Artificial Intelligence In Medicine 149 (2024) 102786

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

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

Research paper

Subspace corrected relevance learning with application in neuroimaging

Rick van Veen [a], Neha Rajendra Bari Tamboli [a], Sofie Lövdal [a,b], Sanne K. Meles [c], Remco J. Renken [d], Gert-Jan de Vries [e], Dario Arnaldi [f,g], Silvia Morbelli [g,h], Pedro Clavero [i], José A. Obeso [j], Maria C. Rodriguez Oroz [k,l,m], Klaus L. Leenders [b], Thomas Villmann [n], Michael Biehl [a,o,*]

[a] Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen, The Netherlands
[b] Department of Nuclear Medicine and Molecular Imaging, University Medical Center Groningen, The Netherlands
[c] Department of Neurology, University Medical Center Groningen, The Netherlands
[d] Department of Biomedical Sciences of Cells & Systems, Cognitive Neuroscience Center, University Medical Center Groningen, The Netherlands
[e] Philips Research, Healthcare, The Netherlands
[f] Department of Neuroscience, University of Genoa, Italy
[g] IRCCS Ospedale Policlinico San Martino, Genoa, Italy
[h] Department of Health Sciences, University of Genoa, Italy
[i] Servicio de Neurología, Complejo Hospitalario de Navarra, Pamplona, Spain
[j] Académico de Número Real Academia Nacional de Medicina de España, Spain
[k] Neurology Department, Clínica Universidad de Navarra, Spain
[l] Neuroscience Program, Center for Applied Medical Research, Universidad de Navarra, Pamplona, Spain
[m] Navarra Institute for Health Research, Pamplona, Spain
[n] Saxon Institute for Computational Intelligence and Machine Learning, University of Applied Sciences Mittweida, Germany
[o] SMQB, Inst. of Metabolism and Systems Research, College of Medical and Dental Sciences, Birmingham, United Kingdom

ARTICLE INFO

Keywords:
Learning vector quantization
Relevance learning
Generalized Matrix Learning Vector Quantization (GMLVQ)
Multi-source data
Neuroimaging

ABSTRACT

In machine learning, data often comes from different sources, but combining them can introduce extraneous variation that affects both generalization and interpretability. For example, we investigate the classification of neurodegenerative diseases using FDG-PET data collected from multiple neuroimaging centers. However, data collected at different centers introduces unwanted variation due to differences in scanners, scanning protocols, and processing methods. To address this issue, we propose a two-step approach to limit the influence of center-dependent variation on the classification of healthy controls and early vs. late-stage Parkinson's disease patients. First, we train a Generalized Matrix Learning Vector Quantization (GMLVQ) model on healthy control data to identify a "relevance space" that distinguishes between centers. Second, we use this space to construct a correction matrix that restricts a second GMLVQ system's training on the diagnostic problem. We evaluate the effectiveness of this approach on the real-world multi-center datasets and simulated artificial dataset. Our results demonstrate that the approach produces machine learning systems with reduced bias - being more specific due to eliminating information related to center differences during the training process - and more informative relevance profiles that can be interpreted by medical experts. This method can be adapted to similar problems outside the neuroimaging domain, as long as an appropriate "relevance space" can be identified to construct the correction matrix.

1. Introduction

Machine learning models require data to make predictions or identify patterns. In the field of medicine, acquiring enough data can be impractical or even risky for patients. Therefore, researchers often combine data from multiple sources to satisfy the data requirements of machine learning systems. However, multi-source data can pose a problem as it may contain sources of variation that are not intrinsic to the classes being distinguished, but rather related to differences

* Corresponding author at: Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen, The Netherlands.
E-mail addresses: r.van.veen133@gmail.com (R. van Veen), barineha29@gmail.com (N.R.B. Tamboli), s.s.lovdal@rug.nl (S. Lövdal), s.k.meles@umcg.nl (S.K. Meles), r.j.renken@umcg.nl (R.J. Renken), gj.de.vries@philips.com (G.-J. de Vries), dario.arnaldi@gmail.com (D. Arnaldi), silviadaniela.morbelli@hsanmartino.it (S. Morbelli), pedro.clavero.ibarra@navarra.es (P. Clavero), jobeso.hmcinac@hmhospitales.com (J.A. Obeso), k.l.leenders@umcg.nl (K.L. Leenders), thomas.villmann@hs-mittweida.de (T. Villmann), m.biehl@rug.nl (M. Biehl).

https://doi.org/10.1016/j.artmed.2024.102786
Received 10 July 2023; Received in revised form 12 January 2024; Accepted 21 January 2024
Available online 24 January 2024

between the data sources themselves. This can result in biased machine learning systems with inflated performance. In this study, we propose an approach to address these issues and improve the Learning Vector Quantization (LVQ) models during training on functional brain images from patients with neurodegenerative diseases and healthy controls.

Neurodegenerative diseases have a significant impact on patients and caregivers as they progressively impair cognitive and/or motor functions. Parkinson's disease (PD) is expected to be one of the leading causes of death by 2040 after cancer [1]. Early and accurate diagnosis is crucial for developing prevention strategies and precision therapeutic measures. However, diagnosing Parkinson's disease based on clinical features is difficult, particularly in its early stages, as the motor symptoms can be subtle and similar to other disorders within the Parkinsonian clinical spectrum [2,3]. Studying patients at different stages of the disease will be vital for early diagnosis.

In the early stages of Parkinson's disease, biomarkers are required to confirm its presence and provide information on the rate of progression and lead time. Combined with advanced computation, imaging the brain with ^{18}F-2-fluoro-2-deoxy-D-glucose Positron Emission Tomography (FDG-PET) may provide a solution. FDG-PET imaging of the brain can provide valuable information about neuronal activity, as the radiotracer FDG reflects the cerebral metabolic rate of glucose [4]. Local pathology can cause decreased FDG uptake, indicating impaired neuronal function in certain brain regions. Spatial covariance analysis of FDG-PET images can be performed using Scaled Subprofile Model and Principal Component Analysis (SSM/PCA), which reduces the large number of voxels for every subject to a limited number of orthogonal dimensions that explain the major sources of variance in the data. SSM/PCA has been used to identify disease-related patterns in specific neurodegenerative conditions, as reviewed by Eidelberg [5] and Meles et al. [6]. These patterns have subsequently been used for the differential diagnosis of Parkinson's disease [7,8] and Alzheimer's disease [9,10]. To improve diagnostic accuracy in multiclass problems in neurodegenerative diseases, SSM/PCA has also been combined with machine learning approaches [11–13].

Ideally, large numbers of patients in different stages of disease should be studied to properly investigate the diagnostic accuracy of FDG-PET. This is especially due to the high dimensionality of the 3D PET images, where a single scan has hundreds of thousands of voxels. It applies particularly to the use of *end-to-end* Deep Learning techniques, which are currently infeasible for the type of problem and set-up considered here [14,15]. In general, larger data sets are required for the application of powerful models with potentially higher classification performance. So far, this has been hampered by the typically small numbers of patients in single-center neuroimaging studies. Combining data from different centers across the world is key, and attempts to do so have already been made for FDG-PET and other neuroimaging markers in large database projects such as ADNI[1] and PPMI.[2] In previous work we have combined FDG-PET data from three neuroimaging centers [12,13]. The initial assumption was that these data did not contain any variation depending on the center they originated from. However, in [12] we showed that the center of origin could be predicted using the data from healthy controls, which would not be possible without center dependent variation. This variation between centers is likely caused by differences in scanners and processing methods or parameters [16]. In other work, Albrecht et al. [17], Mueller et al. [18], Bisenius et al. [19], Martí-Andrés et al. [20] and Cobbinah et al. [21] have applied machine learning techniques to multi-center data. However, no explicit solution for the multi-center data variation was found. Furthermore, the problem in these publications is often limited to a binary problem, i.e., to differentiate between healthy controls (HC) and PD. In this case, center variation can be a problem, but is potentially less strong than the intrinsic variation between HC and PD and therefore less problematic. When the data are mixed, i.e., each class of data originates from several (and the same) neuroimaging centers, center differences become less likely to be picked up by Machine Learning methods, as these do not help to distinguish between the two classes. In contrast, the dataset in this work combines data from healthy controls and different stages of PD each collected at a different neuroimaging center. To the best of our knowledge, no alternative method is available that could be readily employed for the suppression of center-specific bias in the setting considered in this work.

Learning Vector Quantization [22] has since its introduction grown to an important family of supervised learning algorithms. In the training phase, the algorithms determine class-specific prototypes that represent the classes in the data space. Predictions are made based on their distance from the prototypes: A novel sample is classified by computing the distances from the sample to all prototypes and assigning it to the label of the closest prototype. Numerous variants of LVQ exist [23] with success in the biomedical field, medicine, and industry.[3] The approach introduced here is based on Generalized Matrix Learning Vector Quantization (GMLVQ) introduced by Schneider et al. [24]. The authors Schneider et al. [24] extended the work of Sato and Yamada [25] and Hammer and Villmann [26] by introducing an adaptive matrix in the distance measure. The addition of this "relevance matrix" makes it possible to account for correlations of dimensions and rotations of the data space [24], supporting the classification. GMLVQ has demonstrated competitive performance and provided useful insights in a multitude of other biomedical applications, as shown by the relevance matrix analysis [27–31]. The relevance matrix helps to explain the decisions made by the classifier and facilitates the interpretability of the LVQ system. Previous research has shown that in comparable diagnostic problems concerning Parkinsonian Syndromes and SSM/PCA, GMLVQ's performance is superior or at least on par with that of Support Vector Machines and decision trees [12,13,32].

The computational complexity of GMLVQ has been discussed in [24] and follow-up studies. Note that for the present study, which employs low-dimensional representations of small data sets, computational and memory requirements play a minor role and have not been investigated in detail.

The current study introduces a method, referred to as subspace correction [33], that produces a GMLVQ system that can discriminate between early and late stage PD and healthy controls, while restrained from using any of the center-specific variance found in the data.

2. Materials and methods

The goal of this section, specifically Section 2.1, is to present our novel subspace corrected relevance learning procedure based on Generalized Matrix Learning Vector Quantization (GMLVQ). To validate the procedure, we include experiments on an artificial dataset containing four Gaussian clusters (Section 2.3). We apply the method to data obtained from three neuroimaging centers (Section 2.4) containing center-dependent variation [13].

2.1. Generalized matrix LVQ

Subspace corrected relevance learning is based on GMLVQ, an LVQ system which employs an adaptive distance measure as introduced by Schneider et al. [24].

We consider the classification of N-dim. feature vectors $\mathbf{x}_i \in \mathbb{R}^N$ with target class labels $S_i \in \{1, 2, \ldots, C\}$. In LVQ, class assignments are based on the distances of $\mathbf{x}_i$ from a set of M prototypes $\{\mathbf{w}_j \in \mathbb{R}^N\}_{j=1}^M$. Each prototype represents one of C classes as denoted by the labels $S(\mathbf{w}_j) \in \{1, 2, \ldots, C\}$.

[1] https://adni.loni.usc.edu/.
[2] https://www.ppmi-info.org/.
[3] http://www.cis.hut.fi/research/som-bibl/.

GMLVQ incorporates relevance factors in the distance measure by employing a matrix of adaptive parameters which is concurrently optimized with the prototype vectors. Specifically, the adaptive distance measure of GMLVQ is parameterized as

$$d^{\Lambda}(\boldsymbol{w}_j, \boldsymbol{x}_i) = (\boldsymbol{x}_i - \boldsymbol{w}_j)^{\top} \Lambda (\boldsymbol{x}_i - \boldsymbol{w}_j), \tag{1}$$

with the relevance matrix $\Lambda \in \mathbb{R}^{N \times N}$. In order for $d^{\Lambda}(\cdot, \cdot)$ to be a proper non-negative measure, i.e. a semi-metric, the relevance matrix Λ has to be positive semi-definite and symmetric. These properties can be realized by using the parameterization

$$\Lambda = \Omega^{\top} \Omega \tag{2}$$

and optimizing Ω during training instead of Λ directly. In addition, the relevance matrix Λ is normalized by enforcing $\mathrm{Tr}(\Lambda) = 1$ to aid numerical stability and interpretability after training [24].

The updates of the prototypes and relevance matrix can be computed using, for instance, a variety of gradient descent based methods [34]. In the experiments presented here we use *Way Point Gradient Descent Optimization*, see [35] for details, to optimize the objective function introduced by Sato and Yamada [25]

$$E = \sum_{i=1}^{P} f\left(\mu^{\Lambda}(\boldsymbol{x}_i)\right) \tag{3}$$

with a monotonically increasing activation function f. For the results presented in this work we resorted to the simple identity function $f(z) = z$.

The relative difference distance function in Eq. (3) is given by

$$\mu^{\Lambda}(\boldsymbol{x}_i) = \frac{d^{\Lambda}(\boldsymbol{w}_+, \boldsymbol{x}_i) - d^{\Lambda}(\boldsymbol{w}_-, \boldsymbol{x}_i)}{d^{\Lambda}(\boldsymbol{w}_+, \boldsymbol{x}_i) + d^{\Lambda}(\boldsymbol{w}_-, \boldsymbol{x}_i)}. \tag{4}$$

Here, $\mathbf{w}_+$ denotes the *closest correct* prototype with $d(\mathbf{w}_+, \mathbf{x}_i) \leq d(\mathbf{w}_j, \mathbf{x}_i)$ for all j with $S(\mathbf{w}_j) = S_i$. Similarly, $\mathbf{w}_-$ is the *closest wrong* prototype with a label $S(\mathbf{w}_j) \neq S_i$.

The representation of a given matrix Λ by Ω is not unique. This makes the direct comparison of matrices Ω from different systems impossible. Although this is not a problem for the GMLVQ classification, as ultimately Λ is used in the distance function, it is useful to construct a canonical variant of Ω. From Linear Algebra it is known that a real, symmetric and positive semi-definite matrix Λ can be written as

$$\Lambda = \sum_{j=1}^{N} \lambda_j \boldsymbol{v}_j \boldsymbol{v}_j^{\top}, \tag{5}$$

with the N orthonormal eigenvectors $\boldsymbol{v}_j$ and (without loss of generality) ordered eigenvalues λ_j, such that

$$\lambda_1 \geq \lambda_2 \cdots \geq \lambda_J > 0 = \lambda_{J+1} = \lambda_{J+2} = \cdots \lambda_N. \tag{6}$$

In words, Λ has J eigenvectors with non-zero eigenvalues. The eigenvectors correspond to combinations of features or directions in feature space that describe part of the class-dependent differences in the data, which are important for classification. We refer to the J leading eigenvectors as the *relevance space* of the GMLVQ system.

According to Eq. (5), we can always construct a valid canonical representation

$$\hat{\Omega} = \begin{bmatrix} - - \sqrt{\lambda_1}\, \boldsymbol{v}_1^{\top} - - \\ - - \sqrt{\lambda_2}\, \boldsymbol{v}_2^{\top} - - \\ \cdots \cdots \cdots \\ - - \sqrt{\lambda_N}\, \boldsymbol{v}_N^{\top} - - \end{bmatrix}, \tag{7}$$

which also parameterizes the relevance matrix, i.e., $\Lambda = \hat{\Omega}^{\top} \hat{\Omega}$.

It has been shown analytically by Biehl et al. [36] and observed empirically that GMLVQ has a strong tendency to yield singular matrices Λ of very low rank [24,36–38]. Hence, the number of zero (or numerically very small) eigenvalues will be large. Furthermore, because Λ is symmetric, the eigenvectors are orthonormal and can be used to construct a low dimensional visualization of the data by projecting the data on the eigenvectors [13,36,37]. This type of visualization is typically used to identify outliers and find similar data samples [13,36,37,39]. However, in this work we use the visualization only to qualitatively discuss the effect of the suggested correction method.

2.2. Subspace corrected relevance learning

We first use a simplified artificial data set to illustrate the problem and our solution (see Fig. 1). In this example we include artificial HC and PD data from two hypothetical neuroimaging centers, where center-specific variance has been introduced. When we train GMLVQ on this space, to distinguish between HC and PD subjects, we find a "relevance space" that includes a contribution due to the differences between the centers. This relevance space is visualized using a dashed arrow and can be found on the left side of Fig. 1 labeled as "without correction".

The first step of the procedure is to train a GMLVQ system to distinguish between the HCs of the centers. We work under the assumption that healthy controls provide similar data across the cohorts of subjects. The relevance space found by this procedure would be the orange arrow in Fig. 1 which we denote mathematically by Λ_c, parameterized by $\hat{\Omega}_c^{\top} \hat{\Omega}_c$. The subscript "c" indicates a variable being associated with this initial center classification problem. Using this initial result we can construct a correction matrix

$$\Psi_c = \left[I - \sum_{j=1}^{J} \boldsymbol{v}_j \boldsymbol{v}_j^{\top} \right] = \left[\sum_{j=J+1}^{N} \boldsymbol{v}_j \boldsymbol{v}_j^{\top} \right], \tag{8}$$

with eigenvectors $\boldsymbol{v}_j$ of Λ_c, ordered equivalent to Eq. (6), that corresponds to the nullspace of the center problem's relevance space (Λ_c) and represents a projection where no relevant directions from the center classification problem are present. In the second step, with correction, we train to distinguish between the HCs and PD patients while projecting out the relevance space of GMLVQ trained to distinguish between the centers. The effect of this correction is shown in the illustration on the right of Fig. 1, labeled "corrected". The system that is corrected during training is not able use any of the center differences to distinguish between the HCs and PD patients. This results in a relevance space more faithful to the intrinsic differences between HCs and PD patients (dashed arrow).

Variables of the diagnosis problem are indicated with the subscript "d". The correction procedure requires an adaptation of the standard GMLVQ training procedure. After each update of Ω_d, the correction

$$\tilde{\Omega}_d = \Omega_d \Psi_c. \tag{9}$$

is applied. Note that the normalization $\mathrm{Tr}(\Lambda_d) = 1$ needs to be enforced after the correction procedure, which does not influence the orthogonality of Λ_d. We can check that the correction step has the desired effect for all (J) eigendirections $\boldsymbol{v}_l$ relevant in the center classification problem by considering

$$\begin{aligned} \tilde{\Omega}_d \boldsymbol{v}_l &= \Omega_d \left[I - \sum_{k=1}^{J} \boldsymbol{v}_k \boldsymbol{v}_k^{\top} \right] \boldsymbol{v}_l \\ &= \Omega_d \boldsymbol{v}_l - \Omega_d \left[\sum_{k=1}^{J} \boldsymbol{v}_k \underbrace{\boldsymbol{v}_k^{\top} \boldsymbol{v}_l}_{\delta_{kl}} \right] \\ &= \Omega_d \boldsymbol{v}_l - \Omega_d \boldsymbol{v}_l = 0, \end{aligned} \tag{10}$$

with δ_{kl} the Kronecker delta. Hence, we see that after the correction Ω_d does not contain contributions from the eigenvectors that discriminate the centers. Thus, the classifier for the diagnostic problem will not consider these directions in the computation of distances from the prototypes and cannot be confused by the center-specific properties of the data.

In order to measure the effectiveness of the correction procedure, we can quantify if the uncorrected systems indeed use center-dependent

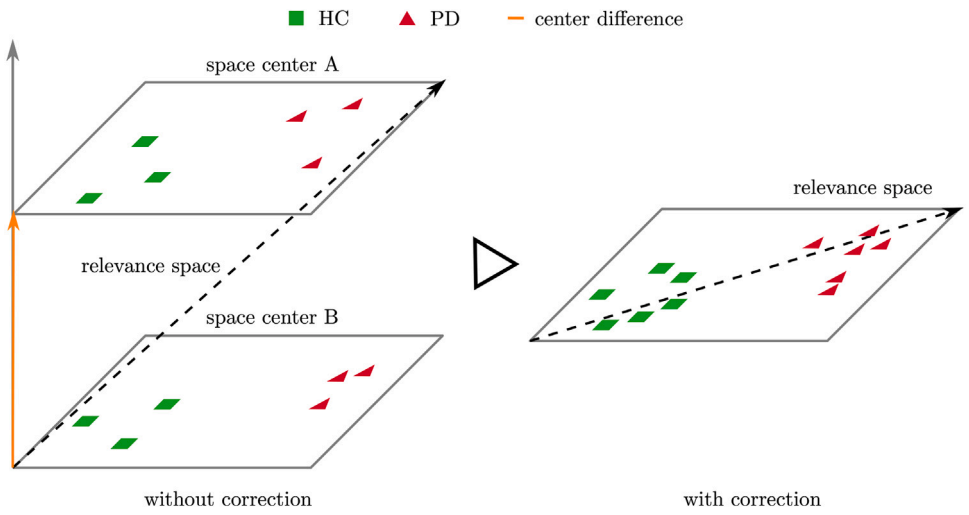


Fig. 1. Illustrative visualization of the problem and the result of the orthogonal learning correction procedure. Both illustration contain two planes representing the different centers, "A" and "B". The orange line denotes the difference between them. The learned relevance space of the systems are drawn as a dashed arrow. On the **left**, center differences will attribute to the relevance space of an uncorrected GMLVQ system. On the **right**, any contribution in the direction of center A is projected out. In this simplified (and ideal) example, this can be interpreted as the data being projected into the space of center B.

differences. A measure can be obtained by computing the angle between the eigenvectors of the "center" problem ($\boldsymbol{v}_c$), and the eigenvectors of the uncorrected "diagnostic" classification system ($\boldsymbol{v}_d$). This angle can be computed using the scalar product of the normalized eigenvectors

$$\phi = \cos^{-1}(\boldsymbol{v}_c \cdot \boldsymbol{v}_d). \tag{11}$$

The function in Eq. (11) returns a value between 0° and 90°. Under the assumption, all center differences are found by GMLVQ and are contained within the eigenvectors of its relevance matrix; these values can be interpreted as follows: 90° indicates no center variations have been used, and 0° means the eigenvectors are entirely overlapping, and thus, only center variations have been used.

2.3. Artificial dataset

We include an artificial dataset of four two-dimensional Gaussian clusters (Fig. 2), with the addition of eight randomly generated noise features (uncorrelated with zero mean and unit variance). The clusters each include 1000 samples and are labeled with a combination of a letter and a number, i.e., A1, A2, B1, and B2. The letter represents a "center" (Fig. 2(a)) and the number a "disease" (Fig. 2(b)). The average of the first two features of the four Gaussian clusters are $\mu_{A1} = [1.5, 0]$, $\mu_{A2} = [0.5, 1]$, $\mu_{B1} = [-0.5, -1]$, and $\mu_{B2} = [-1.5, 0]$. In our concrete example, the covariance matrix for the first two features for all four clusters is

$$\begin{bmatrix} 1.89, & 0.77 \\ 0.77, & 0.47 \end{bmatrix}.$$

To compute the correction matrix, we follow the two-step approach as described in Section 2.1. First, we train a GMLVQ system on the "center" problem, i.e., the discrimination of {A1, A2} from {B1, B2}. We obtain the average Λ_c from a ten times repeated ten-fold cross-validation [40]. We compute the correction matrix and use it in the diagnosis problem, i.e., {A1, B1} vs. {A2, B2}, for which we perform the same cross-validation procedure. We use the sklvq [41] implementation of GMLVQ with the following parameters: A single prototype per class, the way-point gradient descent procedure trained for 50 epochs, with a step size of 0.05 and 0.03 for the prototypes and relevance matrix, respectively. For the activation function the identity was used. All other parameters are left at their default values.[4]

[4] https://sklvq.readthedocs.io/en/0.1.2/api.html.

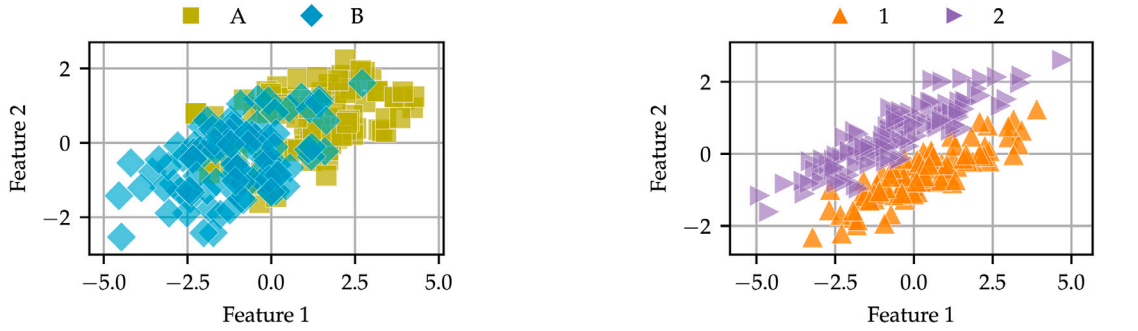
Table 1
General features of the space defining reference groups, used to transform the data.

	UMCG	
	HC (n = 17)	PD (n = 19)
Age, mean (std)	61.3 (7.5)	63.8 (7.5)
Male gender, n (%)	12 (70.6)	13 (68.4)
Disease evolution, mean (std)	–	3 (2)

2.4. Neuroimaging dataset

The dataset we are analyzing consists of brain images acquired using [18F] fluorodeoxyglucose positron emission tomography (FDG-PET) from both healthy controls (HC) and Parkinson's patients (PD). The data were collected at three different centers, namely the Movement Disorder Unit of the Clinica Universidad de Navarra (CUN) [42], the University Medical Center Groningen (UMCG) [43], and the University of Genoa and IRCCS AOU San Martino-IST (UGOSM) [44]. For details on the specific study setups, please refer to the respective publications. General information about the patients and healthy controls is provided in Table 2. Table 2 shows that the PD patients from the three centers are at different stages of the disease evolution. Specifically, the patients from UGOSM are at an early stage of PD while the patients from CUN are in a much later stage of the disease. The patients from UMCG are at a slightly later stage than the patients from UGOSM, resulting in a diverse group of patients with respect to disease evolution. In the three centers, different scanners, settings, and scanning protocols are used [42–44]. This variability has previously been observed to negatively affect the performance of GMLVQ in building a universally applicable model that is independent of the center [12].

We do not deal with the FDG-PET images directly. The images were processed using SSM/PCA based on an independent space-defining reference group, that includes a typical group of 19 Parkinson's patients and 17 healthy controls (see Table 1 for details). All FDG-PET images underwent spatial normalization to an FDG-PET template in Montreal Neurological Institute (MNI) brain space, as described in [45]. To remove voxels that were outside the brain, a 35% threshold of the whole-brain intensity maximum was applied to each FDG-PET image in the reference cohort (Table 1). The resulting masks were multiplied together to create a common mask that included only non-zero values



(a) Visualization of 200 random samples labeled with their "center" (letters) of origin.

(b) Visualization of 200 random samples labeled by their "diagnosis" (numbers).

Fig. 2. The data are a combination of four Gaussian clusters containing 1000 samples each. To increase clarity, the plots show a reduced number of samples.

Table 2
General features of the different groups.

	UMCG		UGOSM		CUN	
	HC (n = 19)	PD (n = 20)	HC (n = 49)	PD (n = 38)	HC (n = 20)	PD (n = 68)
Age, mean (std)	56 (14)	63 (9)	67.8 (11.6)	72 (6.8)	67.9 (3.1)	70.6 (6.4)
Male gender, n (%)	9 (47.4)	–	16 (32.7)	24 (63.2)	11 (55)	37 (54.4)
Disease evolution, mean (std)	–	3(2)	–	1.7(1.6)	–	13.6 (5.1)

for all subjects. This mask was then applied to all images. The masked images were log-transformed, and the subject mean was subtracted from each voxel. Additionally, each voxel was centered around the mean of the healthy controls included in the reference group.

After these steps a principal component analysis (PCA) is applied to the vectorized and preprocessed FDG-PET scans of the space defining reference group. This provides a low-dimensional representation describing the principal sources of variance in the HC and PD patients comprising the reference group. This procedure results in 35 principal components, the last (36th) component does not explain any of the variances in the data and is therefore not used. The data (Table 2) is then projected on these principal components resulting in the feature vectors that serve as input to the GMLVQ system.

We look at the problem within a number of settings, i.e., the two-class setting where we combined HC and PD data of all centers. Second, we test the correction by considering the different stages of disease evolution of the PD patients. Specifically, we looked at the patients from the UGOSM which are at an early stage, and patients from the CUN which are at a late stage in their disease evolution. We perform the analysis with HC included or excluded and report the results for both set-ups.

The correction matrix is obtained by the procedure described for the artificial dataset. We compute the average relevance matrix of GMLVQ trained to distinguish the centers from a ten times repeated ten-fold cross-validation procedure [40]. The data have been balanced by randomly oversampling the minority class(es) and z-transformed based on the training data within each cross-validation run. The corrected and uncorrected diagnostic systems are both trained using the same cross-validation procedures and model parameters. We use the sklvq [41] implementation of GMLVQ with the same parameters as described in Section 2.3. In short, we use the 35-dimensional PCA-based feature vectors corresponding to the FDG-PET scans of the patients listed in Table 2, and train a GMLVQ model first to classify the source (center of HCs), whereafter both an unrestricted and a restricted system is trained for a disease problem.

3. Results and discussion

In this section, we present the results of the experiments performed on the two datasets. Each section presents first the results of the relevant "center" problem, followed by the "diagnostic" problem(s). We report the accuracy and the area under the receiver operating characteristic curve (AUROC).

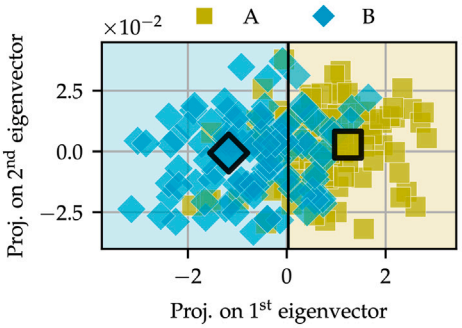
Table 3
Artificial dataset performance metrics. Values are the mean with standard deviation within parentheses of the randomized ten times repeated ten-fold cross-validation procedures [40].

	Center	Uncorrected	Corrected
AUROC	0.83(0.02)	0.99(0.00)	0.99(0.01)
Accuracy (%)	77.55(2.19)	96.14(0.97)	95.19(1.07)

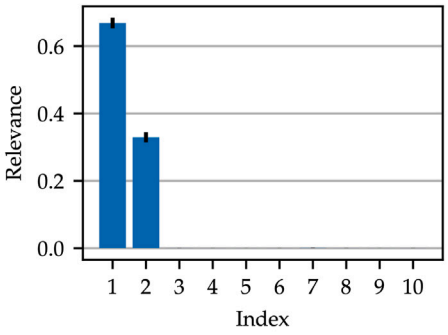
3.1. Artificial dataset

The results of the correction procedure on the artificial data can be found in Fig. 3 and Table 3. Fig. 3(a) presents the projection of the center problem, i.e., "A" vs. "B". The average relevance diagonal of this problem is shown in Fig. 3(b) and reveals that GMLVQ can pick up on the two relevant features, where a higher relevance is given to the first feature. The effect of the correction procedure on the relevance profile of the disease problem will, therefore, likely be more noticeable in the first than the second feature. The relevance space is entirely determined by the first eigenvector, with its corresponding eigenvalue being one. The correction matrix is therefore also based on this single eigenvector. Table 3 shows that the centers can be separated, although not perfectly, with an average AUROC of 0.83.

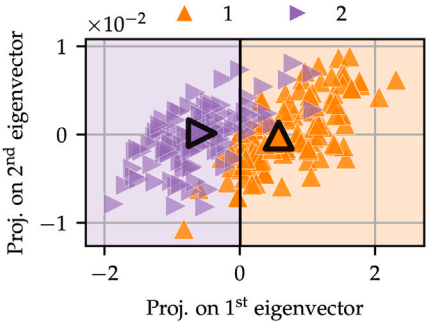
The projection of the uncorrected disease problem and average relevance diagonal are presented in Figs. 3(c) and 3(e). The relevance profile shows us that the second feature is slightly more relevant for classifying the diseases, i.e., "1" vs. "2". With the correction applied (Figs. 3(d) and 3(f)), the differences between the first and second feature are much more apparent. As expected, considering the most relevant features provided by the center system, the relevance of the first feature went down and that of the second went up. The average AUROC and accuracy are included in Table 3. We observe that performance differences between the uncorrected and corrected case are practically negligible, with small favor for the uncorrected case. This slight performance difference may indicate that, in the uncorrected case,



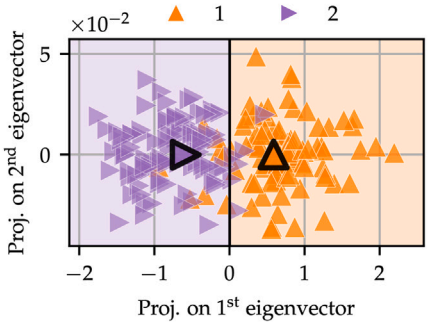
(a) Projection of 200 random samples on the eigenvectors of Λ_c.



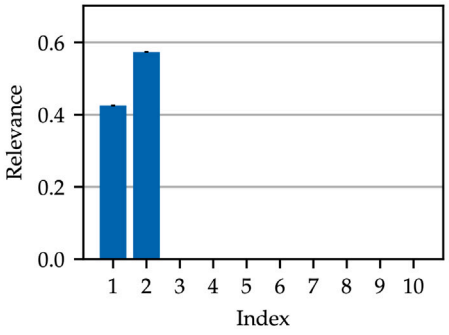
(b) Average diagonal of Λ_c with vertical bars representing the standard deviation.



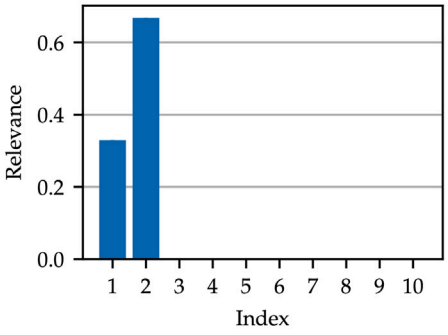
(c) Projection of 200 random samples on the eigenvectors of Λ_d (uncorrected).



(d) Projection of 200 random samples on the eigenvectors of $\tilde{\Lambda}_d$ (corrected).



(e) Average diagonal of Λ_d (uncorrected) with vertical bars illustrating the standard deviation.



(f) Average diagonal of $\tilde{\Lambda}_d$ (corrected) with vertical bars illustrating the standard deviation.

Fig. 3. Projections and relevance diagonals related to the correction procedure performed on the artificial dataset.

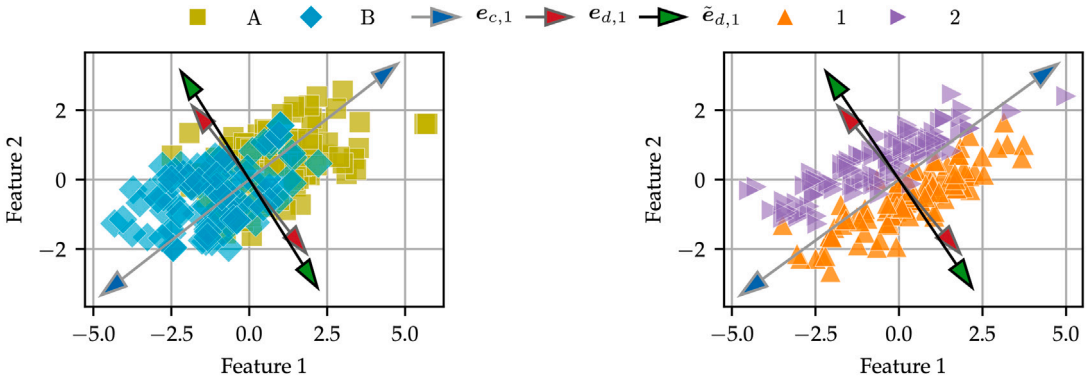
some of the center variances have been used to distinguish between the "diseases". Due to limiting the potential relevance space in the corrected case, these center variances can no longer be used, resulting in slightly less desirable performance but a "cleaner" classification system which does not exploit the purely center-specific bias in the data.

As our artificial data is practically two-dimensional and each of our GMLVQ systems found a relevance matrix that can be parameterized by a single eigenvector, we can visualize them within the original feature space (see Fig. 4). The angle between $\boldsymbol{v}_{c,1}$ and $\boldsymbol{v}_{d,1}$ is 84.36°. As hypothesized, some center variance has been used; see the blue and red arrows in Fig. 4. The corrected system produces by design an eigenvector with an angle of 90° to $\boldsymbol{v}_{c,1}$; the green arrow in Fig. 4. Note that the lengths of the arrows do not hold any significance and are merely different because of aesthetic purposes.

Visualizing the eigenvectors is unique to this case and cannot be done with higher (>3) dimensional data, such as our neuroimaging dataset. However, the angles between eigenvectors can still be computed and allow for a similar interpretation.

3.2. Neuroimaging dataset

Table 4 and Fig. 5 include the performance and plots associated with the three-center classification problem. Specifically, Fig. 5(a) visualizes how well the centers can be discriminated, with an average AUROC of 0.96. Compared to the artificial problem, the number of non-zero eigenvalues and thus eigenvectors used for the correction is different. The first two eigenvalues are 0.72 and 0.17, with near-zero eigenvalues for the rest. In Fig. 5(a), we can also observe that the HCs from the CUN are more straightforward to distinguish from the rest than subjects from the UGOSM and UMCG. As expected, considering the reference group used to construct the feature space, the center difference cannot directly be found in the first few features, see figure Fig. 5(b), as most



(a) Visualization of the eigenvectors on the data labeled by "center"

(b) Visualization of the eigenvectors on the data labeled by "disease"

Fig. 4. Visualization of the "center", uncorrected, and corrected "disease" leading eigenvectors. Note that the length of the eigenvectors does not signify anything; the arrows have different lengths to increase visual clarity.

Table 4
Performance on the uncorrected and corrected HC vs. PD problem, and corresponding center problem. Reported values are the averages with standard deviation within parenthesis stratified from the repeated cross-validation procedure.

	Center	HC vs. PD	
		Uncorrected	Corrected
(a) Training performance			
AUROC	1.00(0.00)	0.89(0.02)	0.89(0.02)
Accuracy (%)	97.32(1.90)	83.00(2.54)	83.09(2.39)
(b) Validation performance			
AUROC	0.96(0.08)	0.85(0.09)	0.85(0.08)
Accuracy (%)	89.93(9.95)	79.38(8.31)	79.72(7.20)

variance in the data is explained by the difference in HC and PD [46–48]. Primarily, features 6, 9, and 35 seem to be more relevant for this case.

Table 4 and Fig. 5 include the performance and plots associated with the two-class setting. In this setting, we have combined the data from the three centers, and labeled the data according to the diagnosis given to the subjects, i.e., HC or PD. The corrected and uncorrected two-class systems perform nearly equally well. Both, the uncorrected and corrected systems have an average AUROC of about 0.85. Also, the eigenvalue "profile" is similar for both systems, with the first eigenvector contributing most to the relevance profiles presented in Figs. 5(e) and 5(f). These relevance profiles do not show an apparent effect, as observed in the artificial case.

It is essential to note that we simplify the assessment here. The effect of the correction cannot be evaluated based on the relevance diagonal only. The relevance diagonal summarizes the information; the correction is based on eigenvectors, i.e., multiple combinations of features.

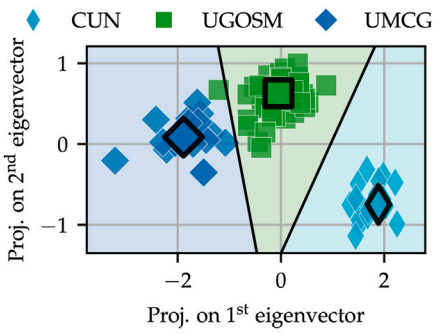
The uncorrected system's first and only relevant eigenvector makes an angle of approximately 83.10° and 78.40° with the center's first and second eigenvectors. Hence, the correction procedure impacts the system, as center difference is used to classify HC vs. PD patients. However, this effect is not very strong, and the corrected system seems to compensate by finding a slightly different relevance space that performs similarly without using the center differences. Furthermore, features 6, 9, and 35, relevant for the center classification, seem to play only a minor role in the disease classification. Comparing the relevance expression of the uncorrected system with the corrected one, we observe a slight increase in the first feature's relevance. These observations underline that, in the "simple" case, i.e., HC vs. PD, the relevant variation is mainly contained within the first few SSM/PCA features, as previously observed in [12,13,46–48].

3.2.1. Early vs. Late stage Parkinson's disease

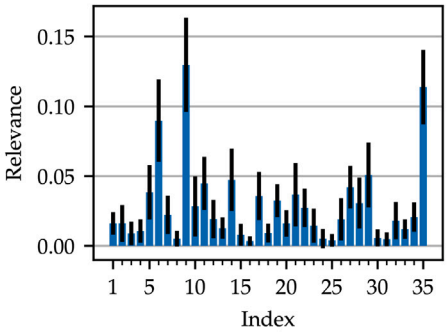
Figs. 6 and 7 contain the projections and average relevance diagonals, considering HC and PD subjects from the CUN and UGOSM centers. However, instead of labeling the subjects by their center of origin in combination with their diagnosis, we have labeled them as early and late-stage PD patients (Table 2) and ignored the center of origin for the HC entirely; Fig. 6 shows the results for this three-class problem. We use the HCs to construct the correction matrix. Therefore, we conducted a second experiment excluding the HCs from the diagnosis problem, thus including only early vs. late-stage PD patients (see Fig. 5). The performance results for this 2-class and 3-class problem are included in Table 5.

In contrast to the three-center problem, classifying HCs from the CUN and UGOSM centers can be achieved with 100% accuracy. The projection is shown in Fig. 6(a) with average relevance diagonal that shows one very dominating feature for the classification, i.e., feature 35 (see Fig. 6(b)). We base the correction matrix on the first eigenvector that primarily dominates the relevance profile with an eigenvalue of approximately 1.0 and near-zero eigenvalues for the remaining eigenvectors.

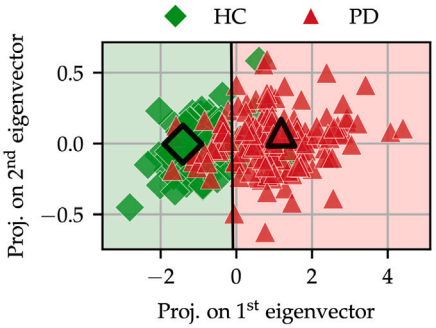
The uncorrected diagnostic projection of the system is included in Fig. 6(c), with average relevance diagonal in Fig. 6(e). The system has two non-zero eigenvalues. In the projection, we see that some center information must have been used in the first and possibly second eigenvector as the HC from the CUN are positioned closer to the PD from the same center than the other HC. Additionally, the relevance diagonal shows a high relevance for feature 35, observed to be associated with the center-dependent differences. These observations are confirmed by the angles to the eigenvector of the center problem. The first and second eigenvectors of the diagnostic system make an angle of 57.85° and 56°. Comparing these plots to the corrected system in Figs. 6(d) and 6(f), we see the HC from the CUN position more closely to the HC from the UGOSM, although not nicely scattered within the group. Similar to the 6-class setting discussed in the previous subsection, this result suggests that not all center-dependent variation has been removed. However, the most relevant feature, 35, has almost entirely reduced to zero for the corrected system. We also observe that in this setting, considering PD subjects at different disease stages, the first feature is no longer the most relevant. This again indicates that later features, i.e., principal components, play an important role when considering more complicated diagnostic problems. The performance of the corrected system is reduced compared to the uncorrected system (Table 5). The reduction in performance can be expected for similar reasons as discussed in the previous discussed problems. The correction procedure limits the variation available to discriminate between the different classes originating from different centers, increasing the



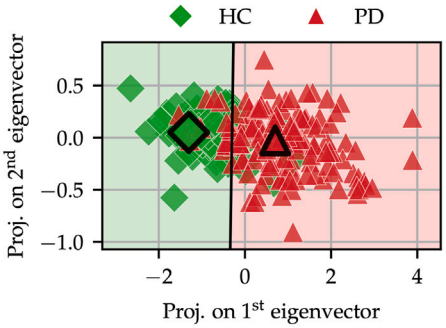
(a) Projection on the first two eigenvectors of the GMLVQ relevance matrix, with eigenvalues 0.72 and 0.16, respectively



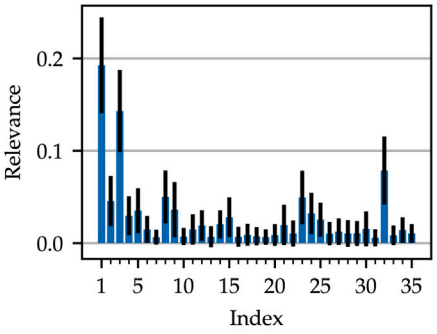
(b) The average relevance diagonal of the GMLVQ system constructed to differentiate the centers.



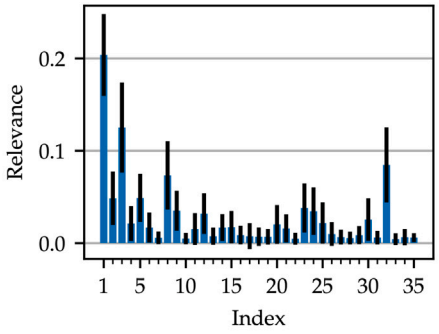
(c) Projection on the first two eigenvectors of the **uncorrected** GMLVQ relevance matrix, with eigenvalues 0.83 and 0.02, respectively



(d) Projection on the first two eigenvectors of the **corrected** GMLVQ relevance matrix, with eigenvalues 0.82 and 0.02, respectively



(e) The average relevance diagonal of the **uncorrected** diagnostic system.



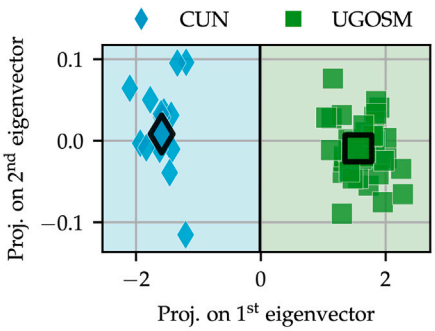
(f) The average relevance diagonal of the **corrected** diagnostic system.

Fig. 5. This figure includes a collection of experiment results performed on the multi-center "2-class" problem. Results are based on the average models from ten times repeated ten-fold cross-validation procedures. Unless otherwise specified in the subfigures, the left column contains the uncorrected and the right column the corrected results. The bar plots represent the average values (the bars) and the standard deviations as the black error lines at the top of the bars.

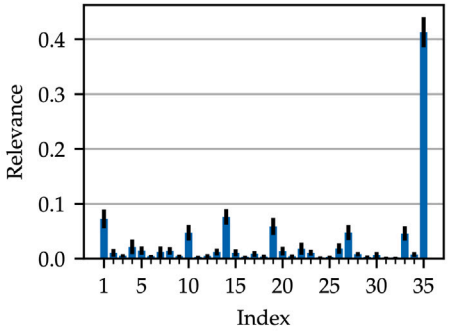
problem's difficulty and reducing performance. However, it also results in a more true discriminative system. E.g., in Fig. 6(d) we see a less separated distribution of early vs late PD, which can be considered more realistic, as PD is assumed to be a continuous spectrum in terms of disease progression [49].

The two-class problem, excluding the HC from the classification task, shows the value of the correction procedure even more clearly. In this case, the PD patients from the CUN and UGOSM are again labeled by their disease evolution stage. Figs. 7(a) and 7(c) include the projection and average relevance diagonal, primarily defined by the first eigenvector with an eigenvalue of 0.95. Together with the near-perfect performance in Table 5, the projection shows that the GMLVQ system found a way to discriminate between the two classes. However, the relevance profile in Fig. 7(c) is very similar to the profile of the center classification task shown in Fig. 6(b). In this specific case, there are two distinct sources of variation in the data. One source of variance due to the difference in disease stage and the center-dependent source. Additionally, the variance between the disease stage difference likely holds less discriminative power than the difference between the center of origin. These observations make it likely that mainly the differences between the centers have been learned by GMLVQ. The angle between the eigenvectors of the diagnostic system and the center system is 40.15°, confirming the observations.

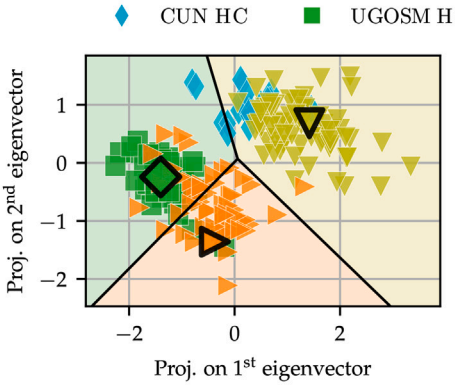
Comparing the uncorrected to the corrected system, we see that performance decreases (Table 5). However, comparing the relevance diagonal, shown in Fig. 7(d), we see that feature 35 entirely vanished and



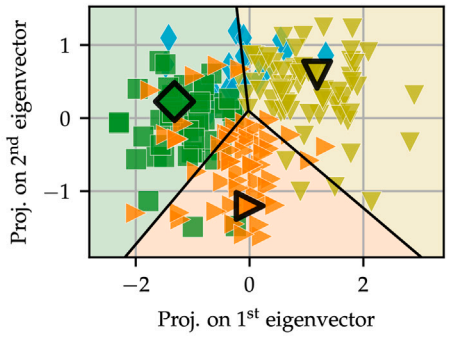
(a) Projection on the first two eigenvectors of the GMLVQ relevance matrix, with eigenvalues 1.00 and 0.00, respectively



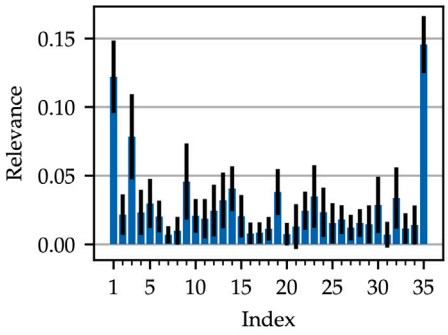
(b) The average relevance diagonal of the GMLVQ system constructed to differentiate the centers.



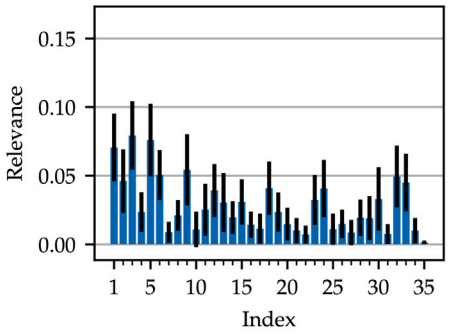
(c) Projection on the first and second eigenvectors of the **uncorrected** GMLVQ relevance matrix, with eigenvalues 0.49 and 0.31, respectively



(d) Projection on the first and second eigenvectors of the **corrected** GMLVQ relevance matrix, with eigenvalues 0.48 and 0.31, respectively



(e) The average relevance diagonal of the **uncorrected** system.

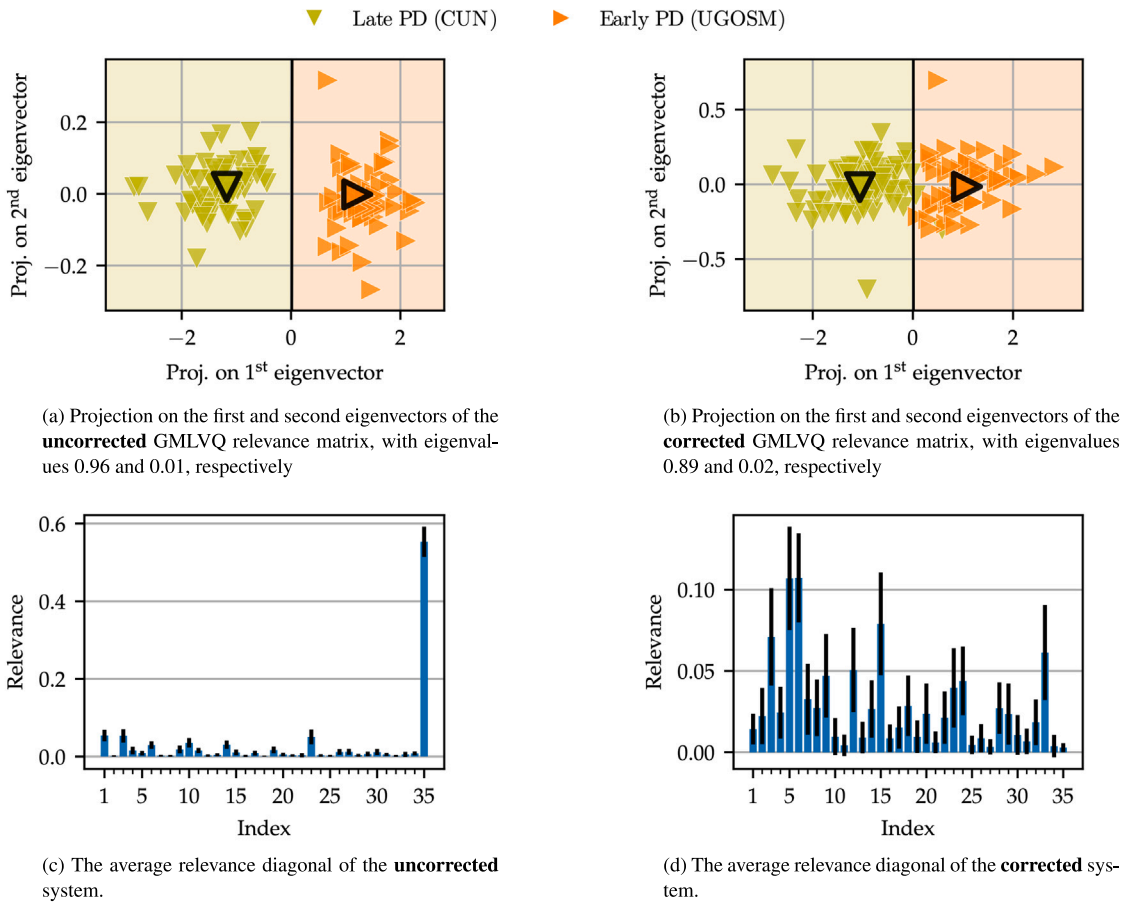


(f) The average relevance diagonal of the **corrected** system.

Fig. 6. This figure includes a collection of experiment results performed on the data of two centers including HC and PD labeled by their disease stage. Results are based on the average models from ten times repeated ten-fold cross-validation procedures. Unless otherwise specified in the subfigures, the left column contains the uncorrected and the right column the corrected results. The bar plots represent the average values (the bars) and the standard deviations as the black error lines at the top of the bars.

produced a more pronounced relevance profile. Clearly, the orthogonal correction procedure enables GMLVQ to produce a more pronounced and defined relevance profile, accompanied by a reduced but more realistic performance. The nominal decrease of the classification accuracy merely reflects the fact that the corrected system is prevented from exploiting the misleading, irrelevant differences between centers in the classification.

In general, a strength of the method is that it is fully explainable, due to its linear basis. This facilitates analytical extensions and transparent interpretation of the model and correction step. Note that the linearity of the approach also facilitates an interpretation of prototypes and relevances in the original high-dimensional image space, as demonstrated in e.g. [39]. However, the linearity also constitutes a potential drawback, leading to limitations in modeling potentially nonlinear effects in the source specific information. Similarly, the assumption that the source specific variance is located in a subspace orthogonal to the information intrinsic to the classes of interest only holds as long as the unwanted (source specific) and wanted (class-specific) information are uncorrelated. The nature of the source-specific variance is depending on the type of data used, and we refer to future work to investigate the properties of the center-specific effects in FDG-PET scans, specifically.



(a) Projection on the first and second eigenvectors of the **uncorrected** GMLVQ relevance matrix, with eigenvalues 0.96 and 0.01, respectively

(b) Projection on the first and second eigenvectors of the **corrected** GMLVQ relevance matrix, with eigenvalues 0.89 and 0.02, respectively

(c) The average relevance diagonal of the **uncorrected** system.

(d) The average relevance diagonal of the **corrected** system.

Fig. 7. This figure includes a collection of experiment results performed on the data of two centers excluding HC and PD labeled by their disease stage. Results are based on the average models from ten times repeated ten-fold cross-validation procedures. Unless otherwise specified in the subfigures, the left column contains the uncorrected and the right column the corrected results. The bar plots represent the average values (the bars) and the standard deviations as the black error lines at the top of the bars.

Table 5
Performance on the uncorrected and corrected early vs. late stage PD problem, and corresponding center problem. Reported values are the averages with standard deviation within parenthesis stratified from the cross-validation procedure.

	Center	2-class		3-class	
		Uncorrected	Corrected	Uncorrected	Corrected
(a) Training performance					
AUROC	1.00(0.00)	1.00(0.00)	0.98(0.01)	0.90(0.02)	0.89(0.02)
Accuracy (%)	100.00(0.00)	99.98(0.13)	97.87(2.65)	78.19(2.65)	75.52(3.05)
(b) Validation performance					
AUROC	1.00(0.00)	1.00(0.00)	0.93(0.08)	0.85(0.07)	0.82(0.07)
Accuracy (%)	100.00(0.00)	98.50(3.12)	86.12(9.14)	70.20(9.70)	66.84(10.96)

4. Conclusion and future work

Our results confirm that center-dependent variation can be at least partially removed using the orthogonal learning correction procedure. The experiments on the artificial dataset show that the problem exists when two sources of variation are introduced in the data and that the procedure can correct this. Furthermore, we provide a measure quantifying how much of the unwanted variation has been used by the uncorrected system, i.e., the angle between the eigenvectors of the relevance matrices.

In the more straightforward cases, i.e., HC vs. PD, not much center variation was used to determine the diagnosis, and the correction procedure had thus limited effect. In the 6-class case, the correction affected the projections and performance since we explicitly asked the system to discriminate between the HCs. The projections of the corrected systems also showed that some center-dependent variation was still left. This observation implies that the correction matrix does not contain all center-dependent relevance space.

Similar conclusions can be drawn for the early versus late-stage PD experiments. When including the HCs, the correction clearly showed that the HCs from the two centers were closer together. However, the groups are still somewhat distinguishable in the plots. The most compelling case is the two-class scenario where the HCs were excluded. Without the correction procedure, the system would not have been able to decide between early and late-stage based on intrinsic differences between early and late-stage PD metabolic profiles. Instead, the results show that the decisions would have been based mainly on center-dependent variations. That means, that without a correction for the center information, we would have obtained a completely biased classifier where both the classification accuracy and the interpretation of the model would in practice have been non-sensical. The fact that the nominal performance or accuracy is lower in the corrected system should not be mistaken as a disadvantage of the method. The seemingly *superior* performance of the uncorrected classifier is based on naively exploiting the misleading center-specific information in the given, biased data set. The corrected system can be trusted to rely on those properties of the data, which truly relate to the target diagnosis. Accordingly, the relevance profile will be more informative about the disease-specific difference between the considered cohorts.

Finally, the experiments have shown encouraging results and identified the method's limitations, which is discussed further in the next section.

4.1. Future work

In the previous paragraphs, we concluded that the correction procedure did not remove all center-variation. This might be caused by GMLVQ not finding all possible directions that explain the center differences. It can be that another set of hyperparameters might resolve this issue by finding a better solution. Another possible resolve could be to apply GMLVQ several iterations and correct each subsequent system until an appropriate (poor) performance in the center classification has been reached. We could then extract the relevant eigenvectors from each of these systems, concatenate them and construct a final correction matrix to train GMLVQ on the diagnostic problem.

In our experiments, we have used an initial relevance matrix equal to the identity matrix. One could initialize the Ω matrix using the non-leading eigenvectors of Λ_c as these already contain directions orthogonal to the leading eigenvectors, likely speeding up convergence and possibly improving performance.

Three variations of the suggested correction method should be good candidates for future comparison. First, instead of correcting the GMLVQ relevance matrix during training, one could project out the center-relevant space beforehand, thereby manipulating the feature vectors and reducing the dimensionality of the data. Theoretically, no contributions in the unwanted directions are possible. However, in practice, due to numerical instability, one might still accumulate contributions over time. Second, one could reduce the application of the correction to only the final update of the relevance matrix. However, this will likely result in decreased performance compared to the other options. The interpretation of the prototypes (by themselves) will be more difficult as they will still contain contributions in the unwanted directions. The contributions will not be considered in the distance though because they would be projected out by the corrected $\tilde{\Omega}$.

Alternatively, the basic method can also be interpreted and implemented as a regularization technique. In contrast to the method we suggested in Section 2.2, where the identified directions are entirely removed during training, one can instead use a tunable (r) penalty term

$$P(\Omega) = r \cdot \sum_{i=1}^{M} \sum_{j=1}^{J} (\boldsymbol{v}_j^\top \omega_i)^2,$$

and add it to the cost function E, such that the new cost function and the partial derivative concerning the columns of Ω become

$$\hat{E} = E + P(\Omega),$$

$$\frac{\partial \hat{E}}{\partial \omega_i} = \frac{\partial E}{\partial \omega_i} + 2r \cdot \sum_{j=1}^{J} \boldsymbol{v}_j^\top.$$

A regularization will enable a less harsh and more controllable correction. In this way, one allows for some magnitude of the unwanted directions to be used but can control how much by increasing or decreasing the scaling parameter r. Moreover, the scaling factor could be a vector such that one can control the contribution to the penalty per eigenvector (when $J > 1$), as likely not all eigenvectors are as crucial to remove.

Finally, Villmann et al. [50] consider a similar approach from a transfer learning perspective. It employs a modified LVQ cost function which corresponds to a weighted combination of the source discrimination and the actual classification task. This single tier approach can be applied also in absence of a separate HC cohort for the analysis of center differences.

Future studies will explore and compare these as well as additional realizations of the basic ideas presented here.

CRediT authorship contribution statement

Rick van Veen: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Neha Rajendra Bari Tamboli:** Writing – review & editing, Writing – original draft, Validation, Investigation. **Sofie Lövdal:** Writing – review & editing, Writing – original draft, Validation, Investigation. **Sanne K. Meles:** Writing – review & editing, Writing – original draft, Resources, Data curation, Conceptualization. **Remco J. Renken:** Writing – review & editing, Writing – original draft, Conceptualization. **Gert-Jan de Vries:** Writing – review & editing, Writing – original draft, Conceptualization. **Dario Arnaldi:** Writing – review & editing, Writing – original draft, Resources. **Silvia Morbelli:** Writing – review & editing, Writing – original draft, Resources. **Pedro Clavero:** Writing – review & editing, Writing – original draft, Resources. **José A. Obeso:** Writing – review & editing, Writing – original draft, Resources. **Maria C. Rodriguez Oroz:** Writing – review & editing, Writing – original draft, Resources. **Klaus L. Leenders:** Writing – review & editing, Writing – original draft, Supervision, Resources, Conceptualization. **Thomas Villmann:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Michael Biehl:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Formal analysis, Conceptualization.

Declaration of competing interest

S.K. Meles, K.L. Leenders, R. van Veen reports financial support was provided by The Michael J Fox Foundation. D. Arnaldi, S. Morbelli reports financial support was provided by Italian Ministry of Health. R. van Veen reports financial support was provided by State of Upper Austria in the frame of SCCH competence center INTEGRATE. S.K. Meles, K.L. Leenders reports financial support was provided by Dutch Stichting Parkinson Fonds. S.K. Meles reports a relationship with The Michael J Fox Foundation that includes: funding grants. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The research reported in this article has been partly funded the Michael J. Fox Foundation (ID 17081), a grant from the Italian Ministry of Health to IRCCS Ospedale Policlinico San Martino (Fondi per la Ricerca Corrente 2019/2020, and Italian Neuroscience network (RIN)), BMK, BMDW, and the State of Upper Austria in the frame of SCCH competence center INTEGRATE [(FFG grant no. 892418)] part of the FFG COMET Competence Centers for Excellent Technologies Programme. Additionally, the authors acknowledge support from the Dutch Stichting ParkinsonFonds (grant number 2022/1891).

References

[1] Gammon K. Neurodegenerative disease: Brain windfall. Nature 2014;515(7526):299–300.
[2] Rizzo G, Copetti M, Arcuti S, Martino D, Fontana A, Logroscino G. Accuracy of clinical diagnosis of parkinson disease: a systematic review and meta-analysis. Neurology 2016;86(6):566–76.
[3] Jellinger KA, Logroscino G, Rizzo G, Copetti M, Arcuti S, Martino D, Fontana A. Accuracy of clinical diagnosis of parkinson disease: A systematic review and meta-analysis. Neurology 2016;87(2):237–8.
[4] Reivich M, Kuhl D, Wolf A, Greenberg J, Phelps Ma, Ido T, Casella V, Fowler J, Hoffman E, Alavi A, et al. The [18F] fluorodeoxyglucose method for the measurement of local cerebral glucose utilization in man.. Circ Res 1979;44(1):127–37.
[5] Eidelberg D. Metabolic brain networks in neurodegenerative disorders: a functional imaging approach. Trends Neurosci 2009;32(10):548–57.
[6] Meles SK, Kok JG, Renken RJ, Leenders KL. From positron to pattern: A conceptual and practical overview of 18f-FDG PET imaging and spatial covariance analysis. In: PET and SPECT in neurology. Springer International Publishing; 2020, p. 73–104.

[7] Rus T, Tomše P, Jensterle L, Grmek M, Pirtošek Z, Eidelberg D, Tang C, Trošt M. Differential diagnosis of parkinsonian syndromes: a comparison of clinical and automated - metabolic brain patterns' based approach. Eur J Nucl Med Mol Imaging 2020;47(12):2901–10.

[8] Tripathi M, Tang CC, Feigin A, Lucia ID, Nazem A, Dhawan V, Eidelberg D. Automated differential diagnosis of early parkinsonism using metabolic brain networks: A validation study. J Nucl Med 2015;57(1):60–6.

[9] Meles SK, Pagani M, Arnaldi D, Carli FD, Dessi B, Morbelli S, Sambuceti G, Jonsson C, Leenders KL, Nobili F. The alzheimer's disease metabolic brain pattern in mild cognitive impairment. J Cereb Blood Flow Metab 2017;37(12):3643–8.

[10] Perovnik M, Tomše P, Jamšek J, Emeršič A, Tang C, Eidelberg D, Trošt M. Identification and validation of alzheimer's disease-related metabolic brain pattern in biomarker confirmed alzheimer's dementia patients. Sci Rep 2022;12(1).

[11] Mudali D, Teune L, Renken R, Leenders K, Roerdink J. Classification of parkinsonian syndromes from FDG-PET brain data using decision trees with SSM/PCA features. Comput Math Methods Med 2015;2015:10.

[12] van Veen R, Talavera Martinez L, Kogan RV, Meles SK, Mudali D, Roerdink J, Massa F, Grazzini M, Obeso J, Rodriguez-Oroz M, Leenders K, Renken R, de Vries J, Biehl M. Machine learning based analysis of FDG-PET image data for the diagnosis of neurodegenerative diseases. In: Applications of intelligent systems. Frontiers in artificial intelligence and applications, vol. 310, IOS Press; 2018, p. 280–9.

[13] van Veen R, Gurvits V, Kogan R, Meles S, de Vries J, Renken R, Rodriguez-Oroz M, Rodriguez-Rojas R, Arnaldi D, Raffa S, de Jong B, Leenders K, Biehl M. An application of generalized matrix learning vector quantization in neuroimaging. Comput Methods Programs Biomed 2020;197:105708.

[14] Goodfellow I, Bengio Y, Courville A. Deep learning. MIT Press; 2016.

[15] Marcus G. Deep learning: A critical appraisal. 2018, CoRR, abs/1801.00631.

[16] Kogan RV, Jong BA, Renken RJ, Meles SK, Snick PJ, Golla S, Rijnsdorp S, Perani D, Leenders KL, Boellaard R. Factors affecting the harmonization of disease-related metabolic brain pattern expression quantification in 18f FDG-PET (PETMETPAT). In: Jovicich J, Frisoni GB, editors. Alzheimer's Dementia: Diag Assess Dis Monit 2019;11(1):472–82.

[17] Albrecht F, Bisenius S, Neumann J, Whitwell J, Schroeter ML. Atrophy in midbrain & cerebral/cerebellar pedunculi is characteristic for progressive supranuclear palsy – A double-validation whole-brain meta-analysis. NeuroImage: Clin 2019;22:101722.

[18] Mueller K, Jech R, Bonnet C, Tintěra J, Hanuška J, Möller HE, Fassbender K, Ludolph A, Kassubek J, Otto M, Růžička E, Schroeter ML. Disease-specific regions outperform whole-brain approaches in identifying progressive supranuclear palsy: A multicentric MRI study. Front Neurosci 2017;11.

[19] Bisenius S, Mueller K, Diehl-Schmid J, Fassbender K, Grimmer T, Jessen F, Kassubek J, Kornhuber J, Landwehrmeyer B, Ludolph A, Schneider A, Anderl-Straub S, Stuke K, Danek A, Otto M, Schroeter ML. Predicting primary progressive aphasias with support vector machine approaches in structural MRI data. NeuroImage: Clin 2017;14:334–43.

[20] Martí-Andrés G, Bommel L, Meles SK, Riverol M, Valentí R, Kogan RV, Renken RJ, Gurvits V, Laar T, Pagani M, Prieto E, Luquin MR, Leenders KL, Arbizu J. Multicenter validation of metabolic abnormalities related to PSP according to the MDS-PSP criteria. Mov Disorders 2020;35(11):2009–18.

[21] Cobbinah BM, Sorg C, Yang Q, Ternblom A, Zheng C, Han W, Che L, Shao J. Reducing variations in multi-center alzheimer's disease classification with convolutional adversarial autoencoder. Med Image Anal 2022;82:102585.

[22] Kohonen T. The self-organizing map. Proc IEEE 1990;78(9):1464–80.

[23] Nova D, Estévez PA. A review of learning vector quantization classifiers. Neural Comput Appl 2014;25(3–4):511–24.

[24] Schneider P, Biehl M, Hammer B. Adaptive relevance matrices in learning vector quantization. Neural Comput 2009;21(12):3532–61.

[25] Sato A, Yamada K. Generalized learning vector quantization. In: Conference on neural information processing systems. NIPS '95, Cambridge, MA, USA: MIT Press; 1995, p. 423–9.

[26] Hammer B, Villmann T. Generalized relevance learning vector quantization. Neural Netw 2002;15(8–9):1059–68.

[27] Arlt W, Biehl M, Taylor AE, Hahner S, Libe R, Hughes BA, Schneider P, Smith DJ, Stiekema H, Krone N, et al. Urine steroid metabolomics as a biomarker tool for detecting malignancy in adrenal tumors. J Clin Endocrinol Metab 2011;96(12):3775–84.

[28] Biehl M, Schneider P, Smith D, Stiekema H, Taylor A, Hughes B, Shackleton C, Stewart P, Arlt W. Matrix relevance LVQ in steroid metabolomics based classification of adrenal tumors. In: Verleysen M, editor. 20th European symposium on artificial neural networks (ESANN 2012). d-side publishing; 2012, p. 423–8.

[29] Yeo L, Adlard N, Biehl M, Juarez M, Smallie T, Snow M, Buckley CD, Raza K, Filer A, Scheel-Toellner D. Expression of chemokines CXCL4 and CXCL7 by synovial macrophages defines an early stage of rheumatoid arthritis. Ann Rheum Dis 2015;75(4):763–71.

[30] Mukherjee G, Bhanot G, Raines K, Sastry S, Doniach S, Biehl M. Predicting recurrence in clear cell renal cell carcinoma: Analysis of TCGA data using outlier analysis and generalized matrix LVQ. In: 2016 IEEE congress on evolutionary computation (CEC). IEEE; 2016.

[31] Biehl M. Biomedical applications of prototype based classifiers and relevance learning. In: Algorithms for computational biology. Springer International Publishing; 2017, p. 3–23.

[32] Mudali D, Biehl M, Leenders KL, Roerdink J. LVQ and SVM classification of FDG-PET brain data. Advances in intelligent systems and computing, vol. 428, Springer International Publishing; 2016.

[33] van Veen R, Tamboli NRB, Biehl M. Orthogonal learning correction. In: Machine learning reports. University of Applied Sciences Mittweida; 2021.

[34] LeKander M, Biehl M, de Vries H. Empirical evaluation of gradient methods for matrix learning vector quantization. In: 12th international workshop on self-organizing maps and learning vector quantization, clustering and data visualization (WSOM). IEEE; 2017, 8 pages.

[35] Papari G, Bunte K, Biehl M. Waypoint averaging and step size control in learning by gradient descent. In: Schleif F, Villmann T, editors. MIWOCI 2011, Mittweida workshop on computational intelligence. Technical report Mlr-2011-06, Leipzig University; 2011, p. 16–26,

[36] Biehl M, Bunte K, Schleif FM, Schneider P, Villmann T. Large margin linear discriminative visualization by matrix relevance learning. In: The 2012 international joint conference on neural networks. IJCNN, 2012, p. 1–8.

[37] Bunte K, Schneider P, Hammer B, Schleif FM, Villmann T, Biehl M. Limited rank matrix learning, discriminative dimension reduction and visualization. Neural Netw 2012;26:159–73.

[38] Biehl M, Hammer B, Villmann T. Prototype-based models in machine learning. Wiley Interdiscip Rev: Cogn Sci 2016;7(2):92–111.

[39] van Veen R, Meles SK, Renken RJ, Reesink FE, Oertel WH, Janzen A, de Vries G-J, Leenders KL, Biehl M. FDG-PET combined with learning vector quantization allows classification of neurodegenerative diseases and reveals the trajectory of idiopathic REM sleep behavior disorder. Comput Methods Programs Biomed 2022;225:107042.

[40] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. second ed.. Springer; 2009.

[41] van Veen R, Biehl M, de Vries JJG. Sklvq: Scikit learning vector quantization. J Mach Learn Res 2021;22(231):1–6.

[42] Garcia-Garcia D, Clavero P, Gasca Salas C, Lamet I, Arbizu J, Gonzalez-Redondo R, Obeso JA, Rodriguez-Oroz MC. Posterior parietooccipital hypometabolism may differentiate mild cognitive impairment from dementia in parkinson's disease. Eur J Nucl Med Mol Imaging 2012;39(11):1767–77.

[43] Teune LK, Bartels AL, de Jong BM, Willemsen ATM, Eshuis SA, de Vries JJ, van Oostrom JCH, Leenders K. Typical cerebral metabolic patterns in neurodegenerative brain diseases. Mov Disord 2010;25(14):2395–404.

[44] Arnaldi D, Morbelli S, Brugnolo A, Girtler N, Picco A, Ferrara M, Accardo J, Buschiazzo A, de Carli F, Pagani M, Nobili F. Functional neuroimaging and clinical features of drug naive patients with de novo parkinson's disease and probable RBD. Parkinsonism Rel Disord 2016;29:47–53.

[45] Della Rosa PA, Cerami C, Gallivanone F, Prestia A, Caroli A, Castiglioni I, Gilardi MC, Frisoni G, Friston K, Ashburner J, et al. A standardized [18 f]-FDG-PET template for spatial normalization in statistical parametric mapping of dementia. Neuroinformatics 2014;12(4):575–93.

[46] Teune LK, Renken RJ, Mudali D, Jong BMD, Dierckx RA, Roerdink JBTM, Leenders KL. Validation of parkinsonian disease-related metabolic brain patterns. Mov Disord 2013;28(4):547–51.

[47] Teune LK, Strijkert F, Renken RJ, Izaks GJ, de Vries JJ, Segbers M, Roerdink JBTM, Dierckx RAJO, Leenders KL. The alzheimer's disease-related glucose metabolic brain pattern. Curr Alzheimer Res 2014;11(8):725–32.

[48] Meles SK, Teune LK, de Jong BM, Dierckx RA, Leenders KL. Metabolic imaging in parkinson disease. J Nucl Med 2016.

[49] Eckert T, Tang C, Eidelberg D. Assessment of the progression of parkinson's disease: a metabolic network approach. Lancet Neurol 2007;6(10):926–32.

[50] Villmann T, Staps D, Ravichandran J, Saralajew S, Biehl M, Kaden M. A learning vector quantization architecture for transfer learning based classification in case of multiple sources by means of null-space evaluation. In: Bouadi T, Fromont E, Hüllermeier E, editors. Advances in intelligent data analysis XX. Cham: Springer International Publishing; 2022, p. 354–64.