{\|\boldsymbol{x}_i\|_\Lambda \|\boldsymbol{w}^L\|_\Lambda} \ , \text{ with } \|\boldsymbol{v}\|_\Lambda = \sqrt{\boldsymbol{v}^\top \Omega^\top \Omega \boldsymbol{v}}, \\ \Lambda =& \Omega^\top \Omega \text{ and } g_\beta(b) = \frac{e^{-\beta(b-1)} - 1}{e^{(2\beta)} - 1} \text{ with } \sum_i \Lambda_{ii} = 1, \text{ and } L \in \{J, K\}. \end{aligned} \tag{3}$$

Here, the exponential function $g_\beta(b)$ transforms the cosine $b = \cos\theta \in [-1, 1]$ into dissimilarities in the range [0,1]. Hyperparameters include the number of prototypes used to represent each class (fixed to one in this contribution), β (we use $\beta = 1$ unless explicitly stated otherwise), and the choice of the metric tensor. Similar to its Euclidean counterparts, the dissimilarity measure d_i^L can be localized with varying potential for further interpretation [31]. The generalization bounds can be estimated using the Rademacher complexity similar to LGMLVQ [54]. The derivatives of Eq. (3) and the local variants can be found in appendix A.1. In the presence of missingness ALVQ the cosine-dissimilarity b and its derivates are computed on the observable dimensions only, similar to its Euclidean predecessor NaNLVQ. However, ALVQ is more robust than Euclidean distance with PDS (NaNLVQ) with higher missingness, as illustrated in [30].

In the presence of class-imbalance one way to counter the impact of majority class samples is the introduction of a penalty-matrix in the cost-function [30; 55]:

$$\hat{E} = \sum_{c=1}^{C} \frac{1}{n_c} \left[\sum_{\boldsymbol{x}_i, s.t. y_i = c} \gamma_{c,\hat{y}_i} \lambda_i \right] \ , \tag{4}$$

where $c = y_i$ is the class label of training sample $\boldsymbol{x}_i$, n_c defines the number of samples within that class, $\hat{y}_i$ is the predicted label and λ_i is the cost-function value of sample i Eq. (1). This allows for stricter penalisation of certain types of misclassifications, such as misclassifications of the minority class samples (e.g., patients) as a majority class (e.g., Healthy).

Sampling-based strategy to address class-imbalance. The widely used imbalanced-handling strategy Synthetic Minority Oversampling (SMOTE) [14] in its original formulation operates in Euclidean space. Hence, for ALVQ, which operates on a hypersphere, we introduced a geodesic variant, referred to as the geodesic SMOTE. To do so it uses the exponential map (from Riemannian geometry), with origin G (Log_G) where the tangent space τ_G of the manifold is constructed [50; 56]. Let ζ be a point on the manifold and $\hat{\zeta}$ the corresponding point in the

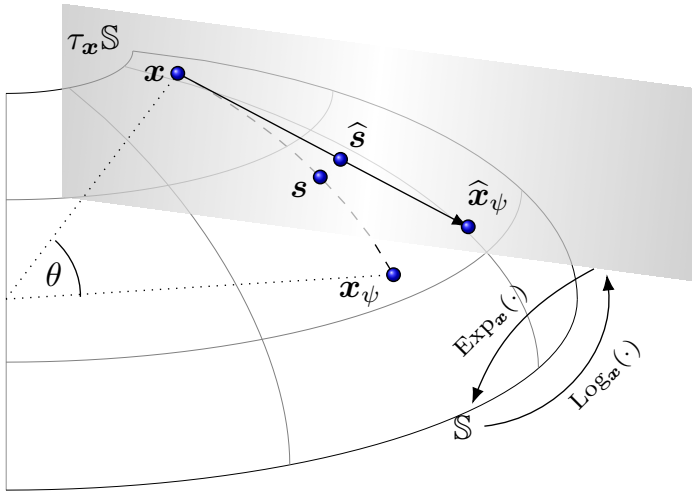


Figure 1: Depiction of geodesic SMOTE to generate synthetic samples $\boldsymbol{s}$ on the hypershpere to oversample minority classes for imbalanced data using Riemannian geometry.

tangent space with $\hat{\zeta} = \text{Log}_G(\zeta)$, $\zeta = \text{Exp}_G(\hat{\zeta})$ and $d_g(\zeta, G) = d_e(\hat{\zeta}, G)$ with d_g being the geodesic distance between the points on the manifold and d_e being the Euclidean distance on the tangent space. Log and Exp denote a mapping of points from the manifold to the tangent space and vice versa. Geodesic SMOTE then presents a point $\boldsymbol{x}$ from class c on the unit sphere with fixed length $\|\boldsymbol{x}\| = 1$, that becomes τ_G. Next, k nearest neighbours ($\boldsymbol{x}_\psi$) of $\boldsymbol{x}$ are found from the same class $\boldsymbol{x}_\psi \in \mathcal{N}_{\boldsymbol{x}}$ using the d_g between the vectors $\theta = \cos^{-1}((\boldsymbol{x}^\top \boldsymbol{x}_\psi)/r^2)$ and $r = 1$. Each $\boldsymbol{x}_\psi$ is then projected onto that tangent space using only the available dimensions

and Log transformed for spherical manifolds:

$$\hat{\boldsymbol{x}}_\psi = \mathrm{Log}_{\boldsymbol{x}}(\boldsymbol{x}_\psi) = \frac{\theta}{\sin\theta}(\boldsymbol{x}_\psi - \boldsymbol{x}\cos\theta) \ . \tag{5}$$

A synthetic sample can either be produced in a two-stage approach or directly as visualized in Fig. 1. For the former the sample is generated on the tangent space similar to the original SMOTE [14] and subsequently projected onto the sphere via Exp [31]. For the latter it is generated on the geodesic directly, using the new angle $\hat{\theta} = \|\widehat{\boldsymbol{x}}_\psi\|$ and the Exp transformation, given by:

$$s = \boldsymbol{x}\cos(\hat{\theta}\alpha) + \frac{\sin(\hat{\theta}\alpha)}{\hat{\theta}} \cdot \widehat{\boldsymbol{x}}_\psi \qquad \text{with } \alpha \in]0,1[\ . \tag{6}$$

This procedure is repeated with other random samples from a minority class until the desired number of training samples is reached.

3. A probabilistic variant of ALVQ

In the medical domain patients can have multiple comorbidities instead of a single crisp condition, or they can have a diagnosis which shows phenotypic similarity or overlap with other conditions. If the classifier could estimate the certainty of a patient belonging to all the conditions it was trained upon, then this can constitute useful information, for further investigations or for treatment planning. Moreover, some diseases may be difficult to diagnose, and might result in different diagnoses given by the different consulting doctors. This can be expressed as probability of a class assignment dependent on the fraction of experts who agree on that class. Therefore, we develop a probabilistic version of ALVQ, which allows to express our model's (un)certainty about the class label, given an input, in the form of conditional probability distribution over the classes. [57], [58] and [59; 60] used information theoretical principles to generalize Robust Soft LVQ (RSLVQ), by using maximum likelihood and the Cross-Entropy (CE) as the cost-function. In our formulation, which is closely related to the CE in [58], we estimate the class when the sample $\boldsymbol{x}$ is given, by minimizing the difference between the true class and our estimate i.e., by minimizing the Kullback-Leibler (KL) divergence (D_{KL}) in the cost-function.

Consider the unknown joint distribution $p(\boldsymbol{x}, c) = p(c|\boldsymbol{x})p(\boldsymbol{x})$ over the inputs and labels that generated our training set $\{(\boldsymbol{x}_i, c_i)\}_{i=1}^N$. Our discriminative model produces an estimate $\hat{p}(c|\boldsymbol{x})$ of $p(c|\boldsymbol{x})$. The expected KL divergence measuring the mismatch between $\hat{p}(c|\boldsymbol{x})$ and $p(c|\boldsymbol{x})$ can be approximated through the training

sample as

$$
\begin{aligned}
H(\hat{p}(c|\cdot)) &= E_{p(\boldsymbol{x})}[D_{KL}(\hat{p}(c|\boldsymbol{x}) \parallel p(c|\boldsymbol{x}))] \\
&\approx \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}\hat{p}(c|\boldsymbol{x}_i)[\ln\hat{p}(c|\boldsymbol{x}_i) - \ln p(c|\boldsymbol{x}_i)] \\
&= \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}\hat{p}(c|\boldsymbol{x}_i)\ln\frac{\hat{p}(c|\boldsymbol{x}_i)}{p(c|\boldsymbol{x}_i)}.
\end{aligned} \tag{7}
$$

Since we do not have access to the true distributions $p(c|\boldsymbol{x}_i)$ the cost-function is often formulated by considering only the generated labels c_i. For the case of the generated sample $(\boldsymbol{x}_i, c_i)$ being noise-free $p(c_i|\boldsymbol{x}_i) = 1$ (for example when the diagnosis is genetically confirmed) other classes have a probability of 0 and KL divergence cannot be used. In such cases one can either simplify the cost-function considering only c_i for $\boldsymbol{x}_i$: $\frac{1}{N}\sum_{i=1}^{N}\hat{p}(c_i|\boldsymbol{x}_i)\ln\hat{p}(c_i|\boldsymbol{x}_i)$ or introduce some noise by subtracting ϵ from the class and adding $\epsilon/(C-1)$ to the others. In the following we assume the latter and provide the detailed derivatives for noisy labels. For sample $\boldsymbol{x}_i$ the $\hat{p}(c|\boldsymbol{x}_i)$ is computed by the following parameterized soft-max function:

$$
\hat{p}(c|\boldsymbol{x}_i) = \frac{g_\Theta\left(\frac{\boldsymbol{x}_i\Lambda\boldsymbol{w}^{c\top}}{\|\boldsymbol{x}_i\|_\Lambda\|\boldsymbol{w}^c\|_\Lambda}\right)}{\sum_j^C g_\Theta\left(\frac{\boldsymbol{x}_i\Lambda\boldsymbol{w}^{j\top}}{\|\boldsymbol{x}_i\|_\Lambda\|\boldsymbol{w}^j\|_\Lambda}\right)} \quad \text{with} \quad g_\Theta(b) = \frac{e^{\Theta(b+1)}-1}{e^{(2\Theta)}-1} \ . \tag{8}
$$

The parameter Θ can be interpreted as $\frac{1}{k_B T}$ where k_B is the Boltzmann constant and T is the absolute temperature, which determines the crispness of the decision-boundaries (see subsec. 6.1). The derivatives of $D_{KL}(\hat{p}(c|\boldsymbol{x}) \parallel p(c|\boldsymbol{x}))$ (Eq. (7)) with $\|\boldsymbol{v}\|_\Omega = \sqrt{\boldsymbol{v}^\top\Omega^\top\Omega\boldsymbol{v}}$ are:

$$
\frac{D_{KL}(\hat{p}(c|\boldsymbol{x}_i) \parallel p(c|\boldsymbol{x}_i))}{\partial\Omega} = \sum_{c=1}^{C}\frac{\partial\hat{p}(c|\boldsymbol{x}_i)}{\partial\Omega}\cdot\left(1+\ln\frac{\hat{p}(c|\boldsymbol{x}_i)}{p(c|\boldsymbol{x}_i)}\right) \tag{9}
$$

$$
\text{and} \quad \frac{D_{KL}(\hat{p}(c|\boldsymbol{x}_i) \parallel p(c|\boldsymbol{x}_i))}{\partial\boldsymbol{w}^j} = \sum_{c=1}^{C}\frac{\partial\hat{p}(c|\boldsymbol{x}_i)}{\partial\boldsymbol{w}^j}\cdot\left(1+\ln\frac{\hat{p}(c|\boldsymbol{x}_i)}{p(c|\boldsymbol{x}_i)}\right) \tag{10}
$$

Now $\frac{\partial\hat{p}(c|\boldsymbol{x}_i)}{\partial\Omega}$ can be expanded to

$$
\frac{\frac{\partial g_\Theta(b_\Omega(\boldsymbol{x}_i,\boldsymbol{w}^c))}{\partial\Omega}\sum_{j=1}^{C}g_\Theta(b_\Omega(\boldsymbol{x}_i,\boldsymbol{w}^j)) - g_\Theta(b_\Omega(\boldsymbol{x}_i,\boldsymbol{w}^c))\sum_{j=1}^{C}\frac{\partial g_\Theta(b_\Omega(\boldsymbol{x}_i,\boldsymbol{w}^j))}{\partial\Omega}}{(\sum_{j=1}^{C}g_\Theta(b_\Omega(\boldsymbol{x}_i,\boldsymbol{w}^j)))^2} \ . \tag{11}
$$

Similarly,

$$
\frac{\partial\hat{p}(c|\boldsymbol{x}_i)}{\partial\boldsymbol{w}^c} = \begin{cases} \frac{\frac{\partial g_\Theta(b_\Omega(\boldsymbol{x}_i,\boldsymbol{w}^c))}{\partial\boldsymbol{w}^c}\cdot\sum_{j=1}^{C}g_\Theta(b_\Omega(\boldsymbol{x}_i,\boldsymbol{w}^j)) - g_\Theta(b_\Omega(\boldsymbol{x}_i,\boldsymbol{w}^c))\cdot\frac{\partial g_\Theta(b_\Omega(\boldsymbol{x}_i,\boldsymbol{w}^c))}{\partial\boldsymbol{w}^c}}{\left(\sum_{j=1}^{C}g_\Theta(b_\Omega(\boldsymbol{x}_i,\boldsymbol{w}^j))\right)^2} & \text{if j=c} \\ \frac{-g_\Theta(b_\Omega(\boldsymbol{x}_i,\boldsymbol{w}^c))\cdot\frac{\partial g_\Theta(b_\Omega(\boldsymbol{x}_i,\boldsymbol{w}^j))}{\partial\boldsymbol{w}^j}}{\left(\sum_{k=1}^{C}g_\Theta(b_\Omega(\boldsymbol{x}_i,\boldsymbol{w}^k))\right)^2} & \text{if} j \neq c \end{cases} \tag{12}
$$

$$\text{where, } \frac{\partial g_\Theta(b_\Omega(\boldsymbol{x}_i, \boldsymbol{w}^j))}{\partial \Phi} = \frac{\Theta e^{\Theta(b_\Omega(\boldsymbol{x}_i, \boldsymbol{w}^j))+1)}}{e^{2\Theta} - 1} \cdot \frac{\partial b_\Omega(\boldsymbol{x}_i, \boldsymbol{w}^j))}{\partial \Phi} \quad \text{with } \Phi \in \{\Omega, \boldsymbol{w}^j\} \ . \tag{13}$$

The partial derivatives $\frac{\partial b_\Omega}{\partial \boldsymbol{w}}$ and $\frac{\partial b_\Omega}{\partial \Omega}$ are the same as the original ALVQ and detailed in Eq. (A. 24) and (A. 25).

4. Geodesic average model

Ensembling is a well known strategy to avoid overfitting and improve on the generalization of ML algorithms [61; 62]. However, the improved performance by combining independently trained models comes at the cost of: (i) increased computational and memory cost needed to keep all the constituent models of the ensemble; and (ii) sacrificing the interpretability offered by the individual models. In this section we propose and investigate a different strategy, namely to build a geodesic average model that retains interpretability while avoiding overfitting effects by combining parameter information of independently trained models.

4.1. Geodesic average over model parameters

In order to build an average of k models we compute the geometric mean of each of the model parameters, namely the trained prototypes of each class $W_c = \{\boldsymbol{w}_{\{c,k\}}\}_{i=1}^k$ and the positive semi-definite matrices Λ_k. We restrict the description for one prototype per class here, each initialized close to the class means. With random initialization one might need to rotate the coordinate system to align the prototypes before averaging. If using several prototypes per class the correct index for averaging can be found using the geodesic distance of the set of prototypes within each class. The β (or Θ) parameter is a positive scalar and typically fixed or found by line search. Classification by geodesic LVQ^A variants (Eqs. (3), (A. 26) and (8)) takes place on the hypershpere and the geometric mean of the model prototypes of each class $\overline{\boldsymbol{w}}_c \in \mathcal{M}$ in the Riemannian interpretation, known as Karcher mean [63], is the point in $\mathcal{M}$ that minimizes the sum of squared geodesic distances:

$$\overline{\boldsymbol{w}}_c = \arg\min_{\boldsymbol{w} \in \mathcal{M}} \sum_{i=1}^{k} d_{\text{geod}}(\boldsymbol{w}_{\{c,i\}}, \boldsymbol{w})^2 \qquad \text{with } c \in \{1, \ldots, C\} \ , \tag{14}$$

with $\boldsymbol{w}_{\{c,i\}}$ being the prototype of class c of individual model i. In the Euclidean LVQ variants, GRLVQ, GMLVQ and LGMLVQ, the geodesic distance is simply Euclidean. In case of $\mathcal{M}$ being the hypersphere the geodesic distance is $d_{\text{geod}}(\boldsymbol{w}_i, \boldsymbol{w}_j) = \cos^{-1}(\frac{\boldsymbol{w}_i \boldsymbol{w}_j}{\|\boldsymbol{w}_i\| \|\boldsymbol{w}_j\|})$. This mean exists and is uniquely defined only as the set of prototypes W_c is contained in an open half-sphere, which means a convexity radius of $\pi/2$, and is typically computed by non-linear optimization methods [63–65]. However, computing the geometric mean of the positive semi-definite (PSD) matrices Λ_k is less straightforward.

The computation of geometric means of positive definite (PD) matrices as proposed by [66] has received considerable attention due to its relevance for numerous applications, ranging from control theory, convex programming, mercer kernels and diffusion tensors in medical imaging. However, the computation of this *Ando mean* is not rank-preserving, resulting almost surely in a rank null for matrices with rank $M < D/2$ [67]. Due to the growing interest in low-rank approximations in large-scale applications, [67; 68] introduced and extended the geometric mean to the set of PSD matrices $S^+(M, D)$ of fixed rank M using a Riemannian framework. Their approach bases on the decomposition of each of the k metric tensors

$$\Lambda_i = U_i R_i^2 U_i^\top \qquad \text{for } i = 1 \dots k \tag{15}$$

exhibiting the geometric interpretation of PSD matrices in $S^+(M, D)$ as flat M-dimensional ellipsoids in $\mathbb{R}^D$. Here U_i is element of the Stiefel manifold $St(M, D)$, which denotes the set of all orthonormal M-frames in $\mathbb{R}^D$. Thus, the columns of each U_i forms an orthonormal basis of the M-dimensional subspace the corresponding flat ellipsoid is embedded in and each R_i^2 is an $M \times M$ PD matrix that defines the ellipsoids shape in that low rank cone. [67] proposes that the Karcher mean of the k M-dimensional subspaces U_i serves as a basis for the mean of the Λ_k where all flat ellipsoids are brought to by a minimal rotation. In that common subspace the problem reduces to the computation of the geometric mean of k rank M PD matrices. The implementation of their proposed mean for an arbitrary number of PSD matrices is outlined in Algorithm 1[1]. For more information about the rank preserving PSD mean and its properties we refer the reader to [67].

4.1.1. Convex combinations of models

LVQ models approximate the solution to non-convex problems and as such may converge to different local optima in independent training runs and the complexity of the problem. We expect that the model resulting from averaging over models from different local optima might exhibit inferior performance compared to its original contributors. Therefore we investigate convex combinations of Matrix LVQ models empirically and propose a clustering strategy to distinguish models to build local averages. For the prototypes of the models the Karcher mean, Eq. (14), can be generalized to a weighted mean or convex combination:

$$\begin{aligned} \widehat{\boldsymbol{w}}_c = &\arg\min_{\boldsymbol{w}\in\mathcal{M}} \sum_{i=1}^{k} \alpha_i d_{\text{geod}}(\boldsymbol{w}_{\{c,i\}}, \boldsymbol{w})^2 \\ &\text{with } c \in \{1, \dots, C\},\ \alpha_i \geq 0 \text{ and } \sum_i \alpha_i = 1 \ . \end{aligned} \tag{16}$$

[1] We provide the Matlab code at https://github.com/kbunte/geodesicLVQ_toolbox

Algorithm 1 Computation of geometric rank preserving PSD mean

procedure PSDmean$(\{\Lambda_i\}_{i=1}^k)$
  for $i = 1 \to k$ **do**
    compute eigenvalue decomposition $\Lambda_i = U_i R_i^2 U_i^\top$
  compute an orthonormal basis V on the Stiefel manifold $St(M, D)$ of the Karcher mean of the k subspaces U_i [a]
  for $i = 1 \to k$ **do**
    compute two orthogonal matrices O_i and O_i^V by SVD of $U_i^\top V$ [b]
    compute bases $Y_i = U_i O_i$
    compute bases $V_i = V O_i^V$
    with $\Psi_i^2 = Y_i^\top \Lambda_i Y_i$ the ellipsoid of Λ_i rotated to the mean subspace is $V_i \Psi_i^2 V_i^\top$
    express the ellipsoids in a common basis V: $T_i^2 = V^\top V_i \Psi_i^2 V_i^\top V$
  compute the ando mean $\overline{A}(T_1^2, \ldots, T_k^2)$ in the low-rank cone [c]
  return the geometric mean $\overline{\Lambda} = V \overline{A}(T_1^2, \ldots, T_k^2) V^\top$

[a] The Karcher mean of a set of M-dimensional subspaces of $\mathbb{R}^D$ on Grassmann manifold $Gr(M, D)$ is unique in a geodesic ball of radius less than $\pi/(4\sqrt{2})$ [75] and can be found by minimal rotation, as provided in the SuMMET package [76].

[b] These bases remove ambiguity in the definition of the PSD mean choosing particular bases Y_i of the fibers $U_i O(M)$ and bases V_i of the mean subspace fiber $V O(M)$ building the endpoints of the geodesic in the Grassmann manifold [67].

[c] Methods are proposed in [66; 77; 78] and we used the mmtoolbox implementation by the latter.

To the best of our knowledge an analytical solution for the weighted mean does not exist and several iterative strategies were proposed [69–73]. Two fast iterative solutions exhibiting linear and quadratic convergence for spheres can be found in [74].

[67] provided an analytical solution for the weighted average of two positive semi-definite matrices Λ_1 and $\Lambda_2 \in S^+(M, D)$, which can be summarized as follows. It bases on the same decomposition as stated in Eq. (15), i.e. $\Lambda_1 = U_1 R_1^2 U_1^\top$ and $\Lambda_2 = U_2 R_2^2 U_2^\top$ defined up to an orthogonal transformation $O \in O(M)$[2] and hence $A_i = U_i R_i^2 U_i^\top = U_i O_i (O_i^\top R_i^2 O_i) O_i^\top U_i^\top$. The equivalence classes $U_i O(M)$, called fibers, denote all bases that correspond to the same M-dimensional subspace $U_i U_i^\top$. While the orthogonal transformations do not affect the Grassmann[3] mean of subspaces they do effect the Ando mean of the low-rank PD matrices $\overline{A}(R_1^2, R_2^2) \neq \overline{A}(R_1^2, O^\top R_2^2 O)$ which causes the problems with the definition of a geometric mean. To deal with the ambiguity [67] proposed to compute particular representatives $Y_1 = U_1 O_1$ and $Y_2 = U_2 O_2$ as bases of the fibers, obtained by SVD of $U_1^\top U_2 = O_1 (\cos \Sigma) O_2^\top$ using the matrix cosine. These two bases correspond

[2] $O(M)$ denotes the general orthogonal group in dimension M

[3] Grassmann $Gr(M, D)$ denotes the space of all M-dimensional linear projectors in $\mathbb{R}^D$

to the endpoints of the geodesic in the Grassman manifold that minimize the distance between two fibers in the Stiefel manifold $St(M, D)$. These are than used to define a geodesic between Y_1 and Y_2 containing the convex combinations or t-weighted mean

$$Y(t) = Y_1 \cos \Sigma t + X \sin \Sigma t \quad \text{with } t \in [0, 1] \ , \tag{17}$$

where Σ is the diagonal matrix containing all principal angles and $X = (Y_2 - Y_1 \cos \Sigma)(\sin \Sigma)^{-1}$. Note that the half-way point $Y(0.5)$ is the Riemannian mean of Y_1 and $Y2$. Than the representative PD matrices for the M-dimensional ellipsoids in the low rank cone in the corresponding subspaces are given by $\Psi_i = Y_i^\top \Lambda_i Y_i$. Following [79] the convex combination (or t-weighted mean denoted by $\#_t$) of these two PD matrices is computed as

$$\Psi_1 \#_t \Psi_2 = \Psi_1^{1/2} \left(\Psi_1^{-1/2} \Psi_2 \Psi_1^{-1/2} \right)^t \Psi_1^{1/2} \tag{18}$$

Finally, having all the necessary ingredients, the convex combination of the SDM matrices Λ_1 and Λ_2 is computed by the t-weighted mean [67]:

$$\Lambda(t) = Y(t)(\Psi_1 \#_t \Psi_2) Y(t)^\top \ . \tag{19}$$

5. Clustering of Matrix LVQ models

In order to avoid averaging across local optima we propose a clustering strategy based on the Grassmann distance between the bases of the fibers U_i from the decomposition of the metric tensors Λ, see Eq. (15) and the text above (17). The Grassmann distance $d_{Gr}(U_i, U_j) = \|\mathbf{\Sigma}\|_2$ is computed using the principal angles $[\Sigma_1, \ldots, \Sigma_m]$, which are collected in the diagonal matrix Σ obtained by SVD of the product of the subspaces $U_i^\top U_j = O_i(\cos \Sigma) O_j^\top$. In case of localized class-wise metric tensors $\Lambda^c = \Omega^{c\top} \Omega^c$ we compute the Grassmann distance for each of the c projectors and use the average distance for clustering. We employ agglomerative hierarchical clustering using Ward Linkage on the pairwise Grassman distances and extract cluster memberships varying the numbers of clusters. Afterwards we compute the geodesic average model using only members of the same cluster and compute the macro averaged accuracy (MAA)[4] on the training set to select the best clustering. Of course different cluster methods could be used as well, such as for example variations of Grassman k-Means [80–82]. Furthermore, the Matlab ManOpt toolbox[5] provides a rich collection of algorithms for a variety of manifold optimization problems. However, we decided to use hierarchical clustering, since we have typically a comparable low number of models, such that the squared complexity with the number of instances does not state a problem and it avoids further introduction of local optima as is expected using k-Means or Gaussian

[4] mean of the classwise accuracies

[5] http://www.manopt.org

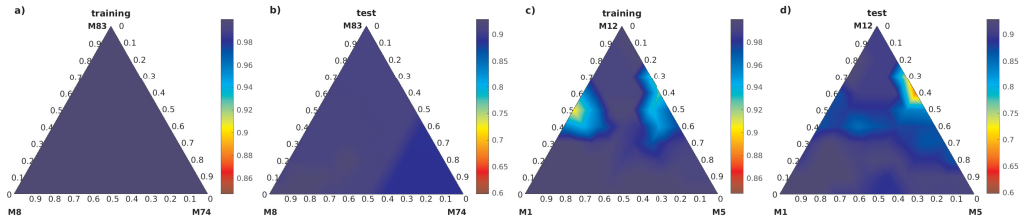


Figure 2: Visualizations of macro averaged accuracies (MAA) in training and test of the convex hull build by three $PLVQ^A$ models from the same (a,b) and three different clusters (c,d)

Mixture Model approaches. Furthermore, the cluster memberships for different numbers of clusters can be easily extracted without the need of re-running the method. Fig. 2 shows the MAA of the convex hull build by three probabilistic LVQ^A models trained on the exact same dataset (that of GCMS on the urinary metabolites, explained in subsec. 7.1) with rank M set to three. The first two panels depict the training and test set performances of the closest models within the same cluster, while the latter 2 panels show the performance of models taken from three different clusters. It can be seen that the convex combination of metric tensors from different clusters can lead to inferior performance, while it can improve using models from the same cluster. Therefore, we propose to extract 2-k clusters, compute the average model of each and look at an elbow in the training performance.

6. Synthetic datasets and experiments

In this section we perform two synthetic experiments to show the influence of the hyper-parameter Θ of the monotone function g_Θ in (8) and to investigate the influence of missingness when the ground truth is known.

6.1. Influence of Θ on classifier confidence

We created a three-dimensional synthetic dataset, each sample of which lies on the surface of a sphere. This toy dataset contained 85000 samples which were distributed into three classes in the proportion of 2:1:1, as shown in the Mollweide projection of this dataset (Fig. 3).

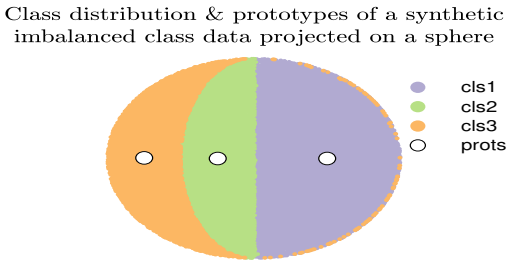


Figure 3: Mollweide projection of a 3D dataset for investigation of the effect of the Θ value

Since we are interested in studying the effect of Θ alone on the region of significant influence and area of regions of uncertainty, we fixed the Ω and the $\boldsymbol{w}^c$ with $c = 1, \ldots, C$ and $C = 3$ of a trained Probabilistic Angle LVQ model and varied only the value of Θ. In Fig. 4 each column corresponds to a value of $\Theta \in \{0.1, 1, 2, 5, 10, 50, 100\}$ and each row depicts the regions of probability for

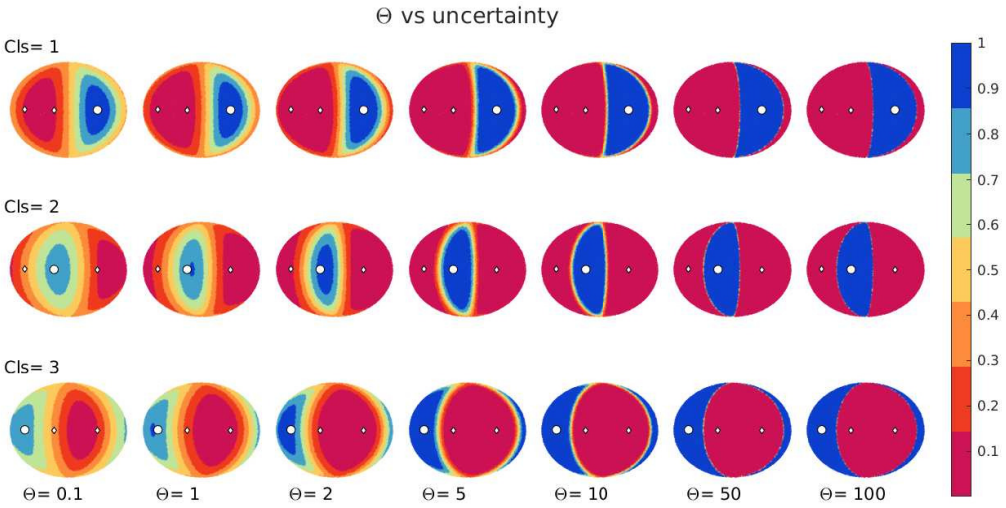


Figure 4: Effect of the value of Θ in classification uncertainty. The **o**-marker represents the prototype of the class highlighted in each row, namely class (Cls) 1, 2 and 3.

class 1, 2 and 3, respectively. In each sub-figure the Mollweide projection of the samples of the toy dataset are coloured according to the confidence of the model in assigning that sample the label of the class whose prototype is highlighted (big white circle). The heatmaps illustrate how with increasing Θ the regions of uncertainty become narrower, resulting in crisper decisions. Since we aimed for non-crisp decisions we set $\Theta < 20$ in our experiments on the real-world datasets.

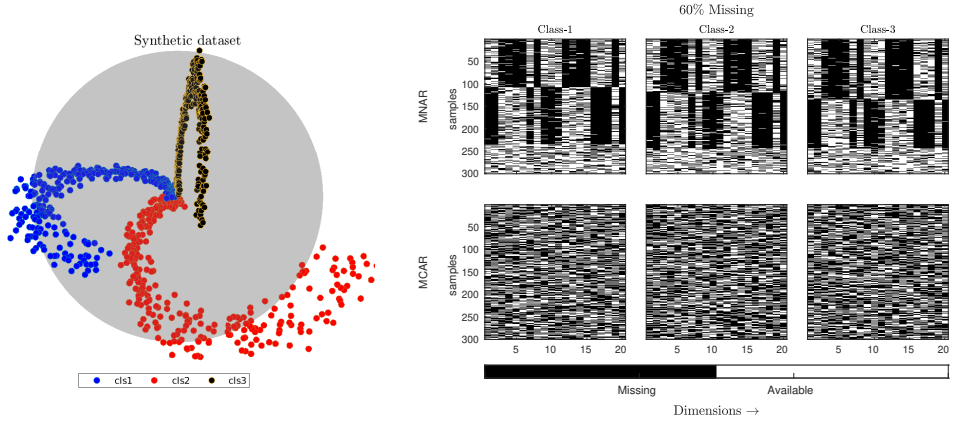


Figure 5: Synthetic data: plot of the 3 informative dimensions (left) and class-wise heatmaps of subject-wise average 60% missingness of type MNAR and MCAR (right).

6.2. Impact of training dataset size, type and amount of missingness

We modelled a synthetic dataset[6] to simulate the influence of the limited availability of training data, given the different type and amounts of missingness, as often encountered in biomedical data analysis. The synthetic dataset χ_{syn} is created with three informative dimensions in which three classes are arranged on two-dimensional manifold arcs bending in 3D and overlapping with their narrow parts in the center of a sphere, as visualized in Fig. 5. Similar to our real-world biomedical dataset the absolute values are not very informative in this arrangement. To increase the complexity we augmented it with four nonlinear transformations of the original three informative dimensions and five dimensions of uniform random noise, resulting in 20 dimensions in total. The non-linear copies were created by taking the (i) base 10 logarithmic transform, and the exponential transforms (ii) e^{χ}, (iii) χ^3 and (iv) χ^5. The dataset is successively treated with increasing amount of missingness, starting from 10% up to 60% in steps of 10%. We considered both the structured missingness mechanism such as the MNAR, as well as the more simplistic MCAR. In particular, the dataset is divided into three groups of group proportions $0.4 : 0.4 : 0.2$, to simulate a dataset that is built from three different laboratories and studies, as described hereafter. They are represented by an additional *"group id"* $\in [1, 2, 3]$. The first two groups of data (studies) measure a few, potentially mutually exclusive, features and the third group measures all features of the 20-dimensions of synthetic data. In case the identity of the group is known or observed (e.g. we know which lab the data items come from), we would have the MAR mechanism. If this is not the case and *"group id"* is unobserved (e.g. the data items from different labs were merged into

[6]Publicly available in https://git.lwp.rug.nl/cs.projects/angleLVQtoolbox.git

a single data set without the lab information), the missingness is of type MNAR. However, with passage of time each of the first two labs started measuring a few more dimensions than they initially used to, and thus we have a time and study-dependent (systematic) missingness. This scenario bears some similarity to our dataset described in subsec. 7.1. In reality randomly missing samples can exist in addition to systematically missing ones. Hence we added 12-15% of the target total missingness as random missingness. The right panel in Fig. 5 shows the most complicated case with 60% MNAR. We generated 300 samples per class (total of 900 samples) for training and validation, and an independent test set consisting of 30,072 samples.

To study the effect of the training set size on the generalization performance of the classifiers under the varying amounts of missingness, we successively reduced the amount of data from 80% to 20% of the original 900 samples in steps of 10%. Next, using 10-fold cross-validation (CV), we compare several strategies for classification in the presence of missing data in the synthetic datasets explained above. The first strategy bases on generative modeling, namely applying PPCA [33] on the data with missingness, followed by classification by LDA, as proposed in [32]. The algorithm is abbreviated by LDA^Q where Q denotes the latent dimension for PPCA. PPCA performed on the full training sets suggests an intrinsic dimensionality of 10 for each percentage of missingness. For classifiers which cannot handle missing data in their original formulation, such as Random Forest (RF) and KNN, we apply the PMM strategy of MICE [23; 83] on the training set to generate 10 imputed training sets and imputation models. These models then generate the corresponding imputed validation sets and imputed hold-out test set[7].

For the KNN classifier we varied number of nearest neighbours k and type of distance used, namely Euclidean and Mahalanobis, and abbreviate the method with $iKNN^{E_k}$ and $iKNN^{M_k}$ respectively. [84] suggests that the value of k should be chosen as the square root of the number of training instances. However, since we varied the size of the training set and simultaneously wanted to eliminate the effect of different values of this hyperparameter for the different sizes of the training set, we selected the upper limit of $k \approx \sqrt{162} \approx 12$ for all the sets according to the smallest set being 20% of the original samples. For RF we selected 150 DTs, which is large enough for a strong ensemble classifier and still smaller than the smallest training set. For PBCs we compare the original Euclidean distance based classifier GMLVQ with rank M on the imputed data, abbreviated by $iLVQ^{E_M}$ and the NaNLVQ capable of handling missing values, accordingly referred to as LVQ^{E_M}. The geodesic Angle LVQ extension (LVQ^{A_M}) is applied

[7] A recent out-of-sample extension for MICE called *mice.reuse* is available at https://github.com/prockenschaub/Misc/tree/master/R/mice.reuse

Table 1: Selected average training T^z_{fN} and hold-out test HO^z_{fN} errors for fraction $f \in [1, 0.2]$ of N original training samples, containing $z\%$ of missingness of type MNAR (on an average).

Classifier	$T^{0\%}_N$	$HO^{0\%}_N$	$HO^{0\%}_{0.2N}$	$HO^{30\%}_N$	$HO^{30\%}_{0.2N}$	$HO^{60\%}_N$	$HO^{60\%}_{0.2N}$
$ikNN^{E_{12}}$	.06 (.01)	0.15 (.01)	0.23 (.05)	0.22 (.03)	0.28 (.04)	0.41 (.01)	0.45 (.02)
$ikNN^{M_{12}}$	.02 (0)	0.08 (.01)	0.23 (.07)	0.23 (.04)	0.34 (.05)	0.45 (.01)	0.52 (.02)
$ikNN^{E_5}$	.05 (.01)	0.17 (.01)	0.26 (.05)	0.26 (.03)	0.30 (.04)	0.43 (.01)	0.47 (.02)
$ikNN^{M_5}$	.02 (0)	0.12 (.01)	0.25 (.06)	0.28 (.04)	0.36 (.05)	0.48 (.01)	0.53 (.02)
iRF_{150}	.00 (0)	0.01 (0)	0.02 (.01)	0.06 (.01)	0.08 (.02)	**0.25**(**.01**)	**0.30**(**.01**)
$iLVQ^{E_{10}}$	.02 (0)	0.02 (0)	0.07 (.04)	0.15 (.02)	0.21 (.04)	0.36 (.01)	0.43 (.03)
$iLVQ^{A_{10}}$	.00 (0)	0.01 (0)	0.08 (.05)	0.14 (.02)	0.20 (.04)	0.35 (.01)	0.41 (.03)
$iLVQ^{E_{20}}$	.02 (.01)	0.02 (.01)	0.07 (.03)	0.15 (.02)	0.21 (.04)	0.36 (.01)	0.43 (.03)
$iLVQ^{A_{20}}$	.00 (0)	0.01 (0)	0.08 (.05)	0.14 (.02)	0.20 (.04)	0.35 (.02)	0.42 (.04)
$LDA^{Q_{10}}$	.01 (.01)	0.17 (.03)	0.26 (.07)	0.25 (.03)	0.30 (.05)	0.38 (.03)	0.40 (.03)
$LVQ^{E_{20}}$	.02 (.01)	0.02 (.01)	0.07 (.03)	0.15 (.01)	0.21 (.04)	**0.30**(**.01**)	**0.35**(**.03**)
$LVQ^{A_{20}}$	.00 (0)	0.01 (.01)	0.07 (.05)	0.14 (.02)	0.20 (.02)	**0.27**(**.01**)	**0.35**(**.05**)
$LVQ^{E_{10}}$	.02 (0)	0.02 (0)	0.07 (.04)	0.15 (.01)	0.23 (.04)	0.31 (.01)	0.37 (.03)
$LVQ^{A_{10}}$	.00 (0)	0.01 (0)	0.08 (.05)	0.14 (.02)	0.21 (.04)	**0.27**(**0**)	**0.35**(**.04**)
$PLVQ^{A^1_{10}}$	.00 (0)	0.01 (0)	0.01 (0)	0.15 (.03)	0.16 (.03)	**0.27**(**.01**)	**0.28**(**.02**)
$LVQ^{LA_{10}}$	.01 (0)	0.05 (.03)	0.13 (.07)	0.13 (.02)	0.23 (.05)	0.24 (.02)	0.36 (.04)

both on the original and the imputed data ($iLVQ^{A_M}$) to show the influence of the imputation on the models. The novel probabilistic ALVQ is abbreviated by $PLVQ^{A^\Theta_M}$ in the following, and the hyperparameter $\Theta = 1$. Additionally we set the rank $M = 10$ for direct comparison with $LDA^{Q=10}$. The PBCs are repeated 5 times with random initialization on each training set.

Tab. 1 reports the performance in terms of classification error (and standard deviation) averaged over the 10 folds CV, when applied on the datasets with MNAR values. The classifier names are abbreviated as introduced before together with the main hyperparameters shown in the subscript and superscript. Prefix i denotes that the classifier is trained and tested on the imputed datasets. In the column names, T^z_{fN} refers to the training error and HO^z_{fN} the corresponding hold-out test error, where f indicates the fraction of the original number of samples N used for training, and z marks the average percentage of missingness per sample. Tab. 1 shows that RF exhibits the lowest error in the hold-out test test. However we also observed that RF suffers significant overfitting. This table further indicates that throughout the experimental settings (variation of amount of missingness and available data for training), the performance of $LVQ^{A_{10}}$ is

Table 2: Selected average training T^z_{fN} and hold-out test HO^z_{fN} errors for fraction $f \in [1, 0.2]$ of N original training samples, containing $z\%$ of missingness of type MCAR (on an average).

Classifier	$T^{0\%}_N$	$HO^{0\%}_N$	$HO^{0\%}_{0.2N}$	$HO^{30\%}_N$	$HO^{30\%}_{0.2N}$	$HO^{60\%}_N$	$HO^{60\%}_{0.2N}$
$ikNN^{E_{12}}$	.06 (.01)	0.15 (.01)	0.23 (.05)	0.23 (.01)	0.23 (.01)	0.38 (.01)	0.38 (.01)
iRF_{150}	.00 (0)	0.01 (0)	0.02 (.01)	0.06 (0)	0.06 (0)	0.23 (.01)	**0.23**(**.01**)
$LDA^{Q_{10}}$	.10 (.01)	0.17 (.03)	0.26 (.07)	0.21 (.03)	0.22 (.03)	0.33 (.03)	0.33 (.03)
$LVQ^{E_{10}}$	.02 (0)	0.02 (0)	0.07 (.04)	0.14 (.01)	0.16 (.02)	0.28 (.01)	0.28 (.02)
$LVQ^{A_{10}}$	.00 (0)	0.01 (.01)	0.08 (.05)	0.13 (.01)	0.14 (.02)	0.28 (.02)	0.28 (.03)
$PLVQ^{A_{10}}$	.00 (0)	0.01 (0)	0.01 (0)	0.15 (.02)	0.16 (.03)	0.29 (.02)	0.29 (.02)
$LVQ^{LA_{10}}$	.01 (0)	0.05 (.03)	0.13 (.07)	0.15 (.02)	0.16 (.03)	0.27 (.02)	0.35 (.03)

more stable than $LVQ^{E_{10}}$ even for the lowest rank of Ω. With regards to the KNN, the choice of distance measure has a stronger effect than the choice of k for this data. Comparing LDA^Q and the LVQs, we find that the effect of the number of PCs is more pronounced in the former than the effect of the rank of Ω for the latter.

Furthermore, we investigate whether the superior performance by RF is due to ensembling. Therefore we train a system of 150 $LVQ^{A_{20}}$ on the exact same imputed subsets of training data that each of the 150 DTs of the RF had trained on, on the most difficult setting (60% MNAR and training set reduced to 20% of its original size). The mean generalization error from the system of $iLVQ^{A_{20}}$ is **0.39 (0.02)** and that from $LVQ^{A_{20}}$ is **0.32 (0.01)** against RF's **0.30 (0.01)**. This additionally confirms that imputation does adversely affect the performance of LVQ^A classifiers. Since ensembling compromises with the interpretability of a classifier we applied geodesic averaging to our classifier, which resulted in a generalization error of 0.31 (0.01), thus comparable to RF with 150 DTs trained on the exact same subset of training data, indicating that ensembling and averaging strategies are indeed beneficial.

Next we compare and discuss the performance of the classifiers on the aforementioned MCAR datasets. For each of the classifiers, only the most promising hyperparameter settings (based on the validation set performance) were applied. Hence, in the following experiments we omit imputation for algorithms that handle the missingness internally. We set both Q for LDA and rank(Ω) to the intrinsic dimensionality estimated by PPCA and eigenvalue decomposition (EVD): 10.

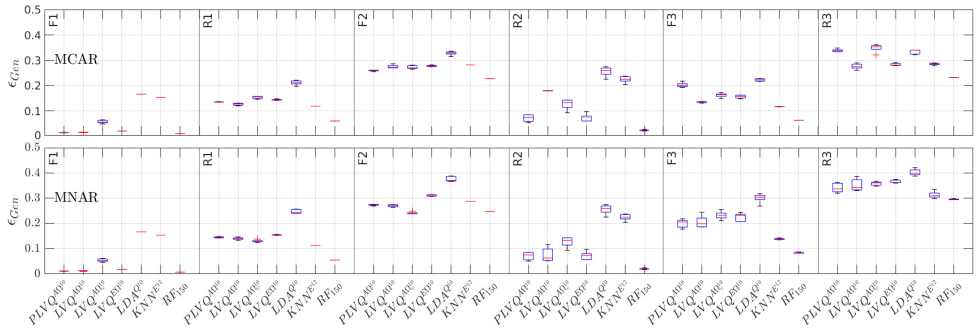


Figure 6: Classification error plots of 6 methods on the hold-out test set for MCAR and MNAR. $F \rightarrow$ full/original size, $R \rightarrow 20\%$ of F, suffixes 1 to 3 denote 0%, 30% and 60% missingness.

Tab. 1, 2, and the visual summary of the performances of $PLVQ^A$, LVQ^A, LVQ^E and LDA trained on unimputed data, and KNN and RF on the imputed dataset, in Fig. 6 illustrate the following: (i) there is prominent superiority in performance of LVQ^A against LVQ^E only when missingness is of type MNAR, while remaining similar for MCAR; (ii) KNN and LDA are less prone to error when the missingness type is MCAR; (iii) even though RF with 150 DTs have a slightly lower error rate than that of the LVQ classifiers, our investigation confirmed that it is because of ensembling. Since the motivation behind Tab. 1 was to show the difference in influence of the MCAR and MNAR type of missingness, we have not repeated the experiment with ensembling for this part.

7. Computer aided diagnosis of Inborn disorders of steroidogenesis

Inborn steroidogenic disorders (referred to henceforth as ISD) are genetic diseases affecting the Endocrine system which synthesizes hormones controlling bodily functions, such as blood pressure regulation, stress response, sex differentiation and puberty. ISDs can cause blockages in hormone production, leading to several forms of Congenital Adrenal Hyperplasia (CAH) and Differences in Sex Development (DSD) [85], which are rare but potentially life-threatening. However as identifying the ISDs involve measuring complex characteristic patterns of the steroidogenic biomarkers, computer-aided-diagnostic approaches are highly desirable for rapid diagnosis, and thereby for efficient treatment planning, selection, its delivery to save lives.

7.1. Urine steroid metabolite GC-MS dataset

The IMSR collected a unique but highly imbalanced dataset of 32 steroid metabolites (biomarkers) extracted from the urine samples of 829 healthy controls (HCs), and 178 patients with ISDs (ISD-1: 22, ISD-2: 12, ISD-3: 30, ISD-4: 26; ISD-5: 37; ISD-6:51), using Gas Chromatography–Mass Spectrometry

(GC-MS). This dataset was compiled across multiple analyses and studies, and over two decades, during which there were enhancement of both the clinicians' understanding of what constitute important biomarkers, and the GC-MS method itself. Consequently, certain biomarkers were only measured in subjects from later studies and resulted in systematic missingness in this dataset (see Fig. 7).

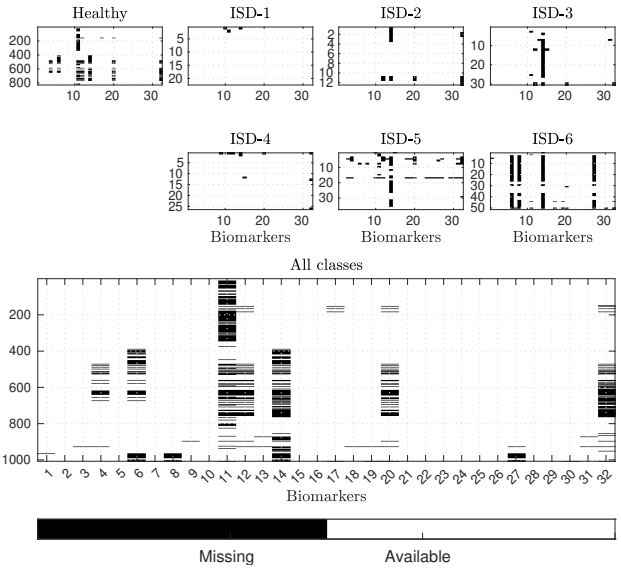


Figure 7: Heat-maps showing the presence of a combination of random and systematic missingness in each of the conditions of ISDs and HCs contained in the GC-MS dataset.

The third challenge is heterogeneous measurements, due to presence of large variations in biomarker profiles across the subjects, even within the same condition, due to individual physiological features, such as age, sex, and the subject's age dependent difference in urine sample collection methods. To combat this, we constructed pairwise combinations of these biomarkers, resulting in 496 ratios, following the proposal of [85–87] that using ratios of biomarkers.

7.2. Experimental setup and selecting hyperparameters.

Due to the limited number of samples for the rare disorders of steroidogenesis it is impossible to keep a hold-out test. Therefore, we validate the performance of the classifiers using 5-fold CV, using stratified sampling to preserve the class distribution. Following the outcome from the series of experiments performed on the synthetic datasets, we did not use imputation on the GC-MS dataset for any algorithm which can handle missing data implicitly. The data is preprocessed by z-score transform with the mean and standard deviation determined by each training set and consecutively used in the corresponding test set. PPCA

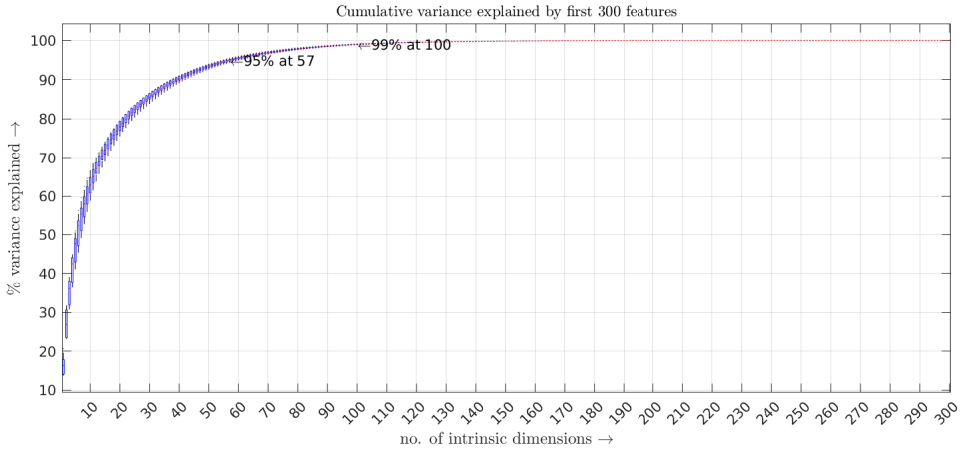


Figure 8: The cumulative variance shows that 57 intrinsic dimensions together explain 95%, and 100 intrinsic dimensions explain more than 99% of the variance of the dataset.

and EVD revealed $Q = 300$ latent dimensions (Fig. 8) and suggest that ≈ 57 and 100 latent dimensions can explain 95% and 99% of the variances of this dataset respectively. EVD of $\Lambda{=}\Omega^\top\Omega$ with $\Omega \in \mathbb{R}^{57\times \mathrm{D}}$ revealed an intrinsic dimensionality M=6. We additionally experimented with $\Omega \in \mathbb{R}^{3\times \mathrm{D}}$ that allows the visualization of the decision-boundaries (satisfying IIC-3). Experiments on each fold were repeated at least 5 times with random initialization of elements $\Omega_{ij} \in [-1, 1]$. Tab. 3 summarizes the experiments performed on the biomarker ratios of the GC-MS dataset. LVQ^E indicates GMLVQ using Euclidean distance, while $LVQ^{A^\beta_M}$ indicates ALVQ (using cosine-based dissimilarity), and M denotes the rank of metric tensor Λ and β the hyperparameter in ALVQ. For this dataset we also experimented with the more complex local variant, using local metric tensors, which are referred to in the tables as LVQ^{LA}. The probabilistic variant is abbreviated by $PLVQ^{A^\Theta_M}$. For block-A all the classifiers, except LDA, were applied on 10 imputed sets of training data per fold since LDA can intrinsically handle missingness. Thus while the variability in the rest of the classifiers in this block arise from both the imputed sets and the oversampling per iteration, the variation in LDA performance is solely due to oversampling. In block-B most experiments used geodesic SMOTE to tackle the class-imbalance in a comparable way. Alternatively a cost-weight matrix (Eq. (4) [30]) can be used, that can penalize certain classification errors more than others. We observed comparable results to the use of geodesic SMOTE and added one result with costs $\gamma_{cp} = 1/7$ for demonstration. The diagonal $\gamma_{c=p}$ and misclassified healthy controls cost is $2/3$, misclassifications of ISD-4 and ISD-5 for any other disease is $1/3$, and the highest penalty is induced by a patient being misclassified as healthy,

Table 3: Experiments on the GC-MS data. All experiments were performed on 5 folds CV using stratified sampling, and repeated random initializations per fold.

Algorithms	**Hyperparameters & experiment**
LDA^{Q_M}	PPCA for latent dimension of M=100 and 57, SMOTE (imbalance)
$iKNN^{E}_{\kappa}$	MICE (imputation), SMOTE (imbalance), and $k \in \{3, 5, 7\}$
iRF_t	MICE (imputation), SMOTE (imbalance), number of DTs $t \in \{7, 50, 100\}$
LVQ^{E_M}	SMOTE (imbalance), 1 prot/class, Rank(Λ) $M \in \{100, 57, 6, 3\}$
$LVQ^{A^{\beta}_M}$	Geodesic SMOTE (imbalance), 1 prot/class, Rank(Λ) $M \in \{100, 57, 6, 3\}$
$LVQ^{LA^{\beta}_M}$	Geodesic SMOTE (imbalance), 1 prot/class, Rank(Λ^L) $M \in \{100, 57, 6, 3\}$
$PLVQ^{A^{\theta}_M}$	Geodesic SMOTE (imbalance), 1 prot/class, Rank(Λ) $M \in \{100, 57, 6, 3\}$
$cPLVQ^{A^{\theta}_M}$	Cost weight matrix γ_{cp} (imbalance), 1 prot/class, Rank(Λ) $M \in \{100, 57, 6, 3\}$
$(cP)LVQ^{A^{\{\beta,\theta\}}_M}_{e\eta}$	Ensembling (majority vote) of $\eta \in \{5, 100\}$ iterations of $(cP)LVQ^{A^{\{\beta,\theta\}}_M}$ for each fold with Rank(Λ) $M \in \{100, 57, 6, 3\}$
$(cP)LVQ^{A^{\{\beta,\theta\}}_M}_{\#\eta_\upsilon}$	Geodesic average of υ clusters over $\eta = 100$ $(cP)LVQ^{A^{\{\beta,\theta\}}_M}$ models for each fold with Rank(Λ) $M \in \{6, 3\}$

setting the corresponding column off-diagonal elements to 1. We concentrate on the best hyperparameter settings (based on training performances only) of the newly presented intrinsically interpretable classifiers (LVQ^A, LVQ^{LA}, $PLVQ^A$ and $cPLVQ^A$) from block-B, to perform ensembling experiments (block-C). This is to ensure that they (i) constitute a fairer performance comparison to RF (which is an ensemble of η DTs), and (ii) enable easy comparison and interpretation of the corresponding average models presented in block-D (see Sec. 4). In block-D, the clustering strategy is employed before building the geodesic average model (see Sec. 5), abbreviated by $\#\eta_\upsilon$ in the subscript, where υ=1 (the default setting) indicates that all models were used in a single cluster. υ >1 indicates that the average was build in subsets of υ clusters. The latter is only beneficial if there are significantly different local optima found, which usually is encountered when the complexity of the model (global metric tensor, rank and/or number of prototypes) is too small for the classification problem, such as the rank 3 restriction that allows visualization of the decision-boundaries.

7.3. Performance comparison on the GC-MS data

In this section we present the results of the best hyperparameter settings (selected based on training performance) described in Tab. 3 for LDA, KNN and RF. For LVQ classifiers, their performances on both imputed and unimputed (original ratios) were compared, however as was seen for the synthetic dataset, imputation has an adverse effect and we do not show them. We performed grid-search to optimize the hyper-parameter settings of all methods with respect to

Table 4: GC-MS: mean validation performance (and standard deviation) across 5 folds. Evaluation measures include sensitivity (Sens, all conditions versus healthy), macro-averaged accuracy (MAA), and class-wise accuracy (cw-Acc). Blocks C and D show the performance of majority vote ensembling ($(cP)LVQ_{e\eta}$) and the fold-wise average model ($(cP)LVQ_{\#\eta}$).

Method	Sens.	MAA	Healthy	ISD-1	ISD-2	ISD-3	ISD-4	ISD-5	ISD-6
A: State-of-the-art traditional, interpretable, flexible ML models.									
iRF_{100}	**94.1**(**.03**)	**92.4**(**.03**)	99.8 (.00)	88.5 (.16)	96.7 (.07)	90.7 (.14)	89.3 (.17)	87.1 (.09)	94.8 (.06)
$iKNN_5^E$	86.2 (.06)	79.6 (.06)	98.0 (.01)	61.1 (.22)	82.7 (.30)	82.7 (.13)	84.9 (.13)	66.6 (.14)	80.9 (.14)
$LDA^{Q_{100}}$	87.7 (.03)	80.3 (.02)	97.7 (.01)	63.0 (.29)	76.7 (.22)	83.3 (.12)	72.7 (.23)	78.6 (.07)	90.2 (.00)
B: LVQ variants with different hyperparameter settings.									
LVQ^{E_6}	89.4 (.04)	86.2 (.06)	99.3 (.00)	78.0 (.14)	93.3 (.15)	83.3 (.12)	76.7 (.33)	78.6 (.15)	94.0 (.05)
$LVQ^{A_6^1}$	95.1 (.04)	91.3 (.03)	98.9 (.01)	84.0 (.22)	98.7 (.03)	93.3 (.10)	89.9 (.12)	78.6 (.17)	95.7 (.05)
$PLVQ^{A_3^{10}}$	95.8 (.02)	89.8 (.03)	98.2 (.01)	81.0 (.25)	95.3 (.10)	90.7 (.13)	86.9 (.14)	81.6 (.16)	94.5 (.06)
$PLVQ^{A_6^{15}}$	**96.6**(**.03**)	**91.8**(**.03**)	98.1 (.01)	85.0 (.24)	100 (.00)	91.3 (.12)	88.8 (.17)	80.5 (.13)	98.8 (.03)
$cPLVQ^{A_6^{15}}$	**97.3**(**.02**)	**91.1**(**.04**)	97.2 (.02)	84.8 (.20)	97.2 (.10)	92.4 (.09)	89.6 (.15)	81.6 (.14)	95.0 (.06)
$LVQ^{LA_3^1}$	95.2 (.03)	91.1 (.03)	99.0 (.01)	85.0 (.18)	97.3 (.06)	92.7 (.09)	88.0 (.18)	78.2 (.12)	97.3 (.05)
C: Ensembling selected LVQ variants.									
$LVQ_{e100}^{A_3^1}$	94.8 (.03)	91.7 (.02)	99.2 (.01)	81.0 (.21)	100 (.00)	96.7 (.07)	88.0 (.18)	78.9 (.14)	98.0 (.04)
$LVQ_{e100}^{A_6^1}$	94.3 (.03)	91.4 (.02)	99.0 (.01)	81.0 (.21)	100 (.00)	96.7 (.07)	88.0 (.18)	78.9 (.14)	96.2 (.05)
$PLVQ_{e100}^{A_3^{10}}$	**96.6**(**.03**)	**93.1**(**.03**)	98.9 (.01)	86.0 (.22)	100 (.00)	96.7 (.07)	88.0 (.18)	81.8 (.17)	100 (.00)
$PLVQ_{e100}^{A_6^{15}}$	**96.6**(**.03**)	**93.4**(**.02**)	98.7 (.01)	86.0 (.22)	100 (.00)	96.7 (.07)	88.0 (.18)	84.3 (.14)	100 (.00)
$cPLVQ_{e100}^{A_6^{15}}$	**97.2**(**.02**)	**92.7**(**.03**)	98.4 (.02)	91.0 (.12)	100 (.00)	93.3 (.09)	88.0 (.18)	81.8 (.17)	96.0 (.05)
$LVQ_{e100}^{LA_3^1}$	94.4 (.02)	90.2 (.02)	99.3 (.01)	76.0 (.25)	100 (.00)	93.3 (.09)	88.0 (.18)	78.9 (.14)	96.2 (.05)
D: Averaging selected LVQ variants.									
$LVQ_{\#100_5}^{A_3^1}$	94.4 (.02)	85.7 (.08)	94.9 (.09)	72.6 (.20)	85.3 (.23)	92.7 (.09)	83.7 (.15)	73.3 (.18)	97.3 (.04)
$LVQ_{\#100_5}^{A_6^1}$	94.6 (.02)	91.4 (.01)	99.0 (.01)	78.5 (.22)	100 (.00)	96.7 (.07)	88.0 (.18)	81.6 (.12)	96.2 (.05)
$PLVQ_{\#100_4}^{A_3^{10}}$	**96.5**(**.01**)	89.0 (.03)	98.1 (.01)	77.8 (.24)	92.5 (.07)	92.5 (.07)	86.0 (.22)	81.0 (.14)	95.5 (.04)
$PLVQ_{\#100_1}^{A_6^{15}}$	**96.6**(**.02**)	**92.6**(**.02**)	98.4 (.01)	86.4 (.20)	99.9 (.00)	93.8 (.08)	88.0 (.17)	82.8 (.14)	99.3 (.00)
$cPLVQ_{\#100_1}^{A_6^{15}}$	**97.8**(**.01**)	**92.9**(**.04**)	98.1 (.01)	91.0 (.12)	100 (.00)	93.3 (.09)	88.0 (.18)	81.8 (.17)	98.0 (.04)
$LVQ_{\#100_1}^{LA_3^1}$	95.5 (.01)	91.2 (.02)	99.3 (.01)	81.0 (.21)	100 (.00)	93.3 (.09)	88.0 (.18)	78.9 (.14)	98.2 (.04)

the training data. For KNN the performance corresponding to Euclidean distance with κ=5 nearest neighbours is reported. For the $PLVQ^{A_M^\Theta}$, Θ=10 and and 15 were found to be good choices for for rank M=3 and 6 respectively. We present the generalization performance from those experimental settings which had best training performance, were easy to interpret, and helped in considerable knowledge gain by the medical community. Tab. 4 shows the most interesting

selection of performances of Angle LVQ (global and local), the newly introduced probabilistic variant $PLVQ^{A^{\Theta}_{M}}$, RF (with 100 trees), imputed KNN with $\kappa = 5$, LDA with latent dimension $Q = 100$, and the original matrix LVQ^{E_M} (imputed and NaNLVQ). Since the class-wise accuracy of the healthy condition is the same as the specificity we only report the former. Tab. 4 shows that the imputed RF is superior to LDA or imputed KNN. Furthermore, the experiments demonstrate that the use of the angular dissimilarity in the LVQ models is beneficial for this dataset. Note, that the ALVQ models are also fairly robust with respect to the hyper-parameter setting, with the exception of the rank 3 models that trade some of the performance for additional interpretability by visualization. Interestingly, $PLVQ^{A^{\Theta}_{M}}$ even with ranks 3 and 6 of Λ, and $LVQ^{LA^{\beta}_{M}}$ with rank 3 local metric tensors Λ^L show better performance than their original formulation and have comparable performance to RF. Following an ensembling strategy with 100 LVQ models leads to a fair comparison. Especially the $PLVQ^{A}$ models achieve similar or superior performance and exhibit higher sensitivity and MAA when compared to RF, albeit with some loss of interpretability. However, the clustered average models in block-D, resolves this by harnessing the power of ensembling while preserving interpretability, as explained in Sec. 4. The overall best performance is achieved with an ensemble of $PLVQ$ models over 100 random initializations, and their corresponding average model in a single cluster. The cost weight matrix $cPLVQ$ is a viable and computationally cheaper alternative to oversampling as it can steer training with prior knowledge without generation of synthetic samples, while maintaining comparable performance.

7.4. Knowledge extraction from Angle LVQ models

7.4.1. Visualisation of decision-boundaries

The ALVQ variants, both probabilistic and deterministic, with Λ of rank 3 can be used to visualize the decision-boundaries between the conditions, the positions of the prototype of each class, and the subjects on a sphere, satisfying IIC-3. The rank-3 model is not complex enough for the GC-MS classification problem resulting in trade-off in performance and several local optima, and requiring clustering before averaging. For Fig. 9 we selected a model from $PLVQ^{A^{10}_{3}}_{\#\eta_\upsilon}$ cluster 4 which averaged over 32 constituent models of fold 1. We visualize the sphere in two dimensions using the Mollweide projection[8]. The reduction of the hypersphere of 496 dimensions to 3 dimensions for visualisation purpose slightly compromised with the sensitivity and class-wise accuracies. However, this illustration provides an effective visual explanation for collaborators of how $PLVQ^{A^{10}_{3}}_{\#100_4}$ performs classification on the hypersphere and highlights the position of subjects which lie close to decision-boundaries which may be challenging to accurately classify.

[8]Matlab code available at https://github.com/SrGh31/classificationSphereMollweide

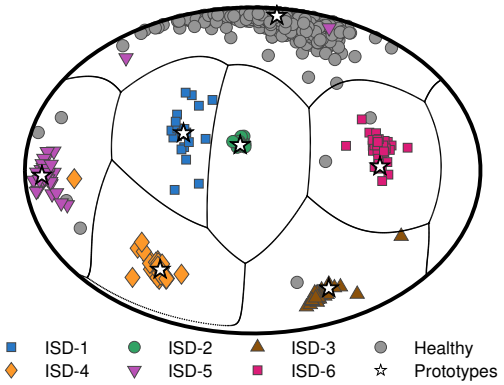


Figure 9: Mollweide projection of the decision-boundaries, prototypes and samples induced by one of the cluster models $PLVQ^{A_3^{10}}_{\#100_4}$ averaged over 32 individual models in fold-1.

7.4.2. Biomarker extraction

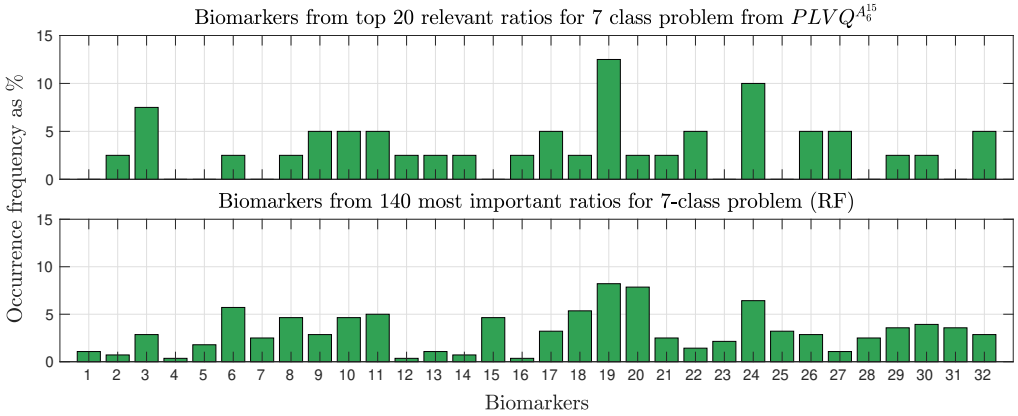


Figure 10: Biomarker relevance for ISD as occurrence frequency extracted from top 20 and 140 most relevant ratios with $PLVQ^A$ and RF respectively. Numbers 1-32 indicate the 32 biomarkers extracted from urine for investigation

Matrix LVQ models enable the extraction of feature relevance information from the given dataset (satisfying IIC-1). One can obtain similar information from RF (feature importance) using MDA strategy. Fig. 10 shows the feature relevance in the form of occurrence frequency of each biomarker measured in urine for the classification of ISD, extracted from the largest 20 ratios on diagonal of Λ in comparison with those extracted from 140 most important ratios of the RF (to cover 20 ratios $\times$ 7 classes). We reassuringly find that both the performance as well as the feature relevance and importance profiles obtained from $PLVQ^A$ and RF to be very similar.

While it is theoretically possible to extract condition specific information from RF, it is not as intuitive or straightforward a process as it is to extract similar information from all LVQ^A variants. Class-specific biomarker information is often vital for the clinicians' understanding and interpretation, but often difficult to obtain from classification models. However PBCs can also extract the magnitude of the disease-specific biomarkers measured, and the (disease) prototype profiles can be obtained by both the global and local versions of LVQ^A. Furthermore, the local version provide clinicians information about a unique "fingerprint" pattern of biomarker ratios (i.e. feature vectors) most specific for distinguishing each condition fulfilling IIC-2. The feature relevance information for each condition is easily extracted from the diagonal of the matrix LVQ variants with classwise local metric tensors Λ^L. However, both probabilistic and deterministic LVQ variants can provide the domain expert with far superior interpretability in this regard. Using the average models we can perform a descriptive analysis of the classification-terms (as detailed in subsubsec. 7.4.3).

7.4.3. Descriptive analysis of Probabilistic LVQ decisions

Any matrix LVQ model allows the analysis of its decision-making based on the classification-terms, which essentially serve similar purpose to that by the classification activation map (CAM) as explained in [3]. For the proposed variant $PLVQ^A$ (Eq. (8)) for example a classification-term is the product of the sample vector dimension $\boldsymbol{x}_{i,d1}$, the relevance matrix element $\Lambda_{d1,d2}$, and a prototype dimension $\boldsymbol{w}^c_{d2}$, together with additional factors or transformations dependent on the dissimilarity measure. A sample $\boldsymbol{x}_i$ is classified as class c if the sum of classification-terms (T^{ic}) including prototype $\boldsymbol{w}^c$ is larger than for any other prototype:

$$\hat{p}(c|\boldsymbol{x}_i) = \frac{g_\Theta\left(\sum_{F1=1,F2=1}^{D} T^{ic}_{F1,F2}\right)}{\sum_j^C g_\Theta\left(\sum_{F1=1,F2=1}^{D} T^{ic}_{F1,F2}\right)} \quad \text{with} \quad T^{ic}_{F1,F2} = \frac{x_{i,F1}\Lambda_{F1,F2}w^c_{F2}}{\|\boldsymbol{x}_i\|_\Lambda \|\boldsymbol{w}^c\|_\Lambda} \ .$$

While the generalization performance is demonstrated in the previous section we show here the decision-making statistics over the full dataset, and hence a descriptive analysis. Therefore, we build one model from the 5 fold $cPLVQ^{A^{15}_6}_{\#100_1}$ models, using the geodesic averaging strategy that represents the average statistics of the trained decision-making process across all folds. Extracting the feature-wise relevances from the diagonal of $\overline{\Lambda}$ and sorting in descending order reveals that 394 biomarker ratios (out of 496) already contain over 95% of the total relevance and equivalent accuracy. We remove the unimportant dimensions resulting in a reduced model for the following analysis, that only misclassifies 15 out of 1007 samples in total. Among the latter are 10 HCs, 1 ISD-1 patient missed as healthy, 2 ISD-3 patients missed as ISD-5 and 2 ISD-5 patients misclassified as ISD-4. The $\hat{p}(c|\boldsymbol{x}_i)$ provides the probability for sample i to belong to class c and we can see

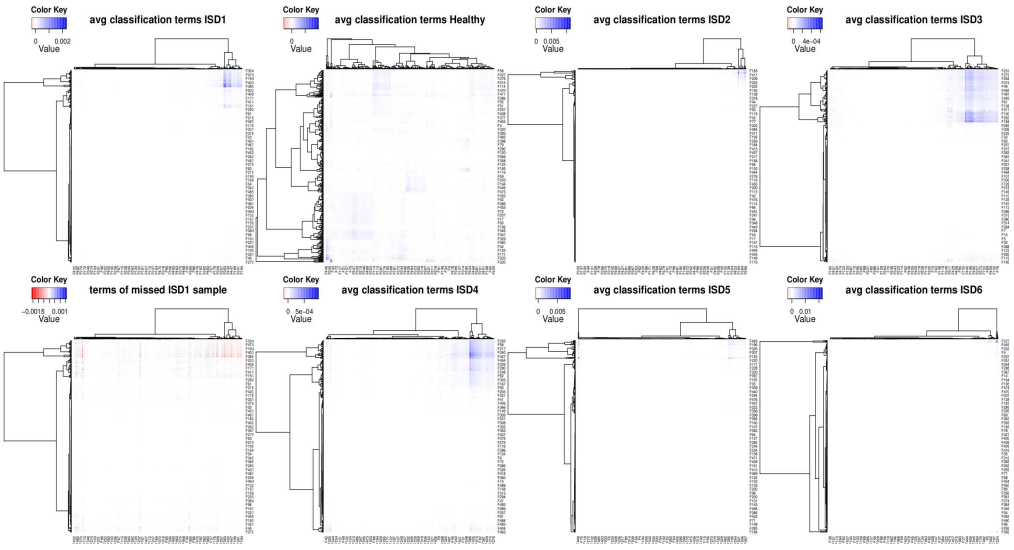


Figure 11: Biclusters of classification terms T^{ic} for the 394 metabolite ratio features F averaged for all samples of $\boldsymbol{x}_i$ with $y_i = c$ for every condition c. The terms of the missed ISD-1 sample (lower left panel) are sorted according to the condition bicluster (top left)

that the second most likely class is often the correct one. However, we are interested in the ratios and biomarkers and how much (on average) they contribute to the decision.

The matrix of classification-terms T^{ic} contain positive or negative entries indicating the correlation of $\boldsymbol{x}_i$ with the prototype $\boldsymbol{w}^c$ induced by the metric tensor Λ. Since the classification decision is based on the biggest sum over all $\arg\max_c(\sum_{F1,F2} T^{ic}_{F1,F2})$ the terms can be sorted. Fig. 11 shows biclusters of classification-terms T^{ic} averaged for all samples of $\boldsymbol{x}_i$ with $y_i = c$ for every condition c. The rows and columns are clustered using the agglomerative Ward2 cluster algorithm [88; 89], grouping similar entries simultaneously in rows and columns. Note that most of the classification decisions for each condition are only based on comparably few biomarker ratios as many terms are close to zero. In contract to HCs the patients show mostly a clear important block of pairwise ratios dominating the decision. The misclassified ISD-1 sample is shown in the lower left panel with the sorting adopted from its condition's average T^{ic} showing clearly that important ratios differ significantly from the respective prototype. Fig. 12 depicts the performance change dependent on the number of top ratios used in the model and the frequency of biomarkers for each class. The left panel shows the respective class specific sensitivity and specificity (as well as overall sensitivity and specificity in terms of healthy vs disease) achieved by the model being reduced to the top x ratios extracted from the sorted average classification-terms T^{ic} per

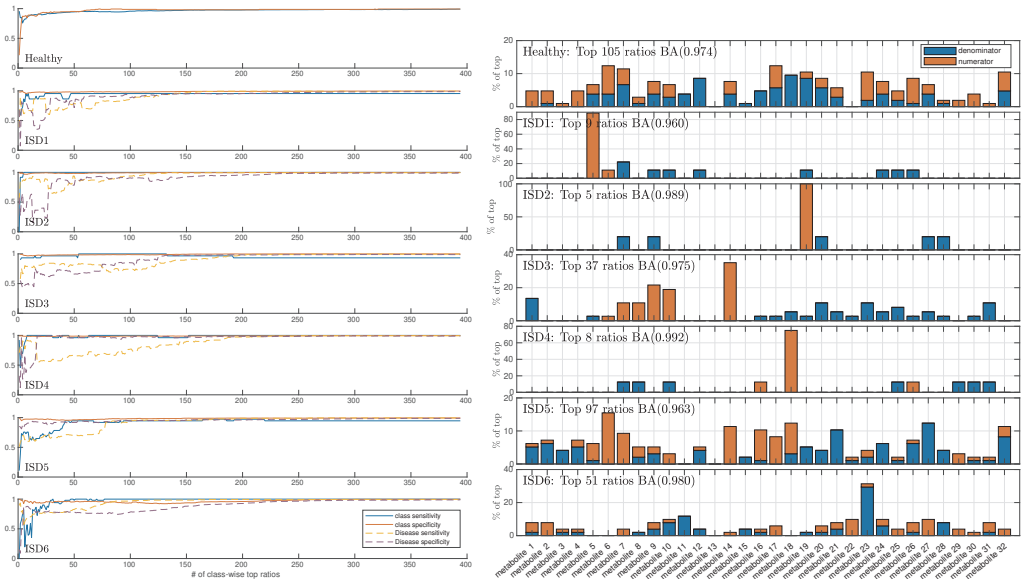


Figure 12: Descriptive analysis of ratio contribution (left) and metabolite frequency among top ratios (right) for the classification of the inborn disorders of steroidogenesis

class $y_i = c$. And the right panel depicts the frequency of biomarkers among the top ratios and their appearance in nominator or denominator, that indicates if it is over- or under-produced in the respective class. We additionally report the number of top ratios per class that achieve 98% of the balanced accuracy (BA, arithmetic mean of the sum of sensitivity and specificity of the class versus the other classes combined) within each panel. Notably, the classification decision for patients of some of the conditions, namely ISD-1, ISD-2, and ISD-4, is very accurate with over 0.95 BA and on average based on fewer than 10 biomarker ratios (9, 5 and 8 respectively). Within those top ratios an overwhelming majority of over 80% contain one specific biomarker in the numerator indicating an excess compared to HCs, which constitutes an interesting biomarker for each of these conditions. Conditions ISD-3, ISD-6, ISD-5, and HC are more heterogeneous and an increasing number of ratios and consequently biomarkers are necessary to distinguish them. In summary, the classifier exhibits excellent performance and readily provides ample insight into the contribution of each feature for the decision of each individual sample, as well as the feature statistics over all samples per class. This transparency constitutes an important characteristic for the purpose of medical education, potentially biomarker discovery and to gain trust for computer-aided-diagnosis with ML within the medical community.

8. Analysis of the UCI (Cleveland) heart disease dataset

8.1. Cleveland heart disease dataset

This dataset (details in [90]) contains 13 features from 164 HCs and 139 heart disease (HD) patients, belonging to 5 different sub-categories: HD1 (55), HD2 (36), HD3 (35) and HD4 (13). Of them, six subjects contain missing values which according to [90] were replaced by -9. Exploratory analysis showed that while there is a very good separation between HCs and HDs in binary classification, the multi-class problem differentiating between the 4 classes of HDs turn out to be remarkably difficult. In this study we investigated this dataset as a five-class problem, as suggested in [5; 31].

8.2. Experimental setup and hyperparameter selection

In this section we compare the most promising classifiers as demonstrated on the GC-MS dataset on the publicly available UCI HD dataset. As mentioned earlier, unlike many of the the existing classifiers the ALVQvariants can learn even in the presence of missing values, thus obviating the need for imputation (even with a pseudo value, such as -9 as suggested for this dataset [90]) or case-deletion. We follow the recommendation with the pseudo value for the RF training. Many previous publications report performances simplifying the five classes, Healthy and HD1-HD4, to the binary disease versus healthy problem, due to the difficulty. Since we were interested in investigating this dataset as a multi-class problem and the smallest minority class contained only 13 subjects, we use 5-fold stratified training-validation split for CV. We use z-score transformation in each fold using the mean and standard deviation of the corresponding training set. As before, we compare two strategies for handling class-imbalance: (a) the original SMOTE [14] or geodesic SMOTE for the ALVQ variants and (b) assigning user-defined variable costs of misclassification (Eq. (4)). For RF we used only (a). We ensured that all the minority classes in the training set were oversampled to contain the same number of samples as the majority class (HC) and based on line search chose $k = 3$ nearest neighbours for both SMOTE and geodesic SMOTE. For option (b) we set the cost-weight entries to 1 to only handle the imbalance. We varied the number of trees for RF_t from t=50-200 and set hyperparameter $\beta = 1$ for $\mathrm{LVQ}^{A^{\beta}_{M}}$ and and $\theta = 2$ for its newly introduced probabilistic counterpart $\mathrm{PLVQ}^{A^{\theta}_{M}}$ after grid-search on the training data. To be comparable to the RF, we trained 100 LVQ^{A} models in each fold and additionally reported the performance of their majority vote ensemble, abbreviated by (cP)LVQ$^{*}_{e100}$. We trained the 100 LVQ models with a full metric tensor rank of $M = 13$ in each fold and also built a one-cluster geodesic average model abbreviated by (cP)LVQ$^{*}_{\#100_1}$ for further analysis. Since the majority of the probabilistic model tensors exhibited a rank of 12 after training, and we need equal rank to build the average model, we limited the rank to 12 for all of them.

Table 5: UCI HD data: mean performance (std), in terms of Sensitivity (HC versus H1-4 combined), MAA and class-wise accuracies (cw-Acc), of RF_t with t trees and Angle LVQ models $(cP)LVQ^{A^{\beta,\theta}_M}$ with rank M, and c indicating the use of cost weights, P probabilistic cost-function (7) and the subscripts $e\eta$ and $\#\eta_\upsilon$ marking the ensemble results with majority vote and average across η models and υ clusters

Method	Sens	MAA	Healthy	HD1	HD2	HD3	HD4
RF_{100}	73.48 (.03)	31.82 (.04)	**86.37**(.03)	15.64 (.07)	19.36 (.10)	25.71 (.11)	12.00 (.12)
RF_{150}	74.45 (.02)	33.60 (.03)	86.25 (.03)	15.27 (.07)	23.14 (.10)	28.00 (.10)	15.33 (.12)
RF_{200}	74.63 (.02)	31.68 (.04)	86.12 (.02)	14.18 (.07)	19.36 (.09)	27.43 (.12)	11.33 (.10)
$PLVQ^{A^2_{12}}$	89.08 (.05)	50.23 (.04)	67.52 (.03)	26.67 (.09)	32.10 (.03)	51.34 (.15)	73.50 (.13)
$cPLVQ^{A^2_{12}}$	88.91 (.05)	51.34 (.06)	71.09 (.03)	25.71 (.09)	29.05 (.07)	52.20 (.15)	78.67 (.17)
$cLVQ^{A^1_{13}}$	83.26 (.07)	30.94 (.07)	63.97 (.15)	21.29 (.12)	21.47 (.14)	15.49 (.14)	32.50 (.22)
$PLVQ^{A^2_{12}}_{e100}$	88.33 (.08)	**58.55**(.08)	74.41 (.03)	32.73 (.14)	30.36 (.11)	**68.57**(.27)	**86.67**(.18)
$cPLVQ^{A^2_{12}}_{e100}$	87.59 (.08)	57.74 (.04)	75.04 (.04)	30.91 (.19)	**33.21**(.07)	62.86 (.26)	**86.67**(.18)
$cLVQ^{A^1_{13}}_{e100}$	82.69 (.09)	33.15 (.10)	82.35 (.06)	16.36 (.12)	21.79 (.18)	8.57 (.08)	36.67 (.41)
$PLVQ^{A^2_{12}}_{\#100_1}$	91.24 (.07)	50.07 (.08)	61.02 (.08)	29.09 (.15)	25.00 (.12)	48.57 (.28)	**86.67**(.18)
$cPLVQ^{A^2_{12}}_{\#100_1}$	**92.75**(.04)	50.95 (.10)	65.34 (.10)	**34.55**(.17)	28.21 (.11)	40.00 (.27)	**86.67**(.18)
$cLVQ^{A^1_{13}}_{\#100_1}$	84.78 (.09)	33.62 (.12)	76.23 (.05)	18.18 (.14)	21.79 (.18)	8.57 (.08)	43.33 (.43)

Tab. 5 summarizes the mean and standard deviations of the method performances measured in Sensitivity (HCs versus all diseases combined), MAA and cw-Acc (HCs and HD1-4). It can be seen immediately that the five class problem is challenging with a MAA not much more than 30% achieved by the RF. The two class problem of all heart disease versus healthy is easier, showing a sensitivity and specificity (cw-Acc of the HCs) of $\approx 74\%$ and $\approx 86\%$. Interestingly this dataset shows a clear difference between the two different cost-functions for the ALVQ. Ensembles of the version inspired by GLVQ (updating only the closest correct and wrong prototypes) exhibit only slightly better performance than RF, while the probabilistic version improves the sensitivity and MAA by more than 10%. This effect might be caused by the influence of all classes in every update due to the use of the parameterized soft-max. Both strategies to handle the imbalance, namely cost-weight matrix and geodesic SMOTE, demonstrate similar performance. The benefits of cost-weighting over oversampling include: (i) faster execution, since it operates on fewer samples; and (ii) provision for the user to indicate priorities for the classification. We observe that increase of accuracy for one class is usually accompanied with a decrease of accuracy of another class. RF exhibits the best accuracy for the HCs at the expense of disease-class

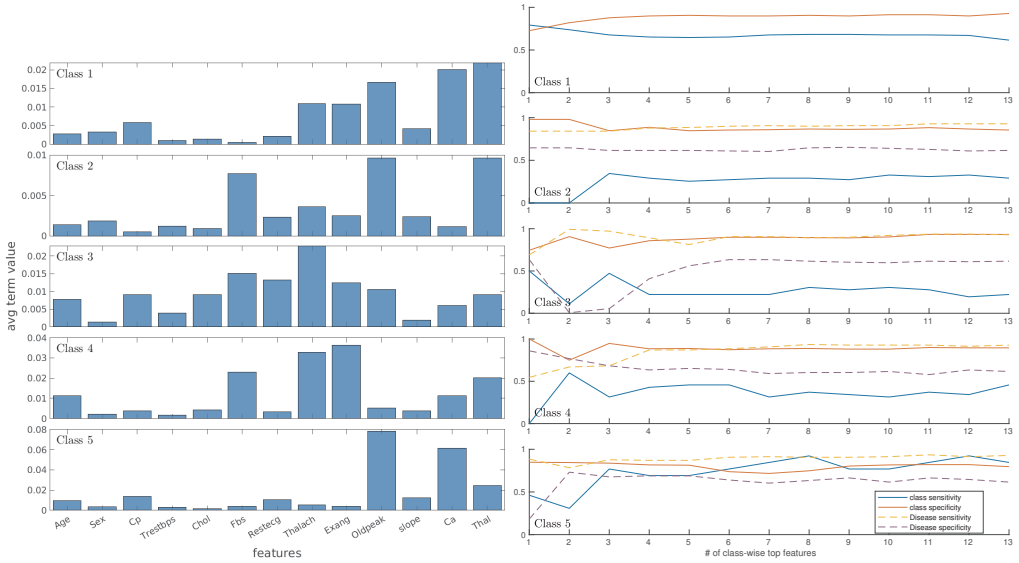


Figure 13: UCI HD: Class-wise feature importance for the classification of each class extracted from the terms (left) and the sensitivity and specificity considering HCs (Class 1) versus disease (Class 2-5) as well as the respective class versus all others combined (right).

accuracies. However, performance with respect to the classwise-accuracies of the disease-classes are clearly surpassed by the Angle LVQ variants, which in turn demonstrates a high sensitivity for the two-class problem. Similar to the previous section we can analyze the statistics of the contribution of the features to the decision of the trained classifiers for each class using the classification-terms of the average model across all folds of the $cPLVQ^{A^2_{12}}_{\#100_1}$ experiment. Fig. 13 illustrates the performance of each class versus all others combined, as well as Healthy versus disease, when including only the top features for the samples of the respective class in the right panel. Since we do not consider ratios we can show the classification-term contribution of the features for each class directly as shown in the left panel. Especially the HC (Class-1) and HD1 (Class-2) exhibit a very similar pattern of important features, which explains the difficulty to distinguish them. Fig. 13 shows that the "Oldpeak", referring to ST depression in the ECG signal, induced by exercise relative to rest, is an important feature to identify HC, HD1 and HD5 from the rest. Contrarily, "Thalach" which refers to the maximum heart rate achieved, is important to discriminate HD3 and HD4 from the rest. While RF can find the overall feature importance and there usually is overlapping agreement with the findings from prototype based methods, the classification-terms from LVQ models help in extraction of class-specific feature relevance as seen in Fig. 13.

9. Conclusion and Future work

This paper demonstrates the performance of ALVQ classifiers when facing heterogeneous measurements, imbalanced classes, limited data for training and systematically missing values. It introduces (2) a probabilistic cost-function that enables training despite the varying certainty in labels, and reflect upon the uncertainty of the classification, thus providing additional insight into the model, and (3) strategies to combine ensembling with interpretability, by computing the geodesic average of matrix LVQ models. This is preceded by a clustering strategy, if multiple local optima is detected. These serve as more *transparent alternatives to a traditional ensemble* and extends to other members of the adaptive metric family, such as the Euclidean LVQ variants and LMNN [43; 44; 91]. We provide our code and a detailed analysis and demonstrate the transparency of our framework and how knowledge is extracted from real-world medical datasets.

Our findings show that in the presence of heterogeneous measurements and systematic missingness the cosine-based adaptive dissimilarity measure (as used in Angle LVQ) appears more robust than the parameterized Euclidean distance. The paper further demonstrates the adverse impact of imputation on distance-based classification in synthetic experiments. The application of RF, the newly developed Probabilistic Angle LVQ, and the model-averaging strategy[9] will be presented in a forthcoming paper written for the medical community. Sec. 8 presents the same on publicly available dataset (subsec. 8.1) for reproducibility and verification. In summary, the proposed strategies show promising results for domains such as healthcare, that are plagued with challenges of imbalanced classes, missing and heterogeneous data, that require not only high performance but also demand algorithmic transparency, estimation of uncertainty, interpretability, and explainability.

Competing interests

No competing interest is declared.

Declaration of no involvement of generative AI in scientific writing

No part of this manuscript is generated from, or has any involvement of any large language model or generative AI model.

[9]Matlab code is publicly available at `https://github.com/kbunte/geodesicLVQ_toolbox`, on the biomedical datasets demonstrated that (i) Angle LVQ variants allow detailed insight into the influence of features on the classification of every sample, every class, the data as a whole and, in special cases, direct visualization of the data and decision-boundaries; and (ii) the geodesic-average of 100 LVQ-models has a comparable or superior performance to RF, while remaining interpretable and facilitating knowledge-extraction. While the knowledge-extracted from the GC-MS dataset (subsec. 7.1)

Author contributions statement

S.Ghosh: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing - Original Draft, Writing - Review & Editing, Visualization. **E.S. Baranowski:** Conceptualization, Methodology, Investigation, Resources, Data Curation, Writing - Review & Editing. **M.Biehl:** Conceptualization, Methodology, Validation, Investigation, Writing - Review & Editing. **W.Arlt:** Conceptualization, Methodology, Investigation, Validation, Resources, Data Curation, Supervision, Funding acquisition, Writing - Review & Editing. **P.Tino:** Conceptualization, Methodology, Validation, Investigation, Writing - Review & Editing. **K.Bunte:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing - Original Draft, Writing - Review & Editing, Visualization, Supervision, Funding acquisition, Project administration.

Ethics approval

Part of this manuscript contains the analysis of data from Human subjects/patients and ethics approval for data collection for research purpose has been provided to the data curators. The data was first de-identified by the data curators before making it available for model training. This paper does not contain or disclose any private information of the subjects.

Consent for publication

Not applicable: this research does not involve personal or private data, and publishing of this manuscript will not result in the disruption of any individual's privacy, security or safety.

Code availability

The MATLAB code for the ALVQ toolbox and average ensemble are available at https://github.com/kbunte/angleLVQtoolbox and https://github.com/kbunte/geodesicLVQ_toolbox. Code for the Mollweide projection and classification sphere (used for creating Fig. 9) is available at https://github.com/SrGh31/classificationSphereMollweide.

Data availability

The synthetic data set is available at https://github.com/SrGh31/syntheticALVQ.git. The Cleveland dataset is available from UCI repository. The GC-MS data is not publicly available and curated by IMSR.

Acknowledgments

We thank (1) the Rosalind Franklin fellowship, co-funded by the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement **600211**, and H2020-MSCA-IF-2014, project ID **659104**, (2) the Center for Information Technology of the University of Groningen for their support and for providing access to the Peregrine High Performance Computing Cluster, (3) the LWP team at University of Groningen who made it possible to work from home during the pandemic, (4) former BSc. student from Short programming project, Ethan Waterink, for his contributions in automating the Mollweide projections.

References

[1] P. J. García-Laencina, J.-L. Sancho-Gómez, A. R. Figueiras-Vidal, Pattern classification with missing data: a review, Neural Computing and Applications 19 (2) (2010) 263–282. doi:10.1007/s00521-009-0295-6.
URL https://doi.org/10.1007/s00521-009-0295-6

[2] A. Holzinger, C. Biemann, C. S. Pattichis, D. B. Kell, What do we need to build explainable ai systems for the medical domain?, arXiv preprint arXiv:1712.09923 (2017). doi:10.48550/arXiv.1712.09923.

[3] E. Tjoa, C. Guan, A survey on explainable artificial intelligence (xai): Toward medical xai, IEEE Transactions on Neural Networks and Learning Systems 32 (11) (2021) 4793—4813. doi:10.1109/tnnls.2020.3027314.

[4] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, Information Fusion 58 (2020) 82–115. doi:10.1016/j.inffus.2019.12.012.

[5] S. Ghosh, Intrinsically interpretable machine learning in computer aided diagnosis, Ph.D. thesis, University of Groningen (2021). doi:10.33612/diss.175627883.

[6] L. v. d. Maaten, G. Hinton, Visualizing data using t-sne, Journal of Machine Learning Research 9 (2008) 2579–2605.

[7] A. Schulz, F. Hinder, B. Hammer, Deepview: Visualizing classification boundaries of deep neural networks as scatter plots using discriminative dimensionality reduction, in: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI, 2020, pp. 2305–2311. doi:10.24963/ijcai.2020/319.

[8] L. Pfannschmidt, C. Göpfert, U. Neumann, D. Heider, B. Hammer, Fri-feature relevance intervals for interpretable and interactive data exploration, in: 2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), IEEE, 2019, pp. 1–10. doi:10.1109/CIBCB.2019.8791489.

[9] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Advances in neural information processing systems, 2017, pp. 4765–4774. doi:10.5555/3295222.3295230.

[10] D. Slack, S. Hilgard, E. Jia, S. Singh, H. Lakkaraju, Fooling lime and shap: Adversarial attacks on post hoc explanation methods, in: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2020, pp. 180–186.

[11] A. Backhaus, U. Seiffert, Classification in high-dimensional spectral data: Accuracy vs. interpretability vs. model size, Neurocomputing 131 (2014) 15–22. doi:10.1016/j.neucom.2013.09.048.

[12] A. Bibal, B. Frénay, Interpretability of machine learning models and representations: an introduction., in: M.Verleysen (Ed.), European Symposium on Artificial Neural Networks (ESANN), 2016, pp. 77–81.
[13] M. Mohammadi, N. Petkov, K. Bunte, R. F. Peletier, F.-M. Schleif, Globular cluster detection in the gaia survey, Neurocomputing 342 (2019) 164–171. doi:10.1016/j.neucom.2018.10.081.
[14] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: synthetic minority over-sampling technique, Journal of artificial intelligence research 16 (2002) 321–357. doi:10.1613/jair.953.
[15] R. O. Mujalli, G. López, L. Garach, Bayes classifiers for imbalanced traffic accidents datasets, Accident Analysis & Prevention 88 (2016) 37–51. doi:10.1016/j.aap.2015.12.003.
[16] E. Alpaydin, Introduction to machine learning, MIT press, 2020. doi:10.1017/S1351324906004438.
[17] C. Bishop, C. M. Bishop, et al., Neural networks for pattern recognition, Oxford university press, 1995. doi:10.7551/mitpress/4923.001.0001.
[18] D. B. Rubin, Inference and missing data, Biometrika 63 (3) (1976) 581–592. doi:10.1093/biomet/63.3.581.
[19] R. J. Little, A test of missing completely at random for multivariate data with missing values, Journal of the American statistical Association 83 (404) (1988) 1198–1202. doi:10.1080/01621459.1988.10478722.
[20] R. J. Little, D. B. Rubin, Statistical analysis with missing data, Vol. 793, John Wiley & Sons, 2019. doi:10.1002/9781119482260.
[21] N. B. Ipsen, P.-A. Mattei, J. Frellsen, not-miwae: Deep generative modelling with missing not at random data, arXiv preprint arXiv:2006.12871 (2020). doi:10.48550/arXiv.2006.12871.
[22] G. Chechik, G. Heitz, G. Elidan, P. Abbeel, D. Koller, Max-margin classification of data with absent features, Journal of Machine Learning Research 9 (Jan) (2008) 1–21.
[23] P. Royston, I. R. White, et al., Multiple imputation by chained equations (mice): implementation in stata, J Stat Softw 45 (4) (2011) 1–20. doi:10.18637/jss.v045.i04.
[24] S. Van Buuren, Flexible imputation of missing data (2018). doi:10.18637/jss.v085.b04.
[25] R. J. Little, Missing-data adjustments in large surveys, Journal of Business & Economic Statistics 6 (3) (1988) 287–296. doi:10.2307/1391878.
[26] J. K. Dixon, Pattern recognition with partly missing data, IEEE Transactions on Systems, Man, and Cybernetics 9 (10) (1979) 617–621. doi:10.1109/TSMC.1979.4310090.
[27] G. Doquire, M. Verleysen, Feature selection with missing data using mutual information estimators, Neurocomputing 90 (2012) 3–11. doi:10.1016/j.neucom.2012.02.031.
[28] E. Eirola, G. Doquire, M. Verleysen, A. Lendasse, Distance estimation in numerical data sets with missing values, Information Sciences 240 (2013) 115–128. doi:10.1016/j.ins.2013.03.043.
URL https://www.sciencedirect.com/science/article/pii/S0020025513002570
[29] R. van Veen, Analysis of missing data imputation applied to heart failure data, Masters thesis, University of Groningen (2016).
URL fse.studenttheses.ub.rug.nl/id/eprint/14679
[30] S. Ghosh, E. Baranowski, R. van Veen, G. de Vries, M. Biehl, W. Arlt, P. Tino, K. Bunte, Comparison of strategies to learn from imbalanced classes for computer aided diagnosis of inborn steroidogenic disorders, in: M. Verleysen (Ed.), Proc. of the 25th European Symposium on Artificial Neural Networks (ESANN), 2017, pp. 199–204.
[31] S. Ghosh, P. Tino, K. Bunte, Visualisation and knowledge discovery from interpretable models, in: 2020 International Joint Conference on Neural Networks (IJCNN), IEEE, 2020, pp. 1–8. doi:10.1109/ijcnn48605.2020.9206702.

[32] B. M. Marlin, Missing data problems in machine learning, Ph.D. thesis, University of Toronto, Toronto, Ont., Canada, Canada, aAINR57898 (2008).
[33] M. E. Tipping, C. M. Bishop, Mixtures of probabilistic principal component analyzers, Neural Computation 11 (2) (1999) 443–482. doi:10.1162/089976699300016728.
[34] K. A. Severson, M. C. Molaro, R. D. Braatz, Principal component analysis of process datasets with missing values, Processes 5 (3) (2017) 38. doi:10.3390/pr5030038.
[35] T. Kohonen, J. Hynninen, J. Kangas, J. Laaksonen, K. Torkkola, Lvq pak: The learning vector quantization program package, Tech. rep., Technical report (1996). doi:10.1109/IJCNN.1992.287101.
[36] A. S. Sato, K. Yamada, Generalized learning vector quantization, in: Advances in Neural Information Processing Systems, Vol. 8, 1996, pp. 423–429.
[37] M. Shumska, K. Bunte, Multispectral texture classification in agriculture, in: M. Verleysen (Ed.), Proc. of the 31th European Symposium on Artificial Neural Networks (ESANN), i6doc.com, Bruges (Belgium) and online event, 2023, pp. 297–302. doi:10.14428/esann/2023.ES2023-110.
URL https://doi.org/10.14428/esann/2023.ES2023-110
[38] M. Shumska, K. Bunte, Towards robust colour texture classification with limited training data, in: N. Tsapatsoulis, A. Lanitis, M. Pattichis, C. Pattichis, C. Kyrkou, E. Kyriacou, Z. Theodosiou, A. Panayides (Eds.), Proc. of the 20th Conference on Computer Analysis of Images and Patterns (CAIP), Springer Nature Switzerland, Limassol, Cyprus, 2023, pp. 153–163. doi:10.1007/978-3-031-44237-7_15.
URL https://doi.org/10.1007/978-3-031-44237-7
[39] H. De Vries, R. Memisevic, A. C. Courville, Deep learning vector quantization, in: European Symposium on Artificial Neural Networks (ESANN), 2016, pp. 503–508.
URL https://api.semanticscholar.org/CorpusID:53231240
[40] T. Villmann, M. Biehl, A. Villmann, S. Saralajew, Fusion of deep learning architectures, multilayer feedforward networks and learning vector quantizers for deep classification learning, in: 2017 12th international workshop on self-organizing maps and learning vector quantization, clustering and data visualization (WSOM), IEEE, 2017, pp. 1–8.
[41] S. Saralajew, L. Holdijk, M. Rees, T. Villmann, Prototype-based neural network layers: incorporating vector quantization, arXiv preprint arXiv:1812.01214 (2018).
[42] J. Ravichandran, M. Kaden, T. Villmann, Variants of recurrent learning vector quantization, Neurocomputing 502 (2022) 27–36.
[43] K. Q. Weinberger, L. K. Saul, Distance metric learning for large margin nearest neighbor classification., Journal of machine learning research 10 (2) (2009).
[44] J. Goldberger, G. E. Hinton, S. Roweis, R. R. Salakhutdinov, Neighbourhood components analysis, Advances in neural information processing systems 17 (2004).
[45] B. Hammer, M. Strickert, T. Villmann, On the generalization ability of GRLVQ networks, Neural Processing Letters 21 (2) (2005) 109–120. doi:10.1007/s11063-004-1547-1.
URL http://dx.doi.org/10.1007/s11063-004-1547-1
[46] D. C. Liu, J. Nocedal, On the limited memory bfgs method for large scale optimization, Mathematical programming 45 (1) (1989) 503–528.
[47] R. H. Byrd, P. Lu, J. Nocedal, C. Zhu, A limited memory algorithm for bound constrained optimization, SIAM Journal on scientific computing 16 (5) (1995) 1190–1208.
[48] C. Zhu, R. H. Byrd, P. Lu, J. Nocedal, Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization, ACM Transactions on mathematical software (TOMS) 23 (4) (1997) 550–560.
[49] D.-J. Kroon, Fminlbfgs: Fast limited memory optimizer—file exchange—matlab central, Mathworks File Exchange (2010).
[50] P. T. Fletcher, C. Lu, S. M. Pizer, S. Joshi, Principal geodesic analysis for the study of nonlinear statistics of shape, IEEE Trans. on Medical Imaging 23 (8) (2004) 995–1005.

doi:10.1109/TMI.2004.831793.
[51] B. Hammer, T. Villmann, Generalized relevance learning vector quantization, Neural Networks 15 (8–9) (2002) 1059 – 1068. doi:http://dx.doi.org/10.1016/S0893-6080(02)00079-5.
URL http://www.sciencedirect.com/science/article/pii/S0893608002000795
[52] P. Schneider, M. Biehl, B. Hammer, Relevance matrices in learning vector quantization, in: M. Verleysen (Ed.), Proc. of the 15th European Symposium on Artificial Neural Networks (ESANN), d-side publishing, Bruges, Belgium, 2007, pp. 37–43. doi:10.4230/DagSemProc.07131.7.
[53] K. Bunte, P. Schneider, B. Hammer, F.-M. Schleif, T. Villmann, M. Biehl, Limited Rank Matrix Learning – Discriminative Dimension Reduction and Visualization, Neural Networks 26 (4) (2012) 159–173. doi:10.1016/j.neunet.2011.10.001.
URL http://dx.doi.org/10.1016/j.neunet.2011.10.001
[54] P. Schneider, M. Biehl, B. Hammer, Adaptive relevance matrices in learning vector quantization, Neural Comput. 21 (12) (2009) 3532–3561. doi:10.1162/neco.2009.11-08-908.
[55] M. Pazzani, C. Merz, P. Murphy, K. Ali, T. Hume, C. Brunk, Reducing misclassification costs, in: Proc. of the 11th ICML, Morgan Kauffmann, San Francisco, 1994, pp. 327–225. doi:10.1016/B978-1-55860-335-6.50034-9.
URL https://www.sciencedirect.com/science/article/pii/B9781558603356500349
[56] R. C. Wilson, E. R. Hancock, E. Pekalska, R. P. W. Duin, Spherical and hyperbolic embeddings of data, IEEE Trans. Pattern Anal. Mach. Intell. 36 (11) (2014) 2255–2269. doi:10.1109/TPAMI.2014.2316836.
URL http://dx.doi.org/10.1109/TPAMI.2014.2316836
[57] S. Seo, K. Obermayer, Soft learning vector quantization, Neural computation 15 (7) (2003) 1589–1604. doi:10.1162/089976603321891819.
[58] A. Villmann, M. Kaden, S. Saralajew, T. Villmann, Probabilistic learning vector quantization with cross-entropy for probabilistic class assignments in classification learning, in: International Conference on Artificial Intelligence and Soft Computing, Springer, 2018, pp. 724–735. doi:10.1007/978-3-319-91253-0_67.
[59] P. Schneider, M. Biehl, B. Hammer, Distance learning in discriminative vector quantization, Neural computation 21 (10) (2009) 2942–2969. doi:10.1162/neco.2009.10-08-892.
[60] P. Schneider, T. Geweniger, F.-M. Schleif, M. Biehl, T. Villmann, Multivariate class labeling in Robust Soft LVQ., in: M. Verleysen (Ed.), European Symposium on Artificial Neural Networks (ESANN), d-side publishing, 2011, pp. 17–22.
[61] L. Breiman, Bagging predictors, Machine learning 24 (2) (1996) 123–140. doi:10.1007/BF00058655.
[62] L. Breiman, Random forests, Machine learning 45 (1) (2001) 5–32. doi:doi.org/10.1023/A:1010933404324.
[63] H. Karcher, Riemannian center of mass and mollifier smoothing, Communications on Pure and Applied Mathematics 30 (5) (1977) 509–541. doi:10.1002/cpa.3160300502.
URL onlinelibrary.wiley.com/doi/abs/10.1002/cpa.3160300502
[64] W. S. Kendall, Probability, Convexity, and Harmonic Maps with Small Image I: Uniqueness and Fine Existence, Proceedings of the London Mathematical Society s3-61 (2) (1990) 371–406. doi:10.1112/plms/s3-61.2.371.
URL doi.org/10.1112/plms/s3-61.2.371
[65] K. A. Krakowski, K. Hüper, J. H. Manton, On the computation of the karcher mean on spheres and special orthogonal groups, in: in Proc. Workshop Robot. Math. (RoboMat) '07, 2007, pp. 119–124.
URL https://hdl.handle.net/1959.11/11040
[66] T. Ando, C.-K. Li, R. Mathias, Geometric means, Linear Algebra and its Applications 385 (2004) 305 – 334, special Issue in honor of Peter Lancaster. doi:10.1016/j.laa.2003.11.

019.
URL sciencedirect.com/science/article/pii/S0024379503008693

[67] S. Bonnabel, A. Collard, R. Sepulchre, Rank-preserving geometric means of positive semi-definite matrices, Linear Algebra and its Applications 438 (8) (2013) 3202 – 3216. doi:10.1016/j.laa.2012.12.009.
URL sciencedirect.com/science/article/pii/S0024379512008646

[68] S. Bonnabel, R. Sepulchre, Riemannian metric and geometric mean for positive semidefinite matrices of fixed rank, SIAM Journal on Matrix Analysis and Applications 31 (3) (2010) 1055–1070. doi:10.1137/080731347.
URL https://doi.org/10.1137/080731347

[69] R. M. Clark, R. Thompson, Statistical comparison of palaeomagnetic directional records from lake sediments, Geophysical Journal International 76 (2) (1984) 337–368. doi:10.1111/j.1365-246X.1984.tb05050.x.
URL https://doi.org/10.1111/j.1365-246X.1984.tb05050.x

[70] G. Wagner, On means of distances on the surface of a sphere (lower bounds)., Pacific J. Math. 144 (2) (1990) 389–398. doi:10.2140/pjm.1990.144.389.
URL projecteuclid.org:443/euclid.pjm/1102645739

[71] G. Wagner, On means of distances on the surface of a sphere. ii. upper bounds., Pacific J. Math. 154 (2) (1992) 381–396. doi:10.2140/pjm.1992.154.381.
URL https://projecteuclid.org:443/euclid.pjm/1102635628

[72] G. S. Watson, Statistics on spheres, University of Arkansas lecture notes in the mathematical sciences.v. 6, Wiley, New York, 1983, "A Wiley-Interscience publication.".
URL http://hdl.handle.net/2027/mdp.39015017408140

[73] P. Alfeld, M. Neamtu, L. Schumaker, Bernstein-Bézier polynomials on spheres and sphere-like surfaces, Comput. Aided Geom. Des. 13 (1996) 333–349. doi:10.1016/0167-8396(95)00030-5.

[74] S. R. Buss, J. P. Fillmore, Spherical averages and applications to spherical splines and interpolation, ACM Transactions on Graphics 20 (2001) 95–126. doi:10.1145/502122.502124.

[75] B. Afsari, Riemannian L_p center of mass: existence, uniqueness, and convexity, Proceedings of the American Mathematical Society 139 (2) (2011) 655–673.
URL http://www.jstor.org/stable/41059320

[76] T. Marrinan, B. Draper, J. R. Beveridge, M. Kirby, C. Peterson, Finding the subspace mean or median to fit your need, in: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR '14, IEEE Computer Society, USA, 2014, p. 1082–1089. doi:10.1109/CVPR.2014.142.

[77] M. Arnaudon, C. Dombry, A. Phan, L. Yang, Stochastic algorithms for computing means of probability measures, Stochastic Processes and their Applications 122 (4) (2012) 1437 – 1455. doi:10.1016/j.spa.2011.12.011.
URL sciencedirect.com/science/article/pii/S030441491100319X

[78] D. A. BINI, B. MEINI, F. POLONI, An effective matrix geometric mean satisfying the ando-li-mathias properties, Mathematics of Computation 79 (269) (2010) 437–452. doi:10.1090/S0025-5718-09-02261-3.
URL http://www.jstor.org/stable/40590410

[79] J. Lawson, Y. Lim, Weighted means and Karcher equations of positive operators, Proceedings of the National Academy of Sciences 110 (39) (2013) 15626–15632. doi:10.1073/pnas.1313640110.
URL https://www.pnas.org/content/110/39/15626

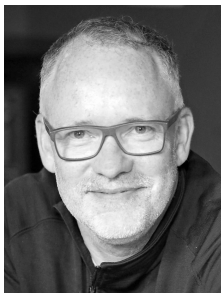
[80] P. Turaga, A. Veeraraghavan, A. Srivastava, R. Chellappa, Statistical computations on Grassmann and Stiefel manifolds for image and video-based recognition, IEEE Trans. Pattern Analysis and Machine Intelligence 33 (11) (2011) 2273–2286. doi:10.1109/TPAMI.

2011.52.
[81] S. Shirazi, M. T. Harandi, C. Sanderson, A. Alavi, B. C. Lovell, Clustering on Grassmann manifolds via kernel embedding with application to action analysis, in: 2012 19th IEEE International Conference on Image Processing, 2012, pp. 781–784. doi:10.1109/ICIP.2012.6466976.
[82] T. Carson, D. G. Mixon, S. Villar, Manifold optimization for k-means clustering, in: 2017 International Conference on Sampling Theory and Applications (SampTA), 2017, pp. 73–77. doi:10.1109/SAMPTA.2017.8024388.
[83] M. J. Azur, E. A. Stuart, C. Frangakis, P. J. Leaf, Multiple imputation by chained equations: what is it and how does it work?, International journal of methods in psychiatric research 20 (1) (2011) 40–49. doi:10.1002/mpr.329.
[84] U. Lall, A. Sharma, A nearest neighbor bootstrap for resampling hydrologic time series, Water Resources Research 32 (3) (1996) 679–693. doi:10.1029/95WR02966.
[85] E. S. Baranowski, W. Arlt, J. Idkowiak, Monogenic disorders of adrenal steroidogenesis, Hormone research in paediatrics 89 (5) (2018) 292–310. doi:10.1159/000488034.
[86] W. Arlt, E. A. Walker, N. Draper, H. E. Ivison, J. P. Ride, F. Hammer, S. M. Chalder, M. Borucka-Mankiewicz, B. P. Hauffa, E. M. Malunowicz, et al., Congenital adrenal hyperplasia caused by mutant p450 oxidoreductase and human androgen synthesis: analytical study, The Lancet 363 (9427) (2004) 2128–2135. doi:10.1016/S0140-6736(04)16503-3.
[87] K.-H. Storbeck, L. Schiffer, E. S. Baranowski, V. Chortis, A. Prete, L. Barnard, L. C. Gilligan, A. E. Taylor, J. Idkowiak, W. Arlt, et al., Steroid metabolome analysis in disorders of adrenal steroid biosynthesis and metabolism, Endocrine Reviews 40 (6) (2019) 1605–1625. doi:10.1210/er.2018-00262.
[88] J. H. W. Jr., Hierarchical grouping to optimize an objective function, Journal of the American Statistical Association 58 (301) (1963) 236–244. doi:10.1080/01621459.1963.10500845.
URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500845
[89] F. Murtagh, P. Legendre, Ward's hierarchical agglomerative clustering method: Which algorithms implement Ward's criterion?, Journal of Classification 31 (2014) 274–295. doi:10.1007/s00357-014-9161-z.
URL https://doi.org/10.1007/s00357-014-9161-z
[90] A. Janosi, W. Steinbrunn, M. Pfisterer, R. Detrano, Heart disease data set, UCI machine learning repository (1988).
URL https://archive.ics.uci.edu/ml/datasets/heart+Disease
[91] D. Ramanan, S. Baker, Local distance functions: A taxonomy, new algorithms, and an evaluation, IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (4) (2011) 794–806. doi:10.1109/ICCV.2009.5459265.

Author Biography

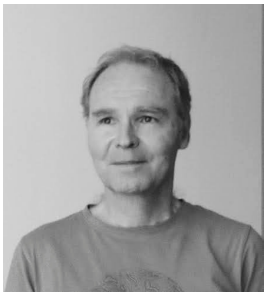
Sreejita Ghosh received a PhD in Machine Learning from the Bernoulli Institute at the University of Groningen, the Netherlands, in September 2021. She gained interdisciplinary experiences within different disciplines of Medicine and Epidemiology through collaboration during her PhD and postdoc positions, and being embedded within such research groups (former postdoc position). Since June 2023 she is a postdoc at the Department of Mathematics and Computer Science of TU Eindhoven, through DAIsy, a ITEA4 project, developing AI solutions for mental healthcare. Her research interests include developing interpretable ML and causal ML methods for human-centric applications. Further information in https://ghoshsreejita.com/.

Dr. Elizabeth Baranowski is a Medical Doctor and subspecialty trainee in Paediatric Endocrinology working at Birmingham Children's Hospital UK. She graduated from University of Birmingham in 2009. She has completed an NIHR awarded Academic Clinical Fellowship, BCH Research Foundation Springboard Fellowship and Medical Research Council awarded Clinical Research Training Fellowship within the Institute of Metabolism and Systems Research, University of Birmingham. She was awarded her PhD in 2022 from the University of Birmingham. Her research interests to date focus on inherited disorders of steroidogenesis, and the fusion of mass spectrometry methods with machine learning analysis for improved biochemical testing.



Michael Biehl received a Ph.D. in Physics from the University of Gießen, Germany, in 1992 and completed a Habilitation in Theoretical Physics at the University of Würzburg, Germany in 1996. He is currently professor of computer science at the Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence of the University of Groningen, The Netherlands. He holds an honorary professorship for machine learning at the Institute of Metabolism and Systems Research, University of Birmingham, UK. His research interest is in the theoretical investigation, development and application of machine learning methods. Further information, preprints etc. are available at http://www.cs.rug.nl/~biehl

Wiebke Arlt is the Director of the Medical Research Council Laboratory of Medical Science (LMS), a national research institute delivering fundamental discovery science to explore mechanisms relevant to human health and disease. She is also a Professor of Transdisciplinary Medicine at Imperial College London and a Consultant Endocrinologist at Imperial College Healthcare NHS Trust. At the LMS, she leads a multidisciplinary research group comprising biochemists, analytical chemists, computer scientists and clinician scientists. Her group investigates the role of steroids in health and disease, with a specific focus on exploring the mechanistic basis for sex differences in the impact of androgens on metabolic dysfunction.



Peter Tino holds a Chair position in Complex and Adaptive Systems at the University of Birmingham, UK. He is fascinated by the possibilities of cross-disciplinary blending of machine learning, mathematical modelling and domain knowledge in a variety of scientific disciplines. He led an EPSRC-funded consortium on developing a new mathematical framework for personalised healthcare and was awarded three Outstanding Paper of the Year Awards from the IEEE Transactions on Neural Networks and the IEEE Transactions on Evolutionary Computation. Peter (co-)chaired Task Force on Mining Complex Astronomical Data and Neural Networks Technical Committee (IEEE Computational Intelligence Society).

Prof. Kerstin Bunte received her Ph.D. in Computer Science in December 2011 from the University of Groningen, The Netherlands. Research visits and postdoctoral positions have taken her to the University of Rochester (USA), Bielefeld University (Germany), Aalto University (Finland) and Université catholique de Louvain (Belgium). She got a European Marie-Curie Fellowship with the University of Birmingham (UK), where she is now Honorary Fellow, and joined as a Rosalind Franklin Fellow at the University of Groningen in 2016, where she is now Associate professor for Machine Learning for interdisciplinary data analysis. Her main interests are developing the theory and practice of Machine Learning methods for interdisciplinary data analysis. This includes efficiency, interpretability, dimensionality reduction, scientific Machine Learning and principled inclusion of expert knowledge. Demonstrators include solutions in Medicine, Astronomy and Smart Industry. Prof. Bunte partnered in the European ITN SUNDIAL, EDUCADO, and is the leader of the NWO Smart Industry project SMART-AGENTS and NWO Vidi project for mechanistic machine learning. She (co)-authored more than 50 papers and is associate editor for IEEE TNNLS, Neural Processing Letters and guest editor for Neurocomputing. She participates regularly in program committees of several machine learning conferences. Further information can be obtained from https://www.cs.rug.nl/~kbunte/.

Appendix A Derivatives of ALVQ variants

A.1 Angle LVQ derivatives

The derivatives of the cost-function of ALVQ (Eq. (3)) are as follows:

$$\frac{\partial E}{\partial \Omega} = \sum_{i=1}^{N} \frac{\partial f}{\partial \mu_i} \frac{\partial \mu_i}{\partial \Omega} \quad \text{and} \quad \frac{\partial E}{\partial \boldsymbol{w}^{L \in [J,K]}} = \sum_{i=1}^{N} \frac{\partial f}{\partial \mu_i} \frac{\partial \mu_i}{\partial \boldsymbol{w}^{L}}, \text{ with} \tag{A. 20}$$

$$\frac{\partial \mu_i}{\partial \Omega} = \frac{2d^K}{(d^J + d^K)^2} \cdot \frac{\partial d^J}{\partial \Omega} - \frac{2d^J}{(d^J + d^K)^2} \cdot \frac{\partial d^K}{\partial \Omega} \ , \tag{A. 21}$$

$$\frac{\partial \mu_i}{\partial \boldsymbol{w}^J} = \frac{2d^K}{(d^J + d^K)^2} \cdot \frac{\partial d^J}{\partial \boldsymbol{w}^J}, \text{ and } \frac{\partial \mu_i}{\partial \boldsymbol{w}^K} = \frac{-2d^J}{(d^J + d^K)^2} \cdot \frac{\partial d^K}{\partial \boldsymbol{w}^K}, \text{ s.t.} \tag{A. 22}$$

$$\frac{\partial d_i^L}{\partial \Omega} = \frac{\partial g_\beta(b)}{\partial b_\Lambda} \cdot \frac{\partial b_\Lambda(\boldsymbol{x}_i, \boldsymbol{w}^L)}{\partial \Omega} \text{ and } \frac{\partial d_i^L}{\partial \boldsymbol{w}^L} = \frac{\partial g_\beta(b)}{\partial b_\Lambda(\boldsymbol{x}_i, \boldsymbol{w}^L)} \cdot \frac{\partial b_\Lambda(\boldsymbol{x}_i, \boldsymbol{w}^L)}{\partial \boldsymbol{w}^L} \ . \tag{A. 23}$$

Partial derivatives w.r.t $b_\Lambda(\boldsymbol{x}_i, \boldsymbol{w}^L)$ are given by:

$$\frac{\partial b_\Lambda(\boldsymbol{x}_i, \boldsymbol{w}^L)}{\partial \boldsymbol{w}^L} = \frac{\boldsymbol{x}_i \Omega^\top \Omega \|\boldsymbol{w}^L\|_\Omega^2 - \boldsymbol{x}_i \Omega^\top \Omega \boldsymbol{w}^L \cdot \boldsymbol{w}^L \Omega^\top \Omega}{\|\boldsymbol{x}_i\|_\Omega \|\boldsymbol{w}^L\|_\Omega^3} \ , \text{ and} \tag{A. 24}$$

$$\begin{aligned} \frac{\partial b_\Lambda(\boldsymbol{x_i}, \boldsymbol{w}^L)}{\partial \Omega_{mn}} =& \frac{x_{i,n} \sum_j^D \Omega_{mj} w_j^L + w_n^L \sum_j^D \Omega_{mj} x_{i,j}}{\|\boldsymbol{x}_i\|_\Omega \|\boldsymbol{w}^L\|_\Omega} \\ & - \boldsymbol{x}_i \Omega^\top \Omega \boldsymbol{w}^L \left[\frac{x_{i,n} \sum_j \Omega_{mj} x_{i,j}}{\|\boldsymbol{x}_i\|_\Omega^3 \|\boldsymbol{w}^L\|_\Omega} + \frac{w_n^L \sum_j \Omega_{mj} w_j^L}{\|\boldsymbol{x}_i\|_\Omega \|\boldsymbol{w}^L\|_\Omega^3} \right]. \end{aligned} \tag{A. 25}$$

where $x_{i,n}$ denotes dimension n of vector $\boldsymbol{x}_i$ and $n = 1, \ldots, M$.

Significance of β in ALVQ. In general, the greater the distance between a sample and a prototype, the lesser is the contribution of that sample towards the update strength of the prototype and Ω. However, β in Eq. (3) parameterizes the strength of this relation between the sample's distance and its contribution towards updating the parameters, such that $\beta \approx 0$ linearizes this relationship.

A.2 Angle LGMLVQ derivatives

In [31] we also introduced the angle variant of the localized GMLVQ (LGMLVQ) and its reduced rank variant. The dissimilarity b in the local extension of ALVQ is written as:

$$b = b_{\Omega^L} = \frac{\boldsymbol{x}_i^\top \Omega^{L\top} \Omega^L \boldsymbol{w}^L}{\|\boldsymbol{x}_i\|_{\Omega^L} \|\boldsymbol{w}^L\|_{\Omega^L}} \ , \tag{A. 26}$$

with corresponding derivatives given by:

$$\frac{\partial b_{\Omega^L}}{\partial \boldsymbol{w}^L} = \frac{\boldsymbol{x}_i \Omega^{L\top} \Omega^L \|\boldsymbol{w}^L\|^2_{\Omega^L} - \boldsymbol{x}_i \Omega^{L\top} \Omega^L \boldsymbol{w}^L \cdot \boldsymbol{w}^L \Omega^{L\top} \Omega^L}{\|\boldsymbol{x}_i\|_{\Omega^L} \|\boldsymbol{w}^L\|^3_{\Omega^L}} \tag{A. 27}$$

$$\begin{aligned}\frac{\partial b_{\Omega^L}}{\partial \Omega^L_{mn}} =& \frac{x_{i,n} \sum_j^D \Omega^L_{mj} w^L_j + w^L_n \sum_j^D \Omega^L_{mj} x_{i,j}}{\|\boldsymbol{x}_i\|_{\Omega^L} \|\boldsymbol{w}^L\|_{\Omega^L}} - \\ & \boldsymbol{x}_i \Omega^{L\top} \Omega^L \boldsymbol{w}^L \left[\frac{x_{i,n} \sum_j^D \Omega^L_{mj} x_{i,j}}{\|\boldsymbol{x}_i\|^3_{\Omega^L} \|\boldsymbol{w}^L\|_{\Omega^L}} + \frac{w^L_n \sum_j^D \Omega^L_{mj} w^L_j}{\|\boldsymbol{x}_i\|_{\Omega^L} \|\boldsymbol{w}^L\|^3_{\Omega^L}} \right] .\end{aligned} \tag{A. 28}$$

This variant is especially efficient when the decision-boundaries are not linearly separable.