IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 1, 2022, accepted February 11, 2022, date of publication February 18, 2022, date of current version March 4, 2022.

*Digital Object Identifier 10.1109/ACCESS.2022.3152789*

# Measuring Group Separability in Geometrical Space for Evaluation of Pattern Recognition and Dimension Reduction Algorithms

**ALDO ACEVEDO[1], CLAUDIO DURÁN[1], MING-JU KUO[1], SARA CIUCCI[1], MICHAEL SCHROEDER[2], AND CARLO VITTORIO CANNISTRACI[1,3]**

[1]Biomedical Cybernetics Group, Biotechnology Center (BIOTEC), Center for Molecular and Cellular Bioengineering (CMCB), Center for Systems Biology Dresden (CSBD), Department of Physics, Technische Universität Dresden, 01307 Dresden, Germany
[2]Bioinformatics Group, Biotechnology Center (BIOTEC), Center for Molecular and Cellular Bioengineering (CMCB), Technische Universität Dresden, 01307 Dresden, Germany
[3]Center for Complex Network Intelligence (CCNI), Tsinghua Laboratory of Brain and Intelligence (THBI), Department of Computer Science, Department of Biomedical Engineering, Tsinghua University, Haidian, Beijing 100084, China

Corresponding author: Carlo Vittorio Cannistraci (kalokagathos.agon@gmail.com)

The work of Carlo Vittorio Cannistraci was supported in part by the Independent Research Group Leader Budget of Biotechnology Center (BIOTEC), Technische Universität Dresden; in part by the Center for Information Services and High Performance Computing (ZIH), Technische Universität Dresden; in part by the Klaus Tschira Stiftung (KTS) GmbH, Germany, under Grant 00.285.20t16; in part by the Zhou Yahui Chair Professorship Award of Tsinghua University, China; in part the Tsinghua Laboratory of Brain and Intelligence, China; and in part by The High Level Talent Program of the Ministry of Science, China. The work of Claudio Durán was supported by the Research Grants–Doctoral Programs in Germany [German Academic Exchange Service (DAAD)], under Grant 57299294. Open Access funding enabled and organized by the Saxon State and University Library Dresden (SLUB).

**ABSTRACT** Evaluating group separability is fundamental to pattern recognition. A plethora of dimension reduction (DR) algorithms has been developed to reveal the emergence of geometrical patterns in a low-dimensional space, where high-dimensional sample similarities are approximated by geometrical distances. However, statistical measures to evaluate the group separability attained by DR representations are missing. Traditional cluster validity indices (CVIs) might be applied in this context, but they present multiple limitations because they are not specifically tailored for DR. Here, we introduce a new rationale called projection separability (PS), which provides a methodology expressly designed to assess the group separability of data samples in a DR geometrical space. Using this rationale, we implemented a new class of indices named projection separability indices (PSIs) based on four statistical measures: Mann-Whitney U-test p-value, Area Under the ROC-Curve, Area Under the Precision-Recall Curve, and Matthews Correlation Coefficient. The PSIs were compared to six representative cluster validity indices and one geometrical separability index using seven nonlinear datasets and six different DR algorithms. The results provide evidence that the implemented statistical-based measures designed on the basis of the PS rationale are more accurate than the other indices and can be adopted not only for evaluating and comparing group separability of DR results but also for fine-tuning DR algorithms' hyperparameters. Finally, we introduce a second methodological innovation termed trustworthiness, a statistical evaluation that accounts for separability uncertainty and associates to the measure of each index a p-value that expresses the significance level in comparison to a null model.

**INDEX TERMS** Pattern recognition, dimension reduction, data embedding, group separability, cluster validity indices.

## I. INTRODUCTION

One of the main current problems in data science, machine learning, and pattern recognition is to develop appropriate visual representations of complex data [1]–[4]. This has a large practical impact and implications, for instance, in the analysis of high-dimensional data in biomedicine (e.g., single-cell [5] or facial imaging analysis [6]) and neuroscience (e.g., visual object space representation in the brain [7] or the representation of pain patterns in brain networks [8]). Indeed, in high-dimensional datasets, the process of discovering patterns is further complicated by the fact

The associate editor coordinating the review of this manuscript and approving it for publication was Zahid Akhtar.

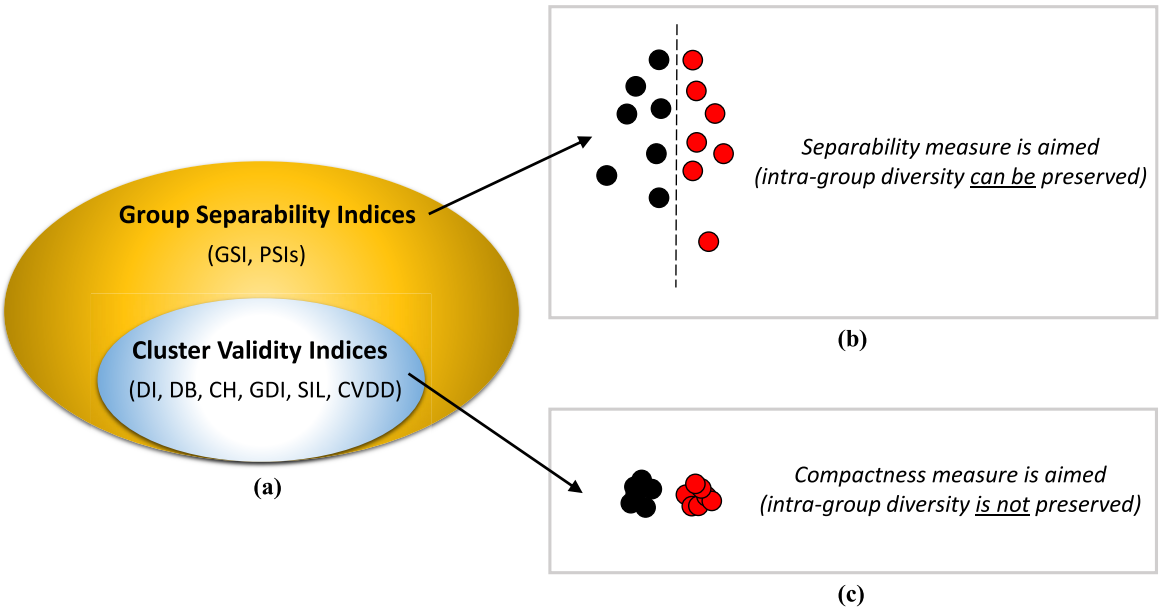



**FIGURE 1.** Group separability and cluster validity While evaluating data segregation, we can define two main concepts. In the blue circle, we have the cluster validity indices (CVIs) that focus their evaluation of data segregation by assessing (mainly) the compactness of the different groups (here, black versus red points). Since the congregation inside the same group is aimed, these indices tend to neglect the intra-group variability. In the yellow circle, we have the group separability indices (GrSIs), which on the other hand, focus their evaluation of data segregation by assessing (only) the separability of the different groups. Thus, indices based on this approach, such as the geometrical separability index (GSI) or the projection separability indices (PSIs) proposed in this study, can conserve the data's intra-group variability, an important feature to retain in the analysis of dimension reduction.

that data often cannot be immediately visualized to determine the similarity and separability of groups of samples. Thus, the development of different techniques for dimension reduction (DR) or data embedding has attracted considerable attention [9]–[11]. These techniques produce a low-dimensional representation where the geometrical distances between samples (data points) preserve the similarities of high-dimensional data together with their relevant structure [12].

Multiple DR algorithms have been developed, such as Principal Component Analysis (PCA) [13], [14], Multidimensional Scaling (MDS) [15], t-Distributed Stochastic Neighbor Embedding (t-SNE) [16], [17], Isometric Feature Mapping (Isomap) [18], Minimum Curvilinear Embedding (MCE) [2], Discriminative Sparse Embedding (DSE) [19], Uniform Manifold Approximation and Projection (UMAP) [20], among others. Despite attempts to preserve the original data structure, many of these algorithms may partially fail. For instance, in parameter-dependent DR algorithms, appropriate tuning of their hyperparameters to optimal values is essential; indeed, a misconfiguration of these inputs might result in a poor low-dimensional representation. The process can be further complicated if a selection between different ways to scale or normalize the data is required. Therefore, a measure or index that can quantify and evaluate the separation of groups of samples (group separability) according to the geometrical patterns revealed by these DR techniques is crucial. In this regard, with the help of the conceptual representation and toy examples presented in Fig. 1, we clarify the difference between two important concepts in data representation analysis: group separability and cluster validity. A measure of group separability aims to maximize the separation between two or more groups without any constraint on their level of internal diversity (Fig. 1b), which can therefore be preserved by dimension reduction representation. A measure of cluster validity is more stringent and requires not only maximizing group separability but also minimizing intra-group diversity to the extent that all members in a group converge to the same point and acquire the same label that represents a cluster assignment. In brief, the simultaneous constraint on these two features (inter-group separability and intra-group diversity) can be interpreted as the general aim of cluster validity to maximize the compactness of the groups (Fig. 1c). Hence, cluster validity measures are a subtype of group separability measure that aims to group compactness (Fig. 1a), which is a sufficient but not a required property for group separability (for details on the different methods to evaluate group separability and the theories behind them, please refer to the initial part of Methods section II.C.2). Indeed, cluster validity measures were designed for the evaluation of clustering [21], and therefore can be aberrant for the evaluation of dimension reduction where the mapping of intra-group diversity is important for an appropriate data representation. Nevertheless, in the past, clustering validity indices (CVIs) were routinely used also for the evaluation of dimension reduction because, as a matter of fact, they are group separability measures that can assess the compactness of groups in the dimension reduction results [22]–[24]. Hence, CVIs can be applied naïvely to validate the results of the DR techniques. Recent studies have relied on these approaches

to validate their results [25]–[27]. Therefore, we considered in this study six representative CVIs (Fig. 1a) for evaluating group separability of DR results such as Dunn index (DI) [28], which relies on the distances among clusters and their diameters; Davies-Bouldin index (DB) [29] based on the idea that for a good partition inter-cluster separation as well as intra-cluster homogeneity and compactness should be high; Calinski-Harabasz index (CH) [30] based on the average between-cluster means and within-cluster sum of squares; Silhouette index (SIL) [31], which validates the clustering performance based on the pairwise difference of between-cluster and within-cluster distances; Generalized Dunn Index (GDI) one of the most reliable improved variants of the Dunn Index described in [32]; and Cluster Validity index based on Density-involved Distance (CVDD) [33], a modern index sensitive to density-separated clusters. Moreover, we included the Geometrical Separability Index (GSI) – also known as Thornton's separability index – [34], which is not a CVI but a geometrically-driven separability approach (Fig. 1a) that calculates the proportion of instances that share the same class label as their first nearest neighbor.

However, our experiments show that the majority of these indices are not suitable for validating the separability given by DR techniques because: i) most of them may not have any relation to the geometrical structure of the data [35] because they mainly focus on assessing the compactness instead of separability (compare Fig. 1a and Fig. 1b); thus, they tend to neglect the intra-group diversity often given by DR results; ii) several of them are boundless (i.e., they do not have an upper or lower bound); thus, they are extremely sensitive to changes in scale, and they are not tailored for comparison of results across different dimensions (for instance, the value of an index applied in two dimensions is not directly comparable with the value of the same index applied in three dimensions); and iii) the few bounded indices do not always address the problem of group separability adequately in a geometrical space (for instance, they are affected by the presence of artificial microclusters). In addition, other limitations have been documented for some CVIs: they can be hardly affected by group overlapping and sample outliers (e.g., DB and SIL, among others) [22], [36], [37]; they can be sensitive to clusters with arbitrary shapes (e.g., their evaluation can be affected while dealing with nonspherical datasets) especially for high-dimensional data (this issue has been encountered with CH, DB, and DI, among others) [33], [35], [38]. We clarify that an arbitrary cluster shape does not necessarily imply nonlinear separability. For instance, it might occur that two clusters have a nonspherical shape but are still linearly separable. Hence, the robustness to nonlinear separability in the presence of different levels of isotropic and anisotropic noise (i.e., outliers) was specifically investigated in this study. Table 1 summarizes the characteristics discussed above (plus some new ones that we will discuss in this study) and reports for each of these indices whether they can satisfy or not some of these fundamental proprieties in the evaluation of group separability. Hence, the investigation of suitable measures for quantifying and evaluating group separability revealed by DR results remains a salient open research topic.

Here, we propose a novel rationale named projection separability (PS), which is specifically designed to assess the group separability of data samples in a geometrical space, such as dimension reduction analyses based on embedding algorithms. Our PS rationale provides a new methodology that states that any statistical separability measure that is bounded or that can be bounded with an adequate transformation function, such as the ones commonly used to measure the performance of a binary classification model, can be used to evaluate the group separability of dimension reduction results based on the geometrical projection of the samples (data points) of two different groups on their projection separability line. The projection separability line is a line that, given two groups of samples in a multidimensional geometrical space, provides a one-dimensional geometrical representation of the data separating the two groups. This projection separability line can be defined according to the different principles, methods, or criteria discussed in the Methods section II.2.C. Then, repeating this procedure for all pairs of groups, for which a separability evaluation is desired, the average separability is considered as the final estimation.

Based on this new rationale, we implemented four statistically driven separability measures called projection separability indices (PSIs). The first index, PSI-P, evaluates group separability using the Mann-Whitney U-test p-value (MWp-value) [39], which is a ranking-based statistical test. The second index, PSI-ROC, performs this evaluation by applying the Area Under the ROC-Curve (AUC-ROC) [40], which provides a measure of the trade-off between the true positive and false positive rates. The third index, PSI-PR, uses the Area Under the Precision-Recall Curve (AUC-PR) [41], which provides a measure of the trade-off between precision and sensitivity (also known as recall). And the fourth index, PSI-MCC, implements the Matthews Correlation Coefficient (MCC) [42], which provides a correlation coefficient between the observed and predicted binary classifications. It should be noted that because these indices are based on bounded statistical measures, they inherit these boundaries. In addition, we introduce a second methodological innovation termed *trustworthiness*, which is a statistical evaluation that accounts for uncertainty and associates to the separability measure provided by each index an empirical p-value that expresses the extent to which this measure is statistically significant in comparison to a null model.

We compared the six representative CVIs and the geometrically-driven separability index GSI against our approaches and demonstrated the effectiveness of the PSIs in evaluating the group separability given by six different dimension reduction techniques on three synthetic and four real (high-dimensional) datasets.

## II. METHODS

Six representative CVIs and one geometrically-driven separability index, which are described in this section, were

**TABLE 1.** Indices properties.

| Index | Characteristics | | | Effective for | | | Robustness to | | |
|---|---|---|---|---|---|---|---|---|---|
| | Objective function | Base on | Bounded | Overlapping | Arbitrary shapes | Linearity detection | Nonlinear (curvilinear) pattern | Isotropic noise | Anisotropic noise (outliers) |
| **DI** [28] | Cluster validity | Distances | - | - | - | - | - | - | - |
| **CH** [30] | Cluster validity | Distances | - | ✓ | - | - | - | - | - |
| **DB** [29] | Cluster validity | Distances | - | - | - | - | - | ✓ | ✓ |
| **GDI** [32] | Cluster validity | Distances | - | - | - | - | - | ✓ | ✓ |
| **SIL** [31] | Cluster validity | Distances | ✓ | ✓ | - | - | - | ✓ | - |
| **CVDD** [33] | Cluster validity | Density | - | - | ✓ | - | - | - | - |
| **GSI** [34] | Group Separability | Nearest Neighbors | ✓ | - | ✓ | - | ✓ | - | ✓ |
| **PSI-P** | Group Separability | PS + Statistics (MW p-value) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| **PSI-ROC** | Group Separability | PS + Statistics (AUC-ROC) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| **PSI-PR** | Group Separability | PS + Statistics (AUC-PR) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| **PSI-MCC** | Group Separability | PS + Statistics (MCC) | ✓ | ✓ | ✓ | ✓ | - | ✓ | ✓ |

*This table presents an overview of the different group separability indices considered in this study. The reference to the scientific article of each index is reported in a square bracket close to its acronym. The symbol ✓ means: Yes, the index can achieve a certain property, and the symbol − means: No, the index cannot achieve a certain property.*

compared with our PSIs. For comparison, we used two synthetic datasets to explain the technicalities and set the basis of the proposed methodology. One synthetic and four real high-dimensional datasets were analyzed using six different dimension reduction (DR) algorithms, considering the first two dimensions of mapping (2D), and in some cases, the first three dimensions of mapping (3D). In general, DR techniques can efficiently reduce the features/variables space to a much smaller number of dimensions without a significant loss of information [43]. We applied these DR techniques based on the idea that the indices may pinpoint different best-performing DR methods (i.e., different methods with the best group separability) owing to the diversity of the datasets. The DR results may change when considering datasets with a small number of samples (few data points), unbalanced sample groups, or noisy data. Under these contexts, some indices may report an incorrect evaluation of the DR methods, selecting those with a nonsignificant or inaccurate group separability as the best.

The applied DR methods are both linear and nonlinear because linear approaches might fail to represent hidden nonlinear relations among the samples in the features/variables space. The techniques used are as follows: Principal Component Analysis (PCA) [13], [14], Non-metric Multidimensional Scaling with Sammon Mapping (NMDS) [44], [45], Multidimensional Scaling based on Bray-Curtis dissimilarity (MDSbc) [46], Minimum Curvilinearity Embedding (MCE) [2], Isometric Feature Mapping (Isomap) [18],

**TABLE 2.** Dataset details.

| Dataset | Samples/Data Points | Features/Variables | Classes/Groups | Type | Main Problem |
|---|---|---|---|---|---|
| **Apple-Stem** | 522 | 2 | 2 | Synthetic | Arbitrary shape |
| **Half-moons** | 1500 | 2 | 2 | Synthetic | Nonlinear (curvilinear) pattern with isotropic/anisotropic noise |
| **Tripartite-Swiss-Roll** | 723 | 3 | 3 | Synthetic | Nonlinearity and discontinuity |
| **Gastric mucosa microbiome** | 24 | 187 | 3 | Real | Small sample size |
| **Radar signal** | 350 | 34 | 3 | Real | Crowding problem |
| **Image proteomics** | 42 | 1947 | 3 | Real | Curse of dimensionality |
| **MNIST** | 300 | 784 | 10 | Real | Nonhierarchical structure |

*The first two datasets are synthetic and represent emblematic cases of arbitrary shapes and nonlinear pattern with isotropic/anisotropic noise. Since they are low dimensional, the application of dimension reduction (DR) methods was not necessary. On the other hand, the third synthetic dataset and all the real datasets are high dimensional. Hence, different dimensional-reduction (DR) methods were applied to embed each dataset into the first two dimensions (2D) or three dimensions (3D) of mapping.*

and t-Distributed Stochastic Neighbor Embedding (t-SNE) [16], [17]. Owing to the high dimensionality and diversity of the datasets, each method can produce a different partitioning or grouping of samples. Moreover, linear approaches might suffer from the absence of sample group separation in the low-dimensional space of representation, which does not necessarily imply the absence of separation in general. Indeed, nonlinear techniques for DR may instead prove it. For this reason, the role of the indices is key to discerning and quantitatively assessing which DR methods provide the best group separability of the samples. Because this study focuses on the separability indices, we do not provide any further explanation of the DR methods (see the corresponding references for more details).

The code for computing the proposed PSIs is available in MATLAB at https://github.com/biomedical-cybernetics/projection-separability-indices and as a Python package at https://pypi.org/project/psis.

### A. DATASET DESCRIPTION

We considered three artificial and four real datasets to compare the proposed PSIs with other indices. It is important to mention that the datasets employed in this study present a nonlinear structure. This selection was made to confirm whether our PSIs can correctly evaluate how the DR methods accurately detect nonlinearity within the data. A complete description of the number of samples, features/variables, and classes/groups for each dataset is presented in Table 2. Their source files are available at https://github.com/biomedical-cybernetics/projection-separability-indices.

#### 1) APPLE-STEM

The first artificial dataset was created using a web tool (available at https://guoguibing.github.io/librec/datagen.html) that allows data points to be drawn in a two-dimensional space and automatically obtains the coordinates and classes associated with each sample. It provides a test scenario with an arbitrary shape, as shown in Fig. 2, where two main groups are drawn: an apple (red) containing 201 data points and its stem (green) containing 321 data points.

#### 2) HALF-MOONS

A second synthetic dataset was created using the Scikit-learn Python module [47]. This dataset contains two interleaving half-circles, commonly named ''half-moons'' (Fig. 3 and Fig. 4). The dataset is defined in a two-dimensional (2D) space and consists of two groups: a red half-moon and a green half-moon, each containing a total of 750 data points. This offers an emblematic example of nonlinear separability (curvilinear) because the two half-moon groups cannot be linearly separated by a line. In addition to the difficulty of analyzing its nonlinearity, we used this dataset to evaluate the stability of the separability indices under two specific contexts.

The first scenario has different increasing levels of isotropic noise (Fig. 3d-n, where six noise levels are considered on the x-axis: 0, 0.075, 0.15, 0.225, 0.3, and 0.375). For its creation, we used the function described in https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_moons.html. As input, we passed the total number of samples (1500 data points), the respective noise

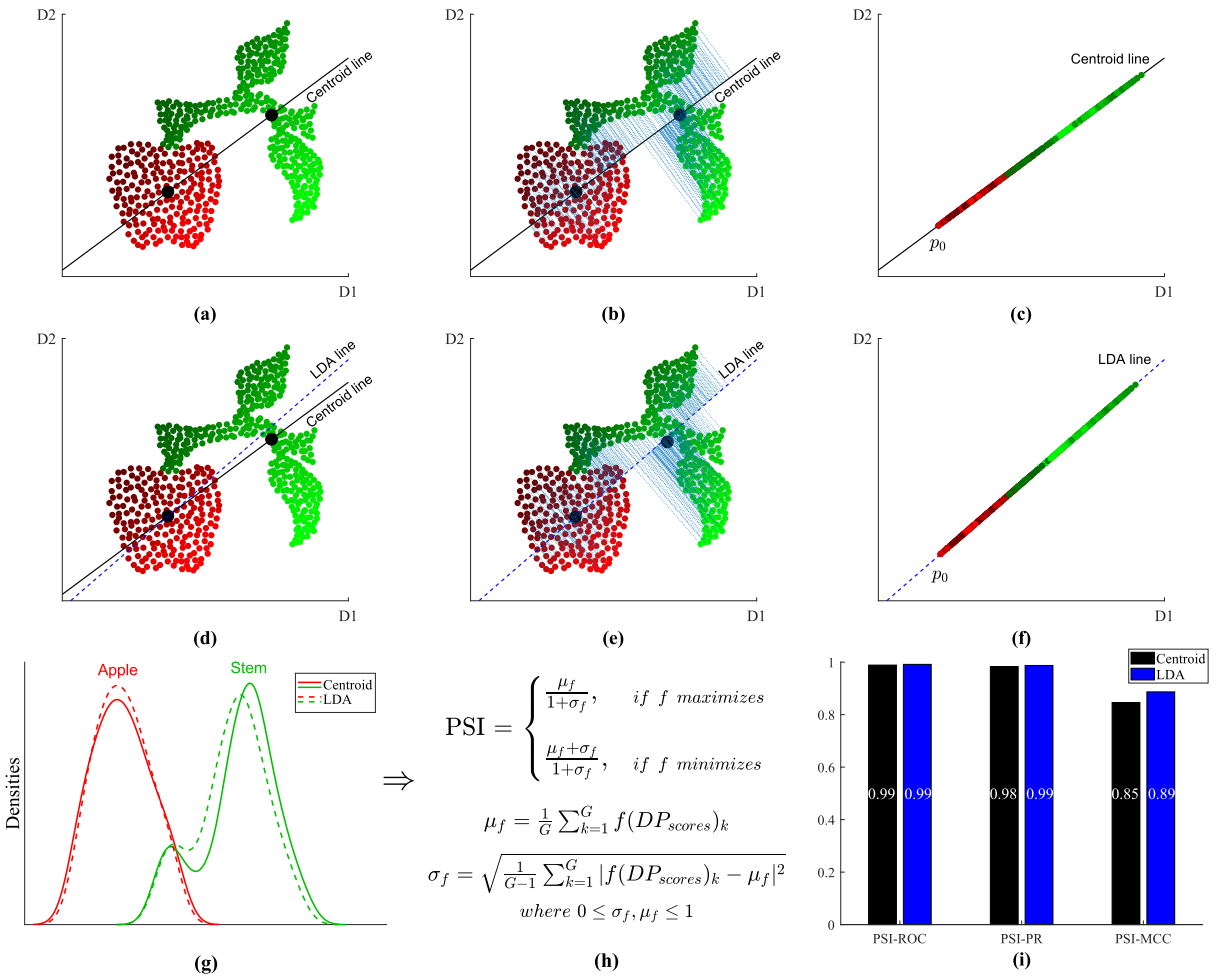



**FIGURE 2.** Workflow for implementing a projection separability index (PSI) based on the projection separability (PS) rationale. In this example, we can visually identify two groups; the apple (regular cluster) in red and its stem(irregular cluster) in green. Two different ways of implementing a projection separability index are represented: 1) Centroid separability line: In (a), the implementation of a PSI starts by determining the centroids (black points) of the clusters (a), then by projecting the sample points on the line that connects the cluster centroids (b). 2) Linear Discriminant Analysis (LDA) separability line: In (d), the implementation of a PSI starts by computing LDA, then by projecting the sample points on the line formed by the first LDA discriminant. Once projected, either via centroid line or LDA line, the sample points are transformed from *D*-dimensional space to 1D space (also referred to as separability line) by fixing one extreme point as reference ($p_0$) and taking the distance from it to the rest of the other points ($DP_i = d(p_0, p_i)$) (c) and (f). Then, considering the array of all distances ($DP_{scores}$) and the labels (green and red), any statistical separability measure $f(\cdot)$ – for instance, the ones commonly used to measure the performance of a classification model - can be applied for evaluating pairwise group separability (h). In the presence of more than two groups (a condition that for simplification is not contemplated in the current figure), the procedure is repeated for all pairs of groups ($G$) in the dataset. Finally, assuming that $f$ is a bounded measure, the respective projection separability index (PSI) is calculated as an overall estimation based on the mean $\mu_f$ and the standard deviation $\sigma_f$ (penalty factor) of the obtained results. For instance, if $f$ maximizes (i.e., the best group separability is given by its upper bound), the first PSI expression is applied. On the other hand, if $f$ minimizes (i.e., the best group separability is given by its lower bound), then the second PSI expression is applied. In (i), three statistical measures were implemented based on the Area Under the ROC-Curve (PSI-ROC), Area Under the Precision-Recall (PSI-PR), and the Matthews Correlation Coefficient (PSI-MCC) for both centroid line (black) and LDA line (blue).

level (each one of the values mentioned previously), and a random seed state equal to 1 (for reproducibility). This procedure generated six variants of half-moons where the injection of isotropic noise in the two clusters "fuzzyfies" their borders and therefore represents an appropriate procedure to test the stability of the group separability indices when the clusters are geometrically perturbed. Moreover, it is important to mention that isotropic noise should not be confused with overlapping because groups with isotropic noise do not necessarily overlap. However, it might occur in specific cases that groups start to overlap their borders when increasing the level of isotropic noise.

The second scenario was generated based on the second noise variant (with an isotropic noise level of 0.075). Here, we introduced anisotropic noise, which can be interpreted as outliers (Fig. 4d-n, where six outlier levels are considered in the x-axis: 0%, 2%, 4%, 6%, 8%, and 10% of the original size). For its creation, each anisotropic noise percentage was progressively added (half-percentage for each group; for instance, 2% means that 1% of outliers were added to each group, with regard to their original size), and the resulting outliers were randomly placed on top of each half-moon. Six variants were generated, where the addition of anisotropic noise was used to test the stability of the separability indices

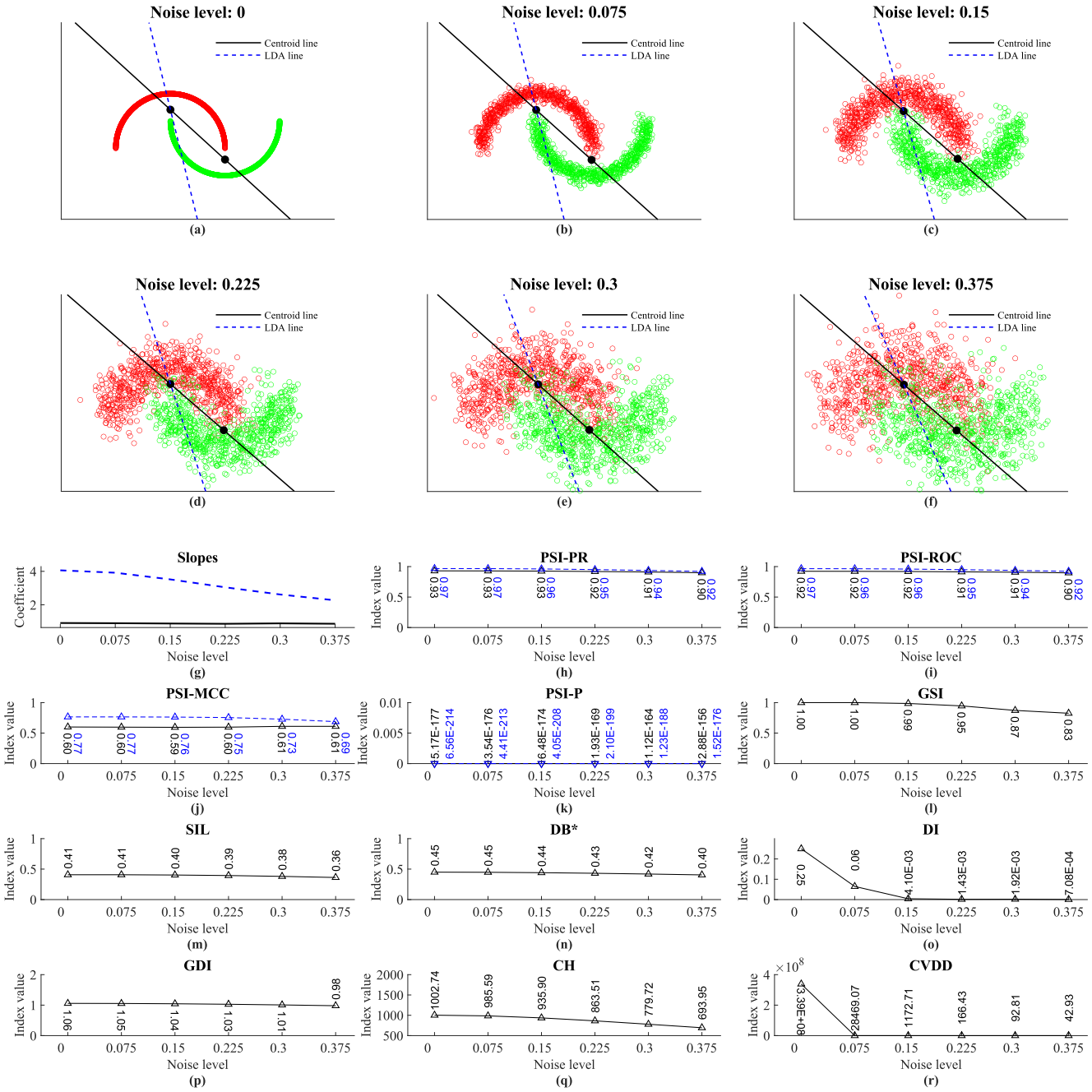



**FIGURE 3.** Half-moons with isotropic noise. In this example, we can visually identify two interleaving half-circles commonly called "half-moons" (a); the top one in red, and the bottom one in green. Different increasing isotropic noise levels were injected, which "fuzzyfies" the groups' borders and represents an appropriate scenario to test the stability of the group separability indices when clusters are geometrically perturbed (b-f). Also, in (a-f), the two main approaches for computing the separability line when implementing a projection separability index (PSI) are represented; the centroid line (black line) and the Linear Discriminant Analysis (LDA) line (blue dashed line). These projection separability lines are differently perturbed by the injection of the isotropic noise, which is represented by the variations on their slope coefficient (g). Different PSIs were implemented for both projection separability lines and used to evaluate the group separability of each noise variant (h-k). Also, the other indices were applied (l-r) to see how their evaluation of separability is perturbed by the injection of noise.

when certain points are placed far from the group to which they belong.

Both isotropic and anisotropic noise scenarios are defined in a two-dimensional (2D) space; thus, they are evaluated without the need to apply any dimension reduction method.

### 3) TRIPARTITE-SWISS-ROLL

The synthetic Tripartite-Swiss-Roll dataset is proposed here by discretizing the Swiss-Roll [18] manifold in a three-dimensional space (3D). It contains 723 sample points divided into three groups characterized by nonlinear

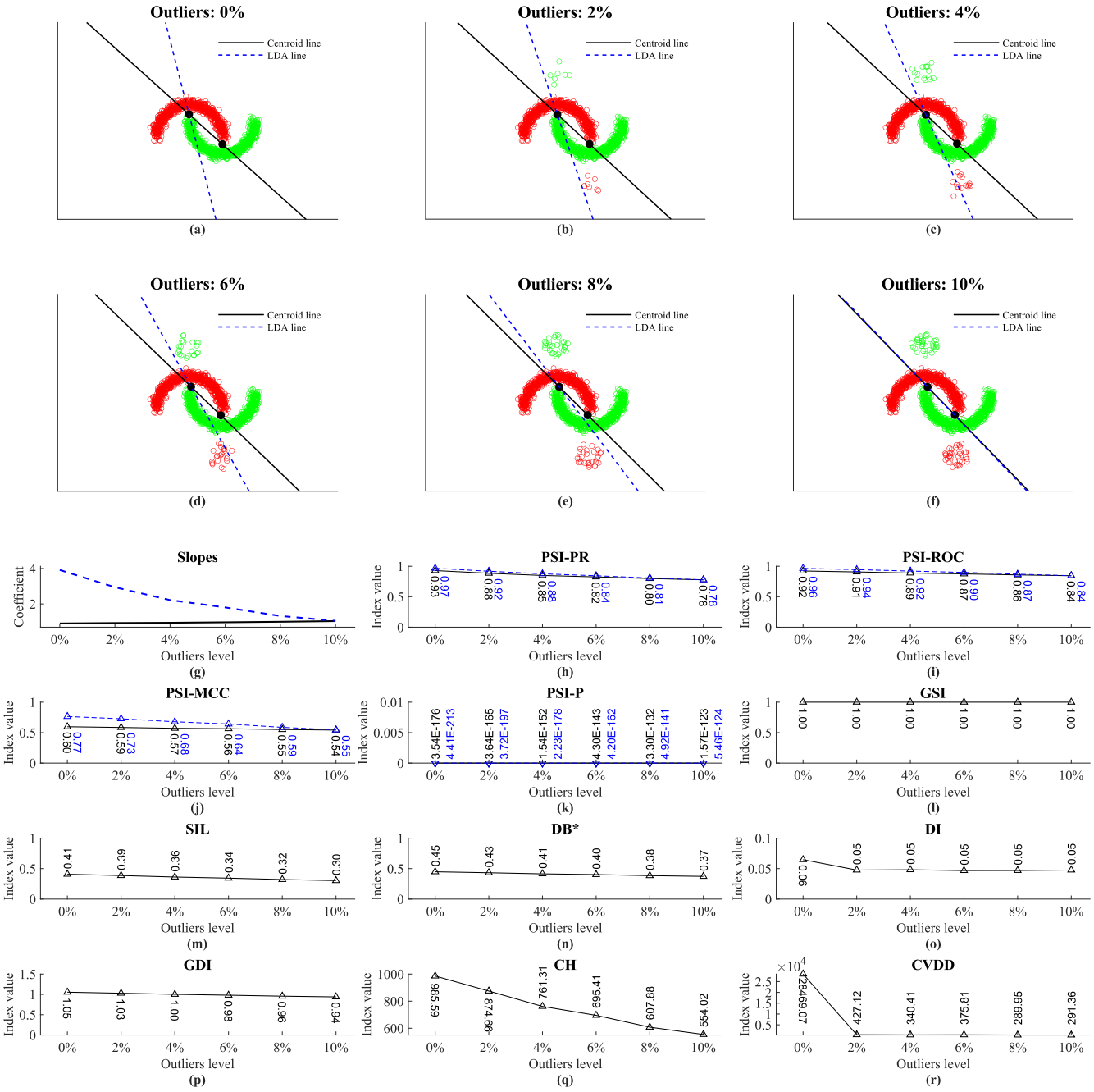



**FIGURE 4.** Half-moons with anisotropic noise (outliers). In this example, we can visually identify two interleaving half-circles commonly called "half-moons" (a); the top one in red and the bottom one in green. Different increasing anisotropic noise (outliers) levels were injected, which represents a scenario to test the stability of the group separability indices when points are misplaced (outside) from the group they belong (b-f). Also, in (a-f), the two main approaches for computing the separability line when implementing a projection separability index (PSI) are represented; the centroid line (black line) and the Linear Discriminant Analysis (LDA) line (blue dashed line). These projection separability lines are differently perturbed by the injection of the anisotropic noise, which is represented by the variations on their slope coefficient (g). Different PSIs were implemented for both projection separability lines and used to evaluate the group separability of each outliers variant (h-k). Also, the other indices were applied (l-r) to see how their evaluation of separability is perturbed by the injection of outliers.

structures. The difficulty in analyzing this dataset lies in the fact that the typical nonlinearity of the Swiss-Roll shape is further impaired by the discontinuity generated by the manifold's tripartition, which is challenging to retain in the 2D representation obtained by dimension reduction.

#### 4) GASTRIC MUCOSA MICROBIOME

An actual 16S rRNA gene sequence dataset was studied by Sterbini *et al.* [48]. It is publicly available at the NCBI Sequence Read Archive (SRA) (http://www.ncbi.nlm.nih.gov/sra, accession number SRP060417), where all details pertaining to the experimental sequencing design are also

reported. It contains 24 biopsy specimens of the gastric antrum from 24 individuals referred to the Department of Gastroenterology of Gemelli Hospital (Rome) with dyspepsia symptoms (heartburn, nausea, epigastric pain and discomfort, bloating, and regurgitation). In this dataset, 12 of these individuals had been taking Proton Pump Inhibitors (PPIs) for at least 12 months (PPIs are increasingly being used for the treatment of the most frequent causes of dyspepsia, besides being part of anti-*Helicobacter pylori* treatment regimens [48]), whereas the others were not being treated (naïve) or had stopped their treatment at least 12 months before sample collection. In addition, 9 individuals (4 treated and 5 untreated) were positive for *H. pylori* infection, and *H. pylori* positivity or negativity was determined by histology and rapid urease tests. Metagenomes were obtained by pyrosequencing the 16S rRNA gene fragments on a GS Junior platform (454 Life Sciences, Roche Diagnostics). Then, the sequence data were processed, replicating the bioinformatics workflow followed by Sterbini *et al*. [48], to obtain the matrix of absolute bacterial abundance. The resulting dataset comprises 24 samples, 187 features, and three groups: untreated dyspeptic patients without *H. pylori* infection (HPneg), untreated dyspeptic patients with *H. pylori* infection (HPpos), and patients treated with Proton Pump Inhibitors (PPI). The difficulty in analyzing this dataset lies in the small sample size compared to the number of features. Under this condition, the data becomes highly sparse, and consequently, it is difficult to discern the actual grouping of samples because the algorithms for dimension reduction encounter issues in appropriately approximating and mapping the hidden geometry of the data.

#### 5) RADAR SIGNAL

A second real dataset was recovered from the UC Irvine Machine Learning Repository (available at http://archive.ics.uci.edu/ml/datasets/Ionosphere). This dataset has been widely described and analyzed in [49]. It contains 351 radar signals targeting free electrons in the ionosphere. Shieh *et al.* [50] studied this dataset using two labeled groups (good and bad radar signals). However, they highlighted that good radar signals are highly similar and bad radar signals are highly dissimilar. Later, Cannistraci *et al.* [51] confirmed that the bad radar signals can be interpreted as two diverse subcategories (two different groups) that are difficult to identify because of their high nonlinearity (elongated and irregular high-dimensional structure). In [51], it was reported that the primary difficulty of this dataset is the crowding problem. This means that after DR, the different groups of samples tend to collapse on top of each other (highly overlapping) in the reduced dimensional space. Thus, evaluating the correct group separability is challenging. Based on these results, we used three labeled groups: good radar signal, bad radar signal 1, and bad radar signal 2.

In the first stage of the analysis, we detected two samples (radar signals) with the same values across all features/variables, so we removed one of them to avoid problems with the calculation of the dissimilarity matrix employed by NMDS. Therefore, the resulting dataset is composed of 350 samples, 34 features, and three groups.

#### 6) IMAGE PROTEOMICS

We used a proteomic dataset obtained from [2]. It was generated by combining a dataset from 2D Electrophoresis (2DE) gel images derived from proteomic Cerebrospinal fluid (CSF) samples of peripheral neuropathic patients [52] and another dataset derived from a neurological study of amyotrophic lateral sclerosis (ALS) patients not affected by neuropathic pain [53]. The resulting dataset contains four main groups divided into healthy control patients (C) with eight samples, patients not affected by neuropathic pain (M) with 19 samples, patients without pain (NP) with eight samples, and patients with a pathological variant of pain (P) with seven samples. In [2], the authors discovered that four patients without pain developed pain after a clinical follow-up at 6-12 months (or $>$ 1 year). Hence, we considered these four patients in the pathological variant with pain (P) group, resulting in 11 samples. The remaining four patients in the NP group were included in the healthy control patient class (C group), ending with 12 samples. Therefore, this dataset is composed of 42 samples, 1947 features/variables, and three groups. Note that this dataset represents a real example of a frequent problem in biomedical datasets, where the number of samples is considerably smaller than the number of features/variables [54] (also known as ''the curse of dimensionality'' [55]). This can affect the performance of the DR methods; thus, it challenges the evaluation given by different separability indices.

#### 7) MNIST

We also included a well-known dataset in the machine learning field called MNIST [56]. This is a large dataset consisting of $28 \times 28$ pixel images of handwritten digits. Every image can be thought of as a 784-dimensional array, where each value represents the intensity of each pixel in greyscale. The different sample groups are numbers between 0 and 9 (i.e., 10 different groups). The number of samples for each group is 980 samples for digit 0, 1135 samples for digit 1, 1032 samples for digit 2, 1010 samples for digit 3, 982 samples for digit 4, 892 samples for digit 5, 958 samples for digit 6, 1028 samples for digit 7, 974 samples for digit 8, and 1009 samples for digit 9; therefore, the total amount of samples in the dataset is 10000.

This dataset has the peculiarity that it does not present a hierarchical organization of the samples. This is explained by the uniqueness of different handwriting styles, where the same digits can be highly dissimilar and different digits can be highly similar. Thus, this dataset can be particularly challenging for dimension reduction methods.

This is an exceptionally large dataset. We want to explore a high number of hyperparameters combinations for Isomap and t-SNE, along with an evaluation of the separability significance (trustworthiness) of the indices for each of these

hyperparameters combinations (this is further explained in Section II.D). Hence, performing these computations considering the entire number of samples in the dataset is not feasible within an acceptable time frame. To reduce the computational time and preserve the validity of the tests applied to this dataset, we randomly selected 30 samples of each digit, resulting in a subdataset with a total of 300 samples, 784 features, and 10 groups.

### B. CLUSTER VALIDITY INDICES (CVIs)

This section describes the six representative cluster validity indices (which are a specific subtype of group separability measures based on compactness estimation, see Fig. 1c) used in this study to assess the group separability of DR results. To facilitate their understanding, we have reported and described their mathematical formulations.

#### 1) SILHOUETTE INDEX

The Silhouette index (SIL) [31] is a measure used for fuzzy clustering validation based on the concept of *silhouette width*, which is defined as follows:

$$\mathrm{SIL} = \frac{1}{N}\sum_{k=1}^{N} \mathrm{SW}\left(C_k\right) \tag{1}$$

where $N$ is the number of clusters and $\mathrm{SW}\left(C_k\right)$ represents the *silhouette width* for the $k$th cluster, calculated as:

$$\mathrm{SW}\left(C_k\right) = \frac{1}{n_k}\sum_{x\in C_k} \mathrm{S}(x) \tag{2}$$

where $n_k$ is the number of data points in the $k$th cluster $C_k$, and $\mathrm{S}(x)$ is the *silhouette width* of sample $x$, which can be expressed as:

$$\mathrm{S}\left(x\right) = \frac{b\left(x\right) - a(x)}{\max(a\left(x\right), b\left(x\right))} \tag{3}$$

Here, $a(x)$ represents the within-cluster mean distance, defined as the average distance between $x$ and the remaining samples belonging to the same cluster. In contrast, $b(x)$ is the smallest of the mean distances of $x$ to the samples belonging to each of the other clusters.

This index takes values between $-1$ and 1, where 1 indicates the best partitioning of the data (i.e., the best group separability).

#### 2) CALINSKI-HARABASZ INDEX

The Calinski-Harabasz index (CH) [30] is based on the ratio between the overall between-cluster distance and the overall within-cluster distance, and is defined as:

$$\mathrm{CH} = \frac{\mathrm{SS}_B}{\mathrm{SS}_W} \times \frac{T-N}{N-1} \tag{4}$$

where $N$ is the number of clusters and $T$ is the total number of data points. $\mathrm{SS}_B$ is the overall between-cluster variance involving the elements of different clusters and is denoted by:

$$\mathrm{SS}_B = \sum_{k=1}^{N} n_k \left\|z_k - q\right\|^2 \tag{5}$$

where $n_k$ is the number of data points in cluster $k$, $z_k$ is the centroid of the $k$th cluster (centroid, which is obtained by taking the mean value of all points within the cluster), $q$ is the overall centroid of the data (i.e., the dataset's grand centroid), and $\|\cdot\|$ denotes the Euclidean distance between $z_k$ and $q$. In (4), $\mathrm{SS}_w$ is the overall within-cluster variance calculated as follows:

$$\mathrm{SS}_W = \sum_{k=1}^{N}\sum_{x\in C_k} \left\|x - z_k\right\|^2 \tag{6}$$

where $x$ is a data point belonging to the $k$th cluster $C_k$, $z_k$ represents the centroid of cluster $k$, and $\|\cdot\|$ denotes the Euclidean distance between $x$ and $z_k$.

For clarity, good clustering has a large overall between-cluster distance ($\mathrm{SS}_B$) and a small overall within-cluster distance ($\mathrm{SS}_W$) [57]. Thus, the best partitioning of the data is achieved by a high ratio of $\mathrm{SS}_B/\mathrm{SS}_W$, that is, a higher value of the CH index represents a better grouping of the samples.

#### 3) DUNN INDEX

The Dunn index (DI) [28] relies on the distances between clusters and cluster diameters. It uses the minimum pairwise distance between samples in different clusters as the inter-cluster separation and the maximum diameter among all clusters as the intra-cluster compactness. This index is calculated as follows:

$$\mathrm{DI} = \min_{k=1,\ldots,N}\left\{\min_{\substack{l=1,\ldots,N\\ l\neq k}}\left(\frac{\delta(C_k, C_l)}{\max\limits_{m=1,\ldots,N}\Delta(C_m)}\right)\right\} \tag{7}$$

In (7), $N$ represents the number of clusters, $\delta\left(C_k, C_l\right)$ represents the minimum distance between clusters $C_k$ and $C_l$ (i.e., the dissimilarity between clusters), which is described as follows:

$$\delta\left(C_k, C_l\right) = \min_{x\in C_k, y\in C_l} d(x, y) \tag{8}$$

where $d(x, y)$ denotes the distance (here Euclidean) between points $x$ and $y$. $\Delta\left(C_m\right)$ is the diameter of the $m$th cluster, represented as:

$$\Delta\left(C_m\right) = \max_{x,y\in C_m} d(x, y) \tag{9}$$

If DI is large, it means that compact and well-separated clusters exist. Thus, the highest value of this index represents the best grouping of samples.

#### 4) GENERALIZED DUNN INDEX

In [32], Bezdek and Pal recognized that DI is very noise sensitive, and proposed several generalizations of this index for clustering validation. Specifically, they proposed five variations in distances and three additional definitions of cluster diameters. Here, we focused on the third distance and diameter variations, which, as concluded by the authors, were the most reliable among the proposed variants [32]. For clarity, we designated the index created out of these variants as the Generalized Dunn Index (GDI).

In particular, the mentioned distance variation concerns (8), which is now defined as:

$$\delta\left(C_k, C_l\right) = \frac{1}{|C_k|\,|C_l|} \sum\nolimits_{x \in C_k, y \in C_l} d(x, y) \tag{10}$$

where $|C_k|$ and $|C_l|$ denote the number of points in the respective clusters and $d(x, y)$ is the distance between the points.

The diameter variation concerns (9), which is now denoted as:

$$\Delta\left(C_m\right) = 2\left(\frac{\sum_{x \in C_m} d(x, z_m)}{|C_m|}\right) \tag{11}$$

where $|C_m|$ denotes the number of points of cluster $m$, $z_m$ represents the centroid of the $m$th cluster, and $d(x, z_m)$ is the distance between point $x$ and centroid $z_m$. Then, $\delta\left(C_k, C_l\right)$ and $\Delta\left(C_m\right)$ are applied, as in (7). Thus, the highest value of this index represents the best-estimated grouping of the samples.

#### 5) DAVIES-BOULDIN INDEX

The Davies-Bouldin index (DB) [29] identifies clusters that are far from each other and compact. This index is calculated as follows:

$$\mathrm{DB} = \frac{1}{N} \sum\nolimits_{k=1}^{N} \max_{k \neq l} \left\{ \frac{\Delta\left(C_k\right) + \Delta(C_l)}{\delta(C_k, C_l)} \right\} \tag{12}$$

In (12), $N$ denotes the number of clusters, $\delta(C_k, C_l)$ is the inter-cluster distance, and $\Delta(C_k)$ and $\Delta(C_l)$ represent the diameters of the clusters $k$ and $l$, respectively. In particular, the diameter $\Delta(C_k)$ (and $\Delta(C_l)$ in the same way, but in relation to the $l$th cluster) can be calculated as:

$$\Delta(C_k) = \frac{1}{n_k} \sum\nolimits_{x \in C_k} d(x, z_k) \tag{13}$$

where $n_k$ is the number of data points of the $k$th cluster, and $d(x - z_k)$ is the distance between point $x$ and the centroid $z_k$ of the cluster (i.e., being their sum, the intra-cluster distance). Moreover, in (12), the inter-cluster distance $\delta(C_k, C_l)$ is defined as follows:

$$\delta\left(C_k, C_l\right) = d(z_k, z_l) \tag{14}$$

where $d(z_k, z_l)$ is the distance (here Euclidean) between the centroids of clusters $k$ and $l$.

In this case, the smallest value of this index indicates the best partitioning of the data (i.e., the clusters are considered to be optimally separated from each other). To facilitate the comparison between the different indices in this study, because DB is the only index where the minimum value represents the best grouping of the samples, we inverted its output using the following formula:

$$\mathrm{DB}^* = \frac{1}{1 + \mathrm{DB}} \tag{15}$$

where DB is the original index value returned in (12). Thus, a higher value of $\mathrm{DB}^*$ represents a better grouping of the samples.

#### 6) CLUSTER VALIDITY INDEX BASED ON DENSITY-INVOLVED DISTANCE

The Cluster Validity index based on Density-involved Distance (CVDD) [33] takes into account two key concepts for addressing separability: core objects and density connectivity. The first concept helps this index deal with outliers, whereas the second allows it to differentiate density-separated clusters. This index is defined as:

$$\mathrm{CVDD} = \frac{\sum_{k=1}^{N} \mathrm{sep}(C_k)}{\sum_{k=1}^{N} \mathrm{com}(C_k)} \tag{16}$$

where $N$ is the number of clusters, $\mathrm{sep}(C_k)$ is the separation between cluster $C_k$ and other clusters, and $\mathrm{com}(C_k)$ is the compactness of cluster $C_k$. In particular, $\mathrm{sep}(C_k)$ is defined for a $\pi$ set of clusters as:

$$\mathrm{sep}\left(C_k\right) = \min_{x_l \in \pi, x_l \in C_k} \mathrm{sep}(C_k, C_l) \tag{17}$$

Thus, the separation between clusters $C_k$ and $C_l$ is given by the minimum pairwise distance between them, as follows:

$$\mathrm{sep}\left(C_k, C_l\right) = \min_{x_k \in C_k, x_l \in C_l} \mathrm{DD}(x_k, x_l) \tag{18}$$

where $\mathrm{DD}(x_k, x_l)$ is the density-involved distance between points $x_k$ and $x_l$ (see [33], Definition 9). On the other hand, the compactness of the cluster $C_k$ is defined as:

$$\mathrm{com}\left(C_k\right) = \frac{1}{|C_k|} \times \mathrm{STD}(C_k) \times \mathrm{Mean}(C_k) \tag{19}$$

where $\mathrm{Mean}(C_k)$ and $\mathrm{STD}(C_k)$ are the mean and standard deviation (respectively) of a cluster with respect to the path-based distance, and $1/|C_k|$ is a penalty factor (see [33], Definition 12).

The higher the value of this index indicates a better group separability.

### C. GROUP SEPARABILITY INDICES

In this section, we describe two different measuring methodologies based on group separability (Fig. 1b). The first methodology was previously reported in the literature, and it is founded on the notion of geometrical separability, and on its basis, a new measure termed the geometrical separability index is defined [34]. The second methodology is the new approach proposed in this study, and it is based on the novel notion of projection separability and, which defines a new set of four measures termed projection separability indices.

#### 1) GEOMETRICAL SEPARABILITY INDEX

The geometrical separability index (GSI) [34] (also known as the Thornton index or TH) is a geometrically-driven separability approach defined as the proportion of a set of data points whose classification labels are the same as those of their first-nearest neighbor. This index is represented as:

$$\mathrm{GSI} = \frac{1}{T} \sum\nolimits_{k=1}^{T} f\left(x_k, x_k'\right) \tag{20}$$

where $x_k'$ is the first nearest neighbor of data point $x_k$, $T$ is the total number of data points, and $f$ is a binary function

that returns either 0 or 1, depending on which class label is associated with $x_k$ and $x'_k$, defined as:

$$f\left(x_k, x'_k\right) = \begin{cases} 1, & \text{if label } x_k = \text{label } x'_k \\ 0, & \text{if label } x_k \neq \text{label } x'_k \end{cases} \quad (21)$$

The value of the GSI is closer to 1 for a set of points in which those with opposite labels exist in well-separated groups. When groups move closer and points from opposite classes begin to overlap geometrically, the value of the index decreases. Finally, if the centroids coincide or the points are uniformly distributed in the geometrical space, this index will be close to 0.5. Thus, for this index, 1 represents the best-estimated grouping of samples.

### 2) PROJECTION SEPARABILITY INDICES

The projection separability (PS) rationale is specifically proposed here to assess the group separability of data samples in a geometrical space, such as in the case of dimension reduction analyses.

The first step is to define a projection separability line for each pair of groups in the dataset. The projection separability line is a line that, given two groups of samples in a multi-dimensional geometrical space, provides a one-dimensional geometrical representation of the data (which can be interpreted as the geometrical ordering of the samples on a line) that separates the two groups (Fig. 2). This projection separability line can be defined according to various principles, methods, or criteria. For instance, a separability line can be obtained by machine learning techniques such as Fisher's linear discriminant analysis (commonly known as LDA) [58] or linear binary soft margin Support Vector Machine (lbSVM) [59], [60]. In the case of LDA, its first discriminant can be used as a projection line, and the criterion of separability is to maximize the ratio of the variance between groups to the variance within groups. In the case of lbSVM, the line orthogonal to the maximum-margin hyperplane (the decision boundary) can be used as a projection line, and the criterion for separability is to maximize the geometrical margin between the two groups. However, both methods have a high time complexity. LDA scales cubically with the number of samples or features [61], and lbSVM scales cubically with the number of samples [60], [62]. Therefore, the application of these approaches is impractical for large datasets. To address this time complexity issue, in this study, we introduce a new methodology called the centroid projection separability line (or centroid line, for simplicity). This new approach scales linearly with the number of samples. It computes the geometrical centroids of each of the two groups, and then considers the line that connects them as a projection line. However, if the theoretical basis for defining the projection separability line by LDA and lbSVM is the optimization of the variance ratio or maximum margin, respectively, what is the theoretical basis for defining the centroid projection line as an appropriate line on which to measure group separability?

We propose the centroid projection separability line as a heuristic technique, whose conceptual roots are in the *parsimony principle* (a.k.a., Occam's razor) that is a problem-solving principle according to which *"entities should be not multiplied behind necessity"* [63], sometimes oversimplified and wrongly interpreted as *"the simplest explanation is usually the best one."* This philosophical principle advocates that when presented with competing hypotheses about the same prediction, one should also investigate the solution with the fewest assumptions.

In contrast to the assumptions of LDA based on the maximization of the variance ratio and lbSVM based on the maximization of the margin between the two groups, the assumption adopted by the centroid projection separability line is extremely simple, and it is not based on the maximization of any measure. When the goal of assessing group separability is to evaluate the extent to which a method for nonlinear dimension reduction is effective, then the centroid projection separability line criteria state that, given two groups, if the projection of the points on the line that connects the two groups' centroids is not representative of the group separability, it means that the applied nonlinear dimension reduction technique failed to provide a linear representation of the data (in the low-dimensional space) that addresses the data nonlinearity in the original feature space. In other words, this suggests that nonlinear dimension reduction did not perform well. The rationale for selecting the two centroids and the projection line that connects them is as follows. The centroid is the center of mass, which is representative of a group of particles in a geometrical space, and according to physic interpretation, it is the center of gravity on which the group of particles relies to represent the mechanics of the complex particle system in a reductionist modeling.

In Fig. 2, we provide a visual example that can help to clarify the essence of our proposed method by considering two main groups with irregular shapes: an apple (red) and its stem (green). First, the centroid of each group is identified (Fig. 2a) by calculating the median. A line segment, the centroid projection separability line, is then drawn between the centroids of the two groups (Fig. 2a). The points are then projected (Fig. 2b) and finally collapsed on the centroid line (Fig. 2c). The comparison between the centroid line and LDA line in Fig. 2d shows that they are very close to each other, and the projection of the points on the respective separability line is comparable (Fig. 2g). For time complexity reasons, in this study, we will only compare the centroid projection line with the LDA projection line, whereas the lbSVM projection line is discussed above, but we leave its investigation to future studies. The computational complexity of computing the centroid line depends on the centrality estimator selected to compute the centroid of each group. For instance, the mean, median, or mode can be considered centroid estimators. In this study, the median was applied for all the presented analyses because it is an estimator robust to noise and outliers, and its time complexity can be $O(N)$ (i.e., scales linearly) [64] in relation to the number of samples ($N$). Given that the median

should be computed for all dimensions of embedding, the final complexity of the median-based centroid projection line is $O(ND)$, where $D$ represents the dimensions of embedding. However, in this study, we considered dimension reduction for data visualization and representation; thus, $D$ is a constant that can take a value of two or three. Hence, the final complexity for computing the centroid projection line in these low-dimensional spaces is $O(N)$.

The second step is to project all the data points onto the projection separability line using the dot product as follows:

$$\text{proj}\,(P, \text{line}\,(A, B)) = A + \frac{(\boldsymbol{AP} \cdot \boldsymbol{AB})}{(\boldsymbol{AB} \cdot \boldsymbol{AB})} \times \boldsymbol{AB} \tag{22}$$

where $P$ is the point to be projected on the projection line that connects centroids $A$ and $B$, and $\boldsymbol{AP}$ and $\boldsymbol{AB}$ are vectors formed by the points in question. The projection of the points on the projection line (Fig. 2c and Fig. 2f) scales the points from a $D$-dimensional space (for instance, in Fig. 2, D=2) to a 1-dimensional (1D) space. Subsequently, the mapping of these points begins by setting a reference point $p_0$, which is one of the two extreme points on the projection line. We clarify that, given a projection line, the two extreme points are always unique, and the choice of one of them as the reference point $p_0$ is irrelevant for computing the ordering of all the other points on the line. By convention, we decided to implement the following procedure to select point $p_0$: the algorithm iteratively checks each dimension of embedding (starting from dimension one) and stops by selecting the first dimension in which at least two points assume two different projected values (i.e., dimensions where all projected points take the same value are neglected). Hence, $p_0$ is defined as the extreme point on the projection line with the lowest distance from the origin (minimum value) in this selected dimension. Finally, the distance from any other point to $p_0$ is calculated as follows:

$$\text{DP}_i = d(p_0, p_i) \tag{23}$$

where $d(p_0, p_i)$ is the Euclidean distance between the reference point $p_0$ and another point $p_i$. Hence, we define $\text{DP}_{scores} = [\text{DP}_1, \text{DP}_2, \ldots, \text{DP}_n]$ as the array that collects all of these distances.

The third and final step is to design different statistical-based validity measures, referred to as projection separability indices (PSIs). The two main formalisms for defining a new PSI are illustrated in Fig. 2h and depend on the fact that a selected statistical measure $f$ is either maximized or minimized for separability evaluation over the projection line. Indeed, the separability value is computed by means of any bounded statistical separability measure $f$, such as the Area Under the ROC-Curve [40], Area Under the Precision-Recall Curve [41], Matthews correlation coefficient [42], Mann-Whitney U-test p-value [39], and many others, such as the F-score [65], which can be investigated in future studies. This procedure is repeated for all pairs of groups for which a separability evaluation is desired, and then, an overall estimation is computed. Thus, we can define any PSI as (Fig. 2e):

$$\text{PSI} = \begin{cases} \dfrac{\mu_f}{1+\sigma_f}, & \text{if } f \text{ maximizes} \\ \dfrac{\mu_f + \sigma_f}{1+\sigma_f}, & \text{if } f \text{ minimizes} \end{cases} \tag{24}$$

where $\mu_f$ is the mean of the results provided by the implemented statistical measure $f$ over the separability line for all pairs of groups and $\sigma_f$ is the standard deviation of the mentioned results (here, used as a penalty factor). Assuming that $f$ is a bounded measure, if the upper bound of $f$ indicates the best group separability, it means that $f$ maximizes; thus, the first definition in (24) should be applied to compute the respective PSI. However, if the lower bound of $f$ indicates the best group separability, it means that $f$ minimizes; thus, the second definition in (24) should be applied to compute the respective PSI. In any of these cases, $\mu_f$ and $\sigma_f$ should be between 0 and 1, and can be calculated as follows:

$$\mu_f = \frac{1}{G}\sum_{k=1}^{G} f(\text{DP}_{scores})_k \tag{25}$$

$$\sigma_f = \sqrt{\frac{1}{G-1}\sum_{k=1}^{G} \left| f(\text{DP}_{scores})_k - \mu_f \right|^2} \tag{26}$$

where $G$ is the total number of pairwise group combinations and $f(\text{DP}_{scores})_k$ is the applied statistical measure for addressing group separability over the array of distances of the $k$th pair of groups.

Regarding the time complexity of a particular PSI. If we define $N$ as the number of points and $G$ as the number of all pairwise combinations of groups. Then, the projection of the points in (22) and the calculation of the distances in (23) are both $O(GN)$. However, in (25) and (26), the time complexity of $f(\text{DP}_{scores})_k$ depends entirely on the applied statistical measure, which is represented as $f(N)$; thus, the time complexity of $\mu_f$ and $\sigma_f$ will be $O(Gf(N))$. Therefore, the overall time complexity of computing a single PSI is $O(GN) + O(Gf(N)) \sim O(Gf(N))$ because $f(N) \geq O(N)$. In this study, we designed four different PSIs to prove the effectiveness of the proposed projection separability rationale in determining the best group separability of the DR results.

The first index, called PSI-P, is defined as:

$$\text{PSI-P} = \frac{\mu_{MWp\text{-}value} + \sigma_{MWp\text{-}value}}{1 + \sigma_{MWp\text{-}value}} \tag{27}$$

This index evaluates group separability by implementing the Mann-Whitney U-test p-value [39], (sometimes called the Mann Whitney Wilcoxon Test or the Wilcoxon Rank-Sum Test). In (27), it is represented by the subscript *MWp-value*. This measure is a nonparametric method used to test whether two independent samples of observations are drawn from the same or identical distributions. The U-test is based on the idea that the pattern exhibited when $n_k$ number of $K$ random variables and $n_l$ number of $L$ random variables are arranged together in an increasing order of magnitude provides information about the relationship between their parent populations [66] (here, a particular pairwise combination of groups). The Mann-Whitney test criterion is based on the

magnitude of the $L's$ in relation to the $K's$ (e.g., the position of $L's$ in the combined ordered sequence). A sample pattern of arrangement where most of the $L's$ are greater than most of the $K's$ or vice versa would be evidence against random mixing [39], [66] (i.e., a significant group separability). This measure is defined as follows:

$$U = \min(U_k, U_l) \tag{28}$$

where $U_k$ and $U_l$ are determined by:

$$U_k = n_k n_l + \frac{n_k(n_k+1)}{2} - R_k \tag{29}$$

$$U_l = n_k n_l + \frac{n_l(n_l+1)}{2} - R_l \tag{30}$$

where $n_k$ is the size of $K$ random variables, $n_l$ is the size of $L$ random variables, $R_k$ is the sum of the ranks for $K$, and $R_l$ is the sum of the ranks for $L$. If the observed value of $U$ is $< U_{crit}$, then the test is significant at the $\alpha$ level (the values of $U_{crit}$ for different $\alpha$ levels are given in the Mann-Whitney Tables [39], [67]); that is, this tends to reject the null hypothesis of identical distribution [66].

Theoretically, MWp-value is bounded between 0 and 1. Thus, because MWp-value is a measure that minimizes, PSI-P reports small values, where the closest value to 0 indicates the best group separability. In [68], it was reported that the time complexity of algorithms based on Mann-Whitney's original recursion formula is $O(M^2N^2)$; thus, if an implementation based on this original recursion is adopted, then the time complexity of PSI-P would be $O(GM^2N^2)$.

The second index, named PSI-ROC, is represented by:

$$\text{PSI-ROC} = \frac{\mu_{AUC\text{-}ROC}}{1+\sigma_{AUC\text{-}ROC}} \tag{31}$$

This index adopts as a separability measure the Area Under the ROC-Curve [40]. In (31), it is represented by the subscript *AUC-ROC*. It provides a measure of the trade-off between the true positive rate (TPR) and false positive rate (FPR) as follows:

$$\text{AUC-ROC} = \int_{x=0}^{1} \text{TPR}\left(\text{FPR}^{-1}(x)\right) dx \tag{32}$$

By definition, AUC-ROC is a measure that maximizes and, is bounded in a range between 0 and 1. Thus, the closest PSI-ROC value to 1 indicates the best group separability. Because AUC-ROC is based on the sum of pairwise losses between examples from different groups, the time complexity of its objective function is quadratic [69]; that is, $O(N^2)$ in the worst scenario. However, optimizing the complexity of computing AUC is a current challenge, and multiple approaches have been proposed [69]–[71]. Thus, in theory, the time complexity of the PSI-ROC based on an unoptimized implementation of AUC-ROC would be $O(GN^2)$, but it can be improved with adequate optimization.

The third index, called PSI-PR, is defined as:

$$\text{PSI-PR} = \frac{\mu_{AUC\text{-}PR}}{1+\sigma_{AUC\text{-}PR}} \tag{33}$$

This index implements as a separability measure the Area Under the Precision-Recall Curve [41]. In (33), represented by the subscript *AUC-PR*. It provides a measure of the trade-off between precision (PREC) and sensitivity (also known as recall, REC) as follows:

$$\text{AUC-PR} = \int_{x=0}^{1} \text{PREC}\left(\text{REC}^{-1}(x)\right) dx \tag{34}$$

In this case, AUC-PR is also a measure that maximizes, reporting values between 0 and 1. Hence, the closest PSI-PR value to 1 indicates the best group separability. Moreover, similar to the previous index, the time complexity of PSI-PR based on an unoptimized version of AUC-PR would be $O(GN^2)$ (however, please note that optimizations for AUC-ROC are not guaranteed to optimize AUC-PR [72]).

Finally, the fourth index, named PSI-MCC, is defined as:

$$\text{PSI-MCC} = \frac{\mu_{MCC}}{1+\sigma_{MCC}} \tag{35}$$

This index uses the Matthews Correlation Coefficient [42] as a separability measure. In (35), represented by the subscript *MCC*. It provides a coefficient between the observed and predicted binary classifications as follows:

$$\text{MCC} = \frac{\text{TP}\times\text{TN} - \text{FP}\times\text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}} \tag{36}$$

where TP states for True Positives, TN for True Negatives, FP for False Positives, and FN for False Negatives.

In general, this index reports values (coefficients) between -1 and 1, where a coefficient of 1 represents a perfect pairwise group separability, and -1 represents an inverted pairwise group separability. However, because we do not know the groups' position on the separability line, we compute MCC in both directions, first assuming that the $k$th group is on the left side and the $l$th group is on the right side, then assuming the opposite. Finally, we select the best MCC result as the actual value to express the pairwise group separability. Thus, in our case, this index returns values between 0 and 1 (not between -1 and 1, as in its formal definition), where 0 means no separation and 1 means perfect separation. Regarding its time complexity, this of MCC implementation is $O(NLogN)$ because sorting is the only expensive function. Thus, the time complexity of the PSI-MCC is $O(GNLogN)$.

As mentioned before, any other statistical-based separability measure can be employed under the proposed PS rationale for designing new projection separability indices that are different from those proposed above. However, in this study, we focus on the described measures because they are widely used in data analysis, they are bounded, and they are also sufficiently diverse between them to cover different types of separability estimations.

### D. DATA ANALYSIS

#### 1) NORMALIZATION AND ALGORITHMS' TUNING

First, to scale and adjust the raw values of the datasets before the analyses, the following normalizations were applied:

DRS, dividing each row by the row (samples) sum; DCS, dividing each column by the column (features/variables) sum; LOG, logarithmic function in base 10 of the matrix values plus 1 (because of zero values); NON, the original data without normalization, were analyzed as well.

For each normalization result, diverse DR methods were applied to obtain different embeddings of data in a low-dimensional space. Thus, the group separability of these results was assessed using different separability indices. For parameter-dependent methods such as Isomap and t-SNE, the selection of the optimal hyperparameters plays a key role in obtaining the correct embedding of the groups of samples. In the case of Isomap, we must input the number of nearest neighbors (hyperparameter $k$) that modifies the construct of the proximity graph. In the case of t-SNE, we need to provide perplexity (hyperparameter $p$), which can be interpreted as a smooth measure of the effective number of neighbors [17] or the balance between local and global aspects of the data. In addition, although not mandatory, we can input the initial number of dimensions (hyperparameter $d$) used to pre-process the data by PCA. To determine the optimal hyperparameters settings for these algorithms, an automated tuning process was implemented to test different configurations and select the one whose embedding provided the best group separability according to each index.

#### 2) TRUSTWORTHINESS: A STATISTICAL EVALUATION OF SEPARABILITY SIGNIFICANCE

This section introduces a methodological innovation proposed to account for the uncertainty of the separability estimation. Briefly, we propose a resampling methodology to build a null model that allows the computation of an empirical p-value to assess the trustworthiness of the separability measure provided by a selected index. Trustworthiness assesses the extent to which a value produced by a certain index is reliable under uncertainty. The procedure is as follows. For a given DR method result (in a certain dataset), we freeze the geometrical position of the samples and reshuffle their group labels uniformly at random. This procedure is repeated 1000 times (we set the default 1000 times because it offers an accurate estimation of the null model, but this value can also be increased in relation to the needs of the users), and the value of a certain index is measured at each round, generating a null model distribution composed of 1000 random values. Subsequently, a p-value (that expresses the extent to which the true index value can be generated at random) was computed as the number of random values that surpassed the initial (true) index value. Thus, for each index, the comparison with the null model is characterized by the following: a) the mean of the 1000 null model random values, b) the standard error of the 1000 null model random values, and c) the separability significance (p-value) to quantify whether the evaluation provided by the index is significant in comparison to the null model. As previously reported, and here expressed in a different form, the separability significance (trustworthiness) represents the fraction of random reshuffling in which the index behaves better than in the true label assignment case. In other words, it measures the likelihood that the index value will be obtained by chance. We considered as significant all the p-values lower than 0.01 with respect to the null model distribution.

In the case of t-SNE and Isomap, we corrected for multiple hypotheses testing all the p-values obtained for each hyperparameters combination using the Benjamini-Hochberg adjustment [73]. We suggest applying this adjustment any time a dimension reduction method has one or more hyperparameters in relation to which their tuning provides multiple embeddings and, therefore, multiple p-values for the same dataset.

#### 3) ASSESSMENT AND VISUALIZATION OF SIMILARITY BETWEEN PERFORMANCES OF DIFFERENT INDICES

We investigated the similarity between the different indices to discover those offering comparable results to our PSIs, regardless of the dataset or embedding method. To do so, we created an array that characterizes each index and its obtained values for all the DR methods' results across all different data normalizations, centered versions, and, in the case of MCE, multiple types of distances (Euclidean distance, Spearman rank correlation distance, and Pearson correlation distance). Thus, we created a matrix that has the indices placed as samples and their values as features/variables (according to each DR technique). Subsequently, the z-score normalization was applied to the rows to scale their values. Finally, PCA was used to visualize and compare similarities between the indices. This procedure was first applied separately for each dataset; second, by merging in a unique matrix, the indices' results for all the datasets.

## III. RESULTS

For time complexity reasons, in this study, we compared the centroid projection line versus the LDA projection line exclusively on the two initial artificial datasets.

Starting with the analysis of the Apple-Stem dataset, which offers an example of clusters with irregular shapes, Fig. 2i shows that the results of the LDA projection line and centroid projection line (according to different evaluation measures) are comparable. This indicates that both approaches for computing the projection line can provide similar results for data with irregular shapes. However, the analyses below of the Half-Moons dataset indicate their underlying differences in the case of data nonlinearity.

In Fig. 3, we report the results of the investigation of the stability of the separability measures with different increasing levels of isotropic noise over the Half-Moons dataset. This offers an emblematic example of nonlinear group separability because the two half-moon groups cannot be linearly separated by a line. Fig. 3a-f provides a visual example of the variants used with isotropic noise. We can see that the centroid line is quite stable, but the LDA line is not stable, as shown by the change in the slope coefficient value in relation to increasing noise levels (blue dashed line in Fig. 3g).

**TABLE 3.** Values of the different indices obtained for the Tripartite-Swiss-Roll dataset.

| Index/Method | Isomap | t-SNE | MCE | HD | PCA | NMDS | MDSbc |
|---|---|---|---|---|---|---|---|
| **PSI-P** | 1.01E-40 | 1.01E-40 | 1.43E-36 | 1.24E-40 | 3.28E-23 | 3.28E-23 | 5.86E-05 |
| **PSI-ROC** | 1.00 | 1.00 | 0.98 | 0.87 | 0.79 | 0.79 | 0.67 |
| **PSI-PR** | 1.00 | 1.00 | 0.98 | 0.78 | 0.68 | 0.65 | 0.41 |
| **PSI-MCC** | 1.00 | 1.00 | 0.85 | 0.63 | 0.50 | 0.46 | 0.27 |
| **DI** | 17533.66 | 0.40 | 0.03 | 0.20 | 1.28E-03 | 1.28E-03 | 2.33E-04 |
| **DB*** | 1.00 | 0.69 | 0.61 | 0.35 | 0.35 | 0.35 | 0.21 |
| **GDI** | 45464.34 | 1.51 | 0.90 | 0.79 | 0.79 | 0.79 | 0.46 |
| **CH** | 1150067129342.85 | 1044.82 | 3421.58 | 100.28 | 91.86 | 91.86 | 42.55 |
| **SIL** | 1.00 | 0.82 | 0.63 | 0.21 | 0.14 | 0.14 | -0.10 |
| **GSI** | 1.00 | 1.00 | 1.00 | 1.00 | 0.85 | 0.82 | 0.79 |
| **CVDD** | 285025577.41 | 785.81 | 490.76 | 209.99 | 1.30 | 1.30 | 1.16 |
| **AVG Rank** | 1.00 | 1.64 | 2.55 | 3.27 | 4.36 | 4.91 | 6.00 |

*This table presents the best DR validation values of the different indices for the Tripartite-Swiss-Roll dataset. This table shows the best DR validation values of the different indices for the Tripartite-Swiss-Roll dataset. In the table, the high-dimensional (HD) space was included in order to show the performance of the indices in the original high-dimensional space. Moreover, an average of rank performances (denoted as AVG Rank) of the indices values was calculated, where the lowest AVG Rank value represents the algorithm evaluated as the best DR method (as the one that offers the best group separability of the samples) by a higher number of indices.*

However, this instability to noise of the LDA line does not importantly affect the evaluation of separability, which is comparable for the centroid line and LDA line according to the different measures (Fig. 3h-k). Fig. 3h-r reports the results of this stability test for each separability measure. For instance, PSI-PR (Fig. 3h) and PSI-ROC (Fig. 3i) can also detect high group separability (their values are always $> 0.90$, and therefore close to 1) in the presence of the nonlinear separability (curvilinear) of this dataset, and they are robust to noise; indeed, their trend is stable and minimally affected by increasing levels of noise. The same can be concluded for PSI-P (Fig. 3k), which provides a p-value that is close to zero, regardless of the level of noise. PSI-MCC (Fig. 3j), SIL (Fig. 3m), and DB* (Fig. 3n) are also stable to noise, but they provide a separability estimation of approximately 0.5, indicating that they are partially affected by the presence of nonlinearity in the geometrical pattern of the clusters. GSI (Fig. 3l) can detect group separability in the presence of nonlinearity (its values are always $> 0.80$), but it seems to suffer more than the other indices in the presence of noise. DI, GDI, CH, and CVDD (Fig. 3o-r) cannot be assessed for their ability to detect separability in the presence of nonlinearity because they are not bounded measures; however, with the exception of GDI, they all seem to suffer to a certain extent from stability issues in the presence of noise. The results of this investigation are summarized in the column *Robustness to nonlinear (curvilinear) pattern* and *Robustness to isotropic noise* in Table 1.

In addition, Fig. 4 reports the results of the investigation of the stability of the separability measures applied to the Half-Moons dataset with different increasing levels of outliers (up to 10% of the number of samples in the dataset). Fig. 4a-f provides visual examples of the variants used and increasing levels of anisotropic noise (outliers). Once more, the centroid line is quite stable, but the LDA line is not stable, as is evident from the change in the slope coefficient in relation to the increasing outliers' levels (blue dashed line in Fig. 4g). However, this instability to anisotropic noise of the LDA line does not importantly affect the evaluation of separability, which is comparable for both the centroid line and LDA line according to the different measures (Fig. 4h-k). Fig. 4h-r reports the results of this stability test for each separability measure. PSI-PR (Fig. 4h) and PSI-ROC (Fig. 4i) can also detect relatively high group separability (their values are always $> 0.75$) in the presence of outliers; indeed, their trend is stable, but is minimally affected by increasing levels of outliers. This makes sense because the presence of outliers represents a form of anisotropic noise; therefore, it should affect the separability of the clusters more than the isotropic noise described previously. PSI-P (Fig. 4k) provides a p-value close to zero, regardless of the level of outliers. PSI-MCC (Fig. 4j), SIL (Fig. 4m), and DB* (Fig. 4n) were also partially affected by outliers. GSI (Fig. 4l) was not affected by the outliers. Regarding the other CVIs, DI and GDI (Fig. 4o and Fig. 4p) were partially affected by outliers. However, CH and CVDD (Fig. 4q and Fig. 4r) seem to suffer to a certain extent from stability issues in the presence of outliers. The results of this investigation are summarized in the column *Robustness to anisotropic noise (outliers)* in Table 1.

In the following results, we will consider how the sample group separability in the low-dimensional embeddings returned (for each dataset) by the DR methods was evaluated using the previously described indices. Thus, the effectiveness of these indices is determined by the correct

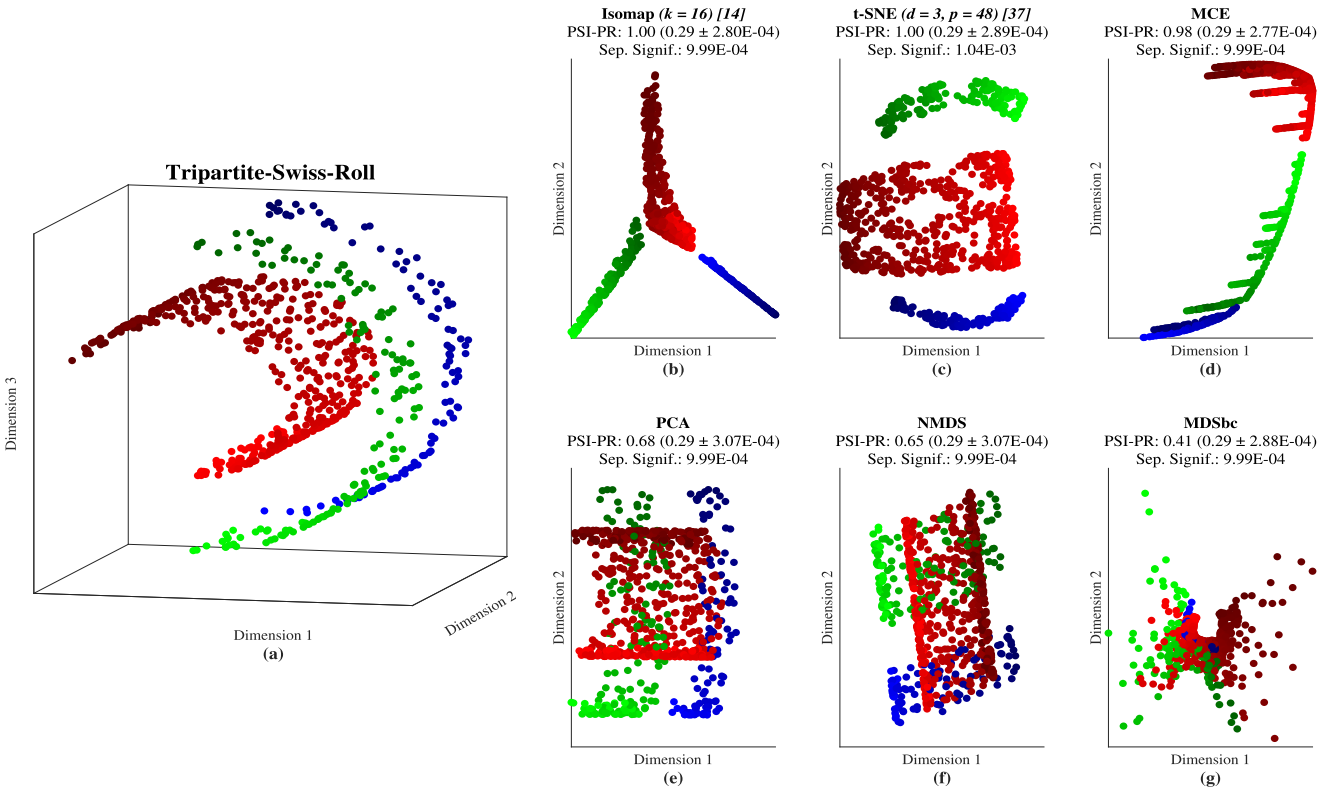



**FIGURE 5.** PSI-PR of the DR methods applied to the Tripartite-Swiss-Roll dataset. In (a) it is displayed the Tripartite-Swiss-Roll dataset in the original 3D-space. The three different colors (red, blue, and green) represent the three partitions of the Swiss-roll manifold, while the color gradient serves as a reference to see how good the DR methods are in preserving the original nonlinear structure. The DR methods were sorted from the best (top left) to the worst (bottom right) in (b)-(g) according to the PSI-PR validation value. In (b) and (c), the optimal hyperparameters are specified in parentheses next to the title of each DR method; moreover, the digit in brackets represents the number of possible solutions with the same validation value. Furthermore, next to the value returned by this index, it is reported in parentheses the random baseline, which is the mean and the standard error of the PSI-PRs computed by randomly re-shuffling the class labels 1000 times. Finally, we added a p-value which indicates the separability significance of each index in comparison to a respective null model computed by random re-shuffling.

identification of the DR method that produces the best low-dimensional representation of the data, i.e., the method that provides the best group separability of the samples. Because of time complexity and considering the large number of tests we performed, only the centroid line PSIs are considered. The results validated by PSI-PR are presented for each dataset in Fig. 5 (Tripartite-Swiss-Roll), Fig. 6 (Gastric mucosa microbiome), Fig. 7 (Radar signal), Fig. 8 (Image proteomics), Fig. 9 (MINST 2D), and Fig. 10 (MNIST 3D). In these figures, we report the PSI-PR value to allow the comparison between the visual perception of separability and its quantification provided by one of the PSIs. For this measure, the value closest to 1 indicates the best group separability. Similar figures obtained by all other indices are shown in the supplementary information (Suppl. Fig. 1-94). Moreover, in all figures, for each index, we report the trustworthiness calculated according to the procedure described in Section II.D.2. In particular, the mean and standard error of the null model and the separability significance (expressed as a p-value) are reported. In the figures, both the mean and standard error are placed in brackets next to the actual value of the indices, and they represent a reference to determine how far the real estimated separation is from a null model average. On the other hand, the p-value used for expressing separability significance is placed under the value of each index, which indicates the significance of the reported index value in comparison to a null model. In other words, the p-value represents the likelihood that the reported value of a certain index can be generated by chance.

The indices values obtained for each DR method on each dataset are shown in Table 3 (Tripartite-Swiss-Roll), Table 4 (Gastric mucosa microbiome), Table 5 (Radar signal), Table 6 (Image proteomics), Table 7 (MNIST 2D), and Table 8 (MNIST 3D). In the tables, the order of the DR methods is obtained by averaging the ranks (named AVG rank) of the values provided by the indices. The indices were also applied to the data in the original high-dimensional space (HD) to determine the effectiveness of the DR methods in preserving the group separability present in the original multidimensional space.

### A. TRIPARTITE-SWISS-ROLL

Tripartite-Swiss-Roll is a synthetic dataset that contains three groups, each characterized by a nonlinear structure (Fig. 5a). For this dataset, all indices evaluated Isomap, MCE, and t-SNE as those that provided the best group separability (Table 3). In addition, in Table 3, we can see how these DR methods outdid the HD results, providing a clear hint that

**TABLE 4.** Values of the different indices obtained for the Gastric mucosa microbiome dataset.

| Index/Method | MCE | t-SNE | Isomap | HD | PCA | MDSbc | NMDS |
|---|---|---|---|---|---|---|---|
| **PSI-P** | 0.01 | 3.73E-03 | 0.01 | 0.01 | 0.02 | 0.05 | 0.02 |
| **PSI-ROC** | 0.91 | 0.90 | 0.87 | 0.88 | 0.85 | 0.81 | 0.85 |
| **PSI-PR** | 0.96 | 0.95 | 0.94 | 0.94 | 0.91 | 0.86 | 0.90 |
| **PSI-MCC** | 0.61 | 0.72 | 0.61 | 0.61 | 0.61 | 0.59 | 0.61 |
| **DI** | 0.04 | 0.28 | 0.09 | 0.32 | 0.03 | 0.10 | 0.01 |
| **DB*** | 0.47 | 0.47 | 0.45 | 0.38 | 0.43 | 0.41 | 0.43 |
| **GDI** | 0.77 | 0.94 | 0.85 | 0.72 | 0.72 | 0.77 | 0.72 |
| **CH** | 38.81 | 14.99 | 15.86 | 9.96 | 15.97 | 10.30 | 14.19 |
| **SIL** | 0.42 | 0.40 | 0.32 | 0.20 | 0.22 | 0.21 | 0.22 |
| **GSI** | 0.79 | 0.79 | 0.79 | 0.63 | 0.63 | 0.54 | 0.63 |
| **CVDD** | 12.56 | 0.32 | 5.08 | 15.69 | 1.08 | 6.79 | 0.56 |
| **AVG Rank** | 1.82 | 2.27 | 3.09 | 3.73 | 4.09 | 4.82 | 4.91 |

*This table shows the best DR validation values of the different indices for the Gastric mucosa dataset. In the table, the high-dimensional (HD) space was included in order to show the performance of the indices in the original high-dimensional space. Moreover, an average of rank performances (denoted as AVG Rank) of the indices values was calculated, where the lowest AVG Rank value represents the algorithm evaluated as the best DR method (as the one that offers the best group separability of the samples) by a higher number of indices.*

these algorithms preserved the original group separability present in HD well. All indices yielded Isomap as the DR method that provided the best group separability of the data over the other techniques. In Table 3, we can see that in the case of Isomap, the values obtained by DI, GDI, CH, and CVDD are incredibly high in comparison with the values of the other DR methods for the same indices. Thus, we might believe that these indices are perfectly accurate for assessing the best group separability of this DR method. However, careful analysis of the visual representation of the Isomap results evaluated using these indices (Suppl. Fig. 4, 6, 7, and 10, respectively) reveals that they evaluated, with a good group separability, a 2D-embedding which does not correctly preserve the original intrinsic structure of the Tripartite-Swiss-Roll (Fig. 5a), a situation that was replicated by DB* and SIL (Suppl. Fig. 5 and 8). Indeed, in the figures, we can see that all sample points of each respective synthetic group collapse in the same position. This explains why these indices obtained such high values because they maximize inter-group separation and neglect intra-group variability, which is an important property to retain. Moreover, the aforementioned indices found only one optimal value of the hyperparameter for Isomap ($k = 4$). In contrast, all PSIs (PSI-P, PSI-ROC, PSI-PR, and PSI-MCC) and GSI detected multiple optimal values for this hyperparameter (14 different configurations of $k$, Suppl. Fig. 11-15), where six of these options presented perfect group separability and also preserved the intrinsic original data structure in the first two dimensions of embedding. Hence, these solutions conserved intra-group variability, which is visible because of the preservation of the color gradient present in the original high-dimensional shape of Tripartite-Swiss-Roll (Fig. 5a). Thus, this is the first evidence that the PSIs perform better than most of the other indices, not only while evaluating group separability but also while automatically identifying the optimal hyperparameters of parameter-dependent algorithms, such as Isomap.

Another result related to the analysis of this dataset is that none of the indices evaluated PCA as the dimension reduction method with the best group separability. As already proven in [18], PCA is a linear transformation that cannot address the type of data nonlinearity present in this dataset. Therefore, this result was expected, and it indicates that all indices were able to detect PCA's inability to deal with nonlinear high-dimensional data.

### B. GASTRIC MUCOSA MICROBIOME

This dataset provided a scenario with a small number of samples. This might represent a problem for evaluating group separability because, in the case of a low number of data points, they become highly sparse in the high-dimensional space. Thus, after dimension reduction, the geometrical localization of these points can be close to their random distribution. Hence, we evaluated whether the DR methods could correctly separate the three main groups: untreated dyspeptic patients without *H. pylori* infection (HPneg), untreated dyspeptic patients with *H. pylori* infection (HPpos), and patients treated with Proton Pump Inhibitors (PPI). In Table 4, we can see that the indices evaluated different DR methods with the best group separability, namely t-SNE by PSI-P, PSI-MCC, DI, and GDI; MCE by PSI-PR, PSI-ROC, CH, DB*, CVDD, and SIL; and Isomap by GSI. In particular, the DI's evaluations (Suppl. Fig 24) erroneously indicated that all embeddings provide a separability that is not statistically significant according to trustworthiness (i.e., this index reported a separability significance with a p-value larger than 0.01 for these embeddings), meaning that the evaluations

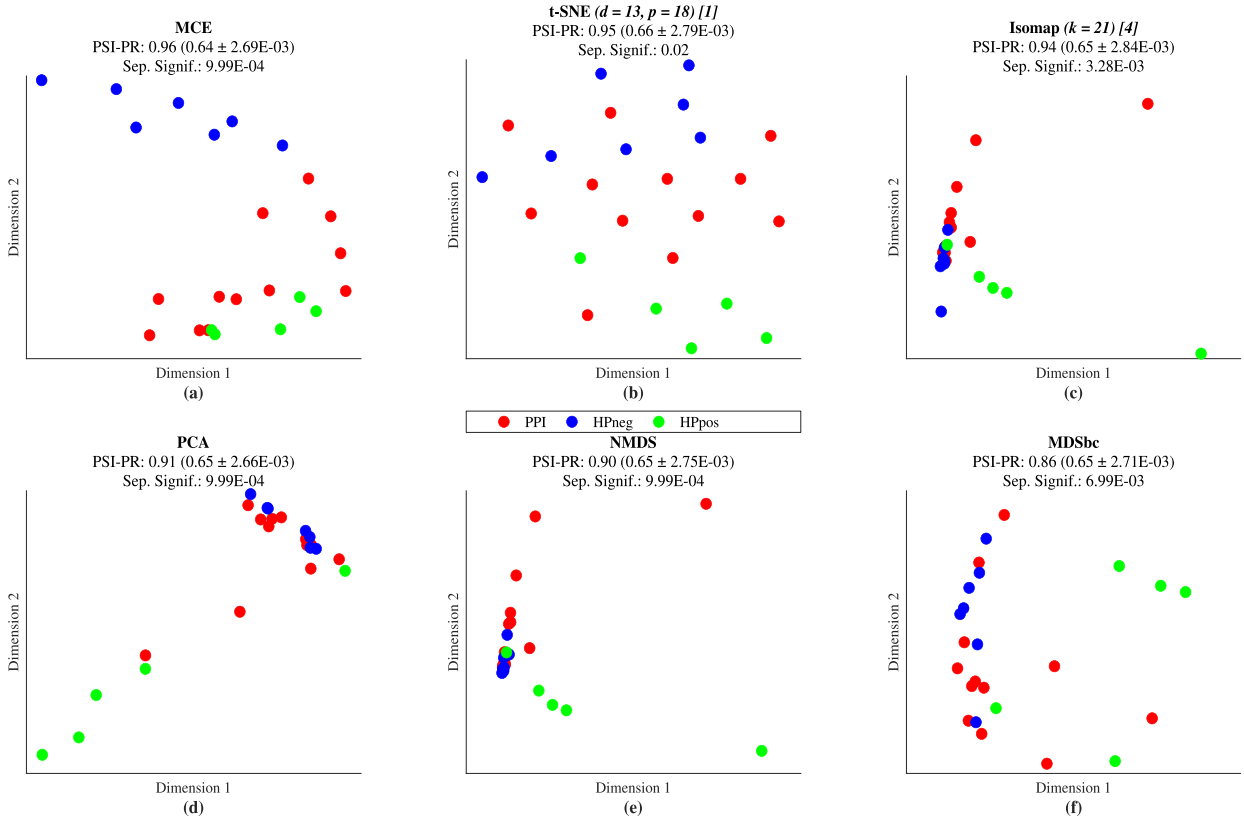



**FIGURE 6.** PSI-PR of the DR methods applied to the Gastric mucosa microbiome dataset. The DR methods were sorted from the best (top left) to the worst (bottom right) in (a)-(f) according to the PSI-PR value. In (b) and (c), the optimal hyperparameters are specified in parentheses next to the title of each DR method; moreover, the digit in brackets represents the number of possible solutions with the same validation value. Furthermore, next to the value returned by this index, it is reported in parentheses the random baseline, which is the mean and the standard error of the PSI-PRs computed by randomly re-shuffling the class labels 1000 times. Finally, we added a p-value which indicates the separability significance of each index in comparison to a respective null model computed by random re-shuffling.

given by DI on the applied DR methods are not reliable. This result is erroneous because, while most of the other indices can detect separability that is statistically significant, DI cannot select any embedding that is meaningful from a statistical perspective. CVDD also evaluates all embeddings as not significant, except for MCE, which was evaluated by CVDD as the only method to provide a statistically significant embedding (Suppl. Fig 30). This result indicates that CVDD is more robust than DI on this dataset because it agrees with PSI-PR, PSI-ROC, CH, DB*, and SIL in that MCE provides the first and meaningful embedding. GDI detected significant separability for most of the DR methods, except for Isomap, PCA, and NMDS (Suppl. Fig 26). In the case of GSI, all the methods provided significant embeddings, except for t-SNE and MDSbc (Suppl. Fig. 29).

Arguably, in Fig. 6, some of the embeddings evaluated as significant by PSI-PR contain outliers (for instance, Fig. 6c for Isomap and Fig. 6e for NMDS). These results are significant because, in general, PSI-PR is not perturbed by outliers if they are ordered on the projection line in a position that agrees with the separability of the group to which they belong. This is valid for all PSIs based on statistics that consider only the ordering of the points and not their relative distance on the projection line, such as the Area Under the Precision-Recall Curve (AUC-PR) or the Area Under the ROC-Curve (AUC-ROC). On the one hand, this avoids overestimating the positive contribution to separability of outliers that are geometrically far and located in the same separability region of their group. On the other, this reduces the negative impact of outliers that are geometrically far and opposite to the separability region of their group. Most CVIs (for instance, DI) are instead mathematically designed to aim compactness by strongly penalizing for outliers "regardless" (or minorly accounting) on the fact that they contribute positively or negatively to the separability estimation of the groups to which they belong.

In general, t-SNE, Isomap, and MCE outdid HD on this dataset (Table 4), confirming their capability to detect the original group separability. However, according to PSI-PR (Fig. 6), MCE was the best method for identifying clear visual separation of the groups. Moreover, the aforementioned DR methods did not show any internal separation of PPI-treated patients according to their *H. pylori* infection. It is known in the literature that PPIs cause a significant perturbation in the gastric tissue microbiota of dyspeptic patients regardless of the initial pathological infection due to *H. pylori* [48]. To confirm our results, we repeated this analysis by restricting the dataset samples to only two PPI subgroups

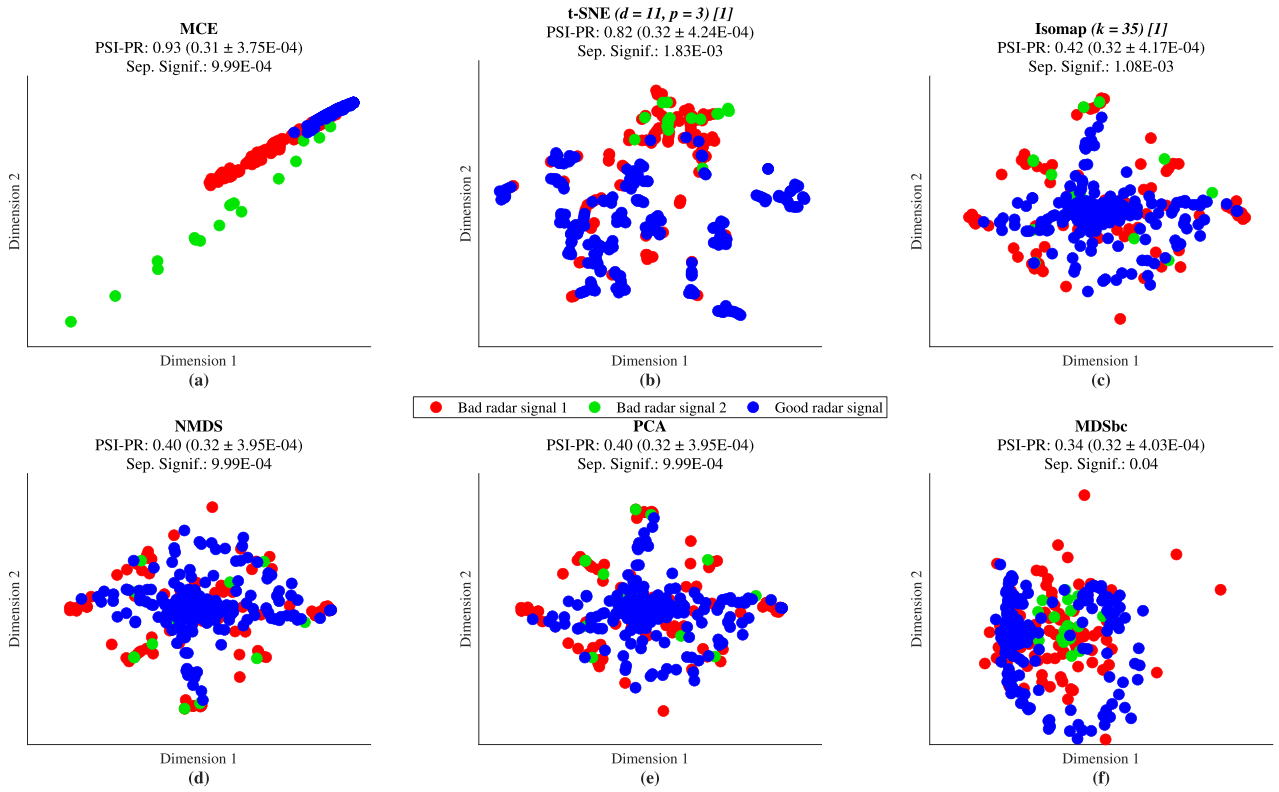



**FIGURE 7.** PSI-PR of the DR methods applied to the Radar signal dataset. The DR methods were sorted from the best (top left) to the worst (bottom right) in (a)-(f) according to the PSI-PR validation value. In (b) and (c), the optimal hyperparameters are specified in parentheses next to the title of each DR method; moreover, the digit in brackets represents the number of possible solutions with the same validation value. Furthermore, next to the value returned by this index, it is reported in parentheses the random baseline, which is the mean and the standard error of the PSI-PRs computed by randomly re-shuffling the class labels 1000 times. Finally, we added a p-value which indicates the separability significance of each index in comparison to a respective null model computed by random re-shuffling.

(PPIneg and PPIpos). However, none of the indices provided a significant group separability, meaning that the DR methods were not able to significantly separate these two subgroups (Suppl. Fig. 36-47).

### C. RADAR SIGNAL

As proposed in [51], we analyzed the Radar signal dataset using three main groups: good radar signal, bad radar signal 1, and bad radar signal 2. The aim of analyzing this dataset is to verify whether the indices can evaluate the results of the DR methods in the presence of the crowding problem (i.e., highly overlapped groups of samples in the reduced dimensional space). Table 5 shows that GSI, SIL, CH, DB*, PSI-PR, PSI-ROC, PSI-MCC, and PSI-P evaluated MCE as the DR method with the best group separability. In terms of visualization, the evaluations provided by GSI (Suppl. Fig. 56), SIL (Suppl. Fig. 55), CH (Suppl. Fig. 54), DB* (Suppl. Fig. 52), PSI-PR (Fig. 7), PSI-MCC (Suppl. Fig. 50), PSI-ROC (Suppl. Fig. 49), and PSI-P (Suppl. Fig. 48) in relation to MCE, sorted out the crowding problem. The fact that MCE can solve the crowding problem was proposed in [51], and we confirmed this hypothesis based on these indices. Instead, other DR methods such as PCA, Isomap, NMDS, MDSbc, and t-SNE tend to mix the different groups (crowded embedding results); thus, the mentioned indices successfully detected this situation and evaluated MCE as the DR method with the best group separability.

In particular, DI and GDI did not provide reliable results for this dataset. In the case of GDI, this index reported values under the random baseline for Isomap, PCA, NMDS, and MDSbc. In the case of MCE, the value obtained by this index is equal to the random baseline, and its separability is not significant (p-value = 0.5), indicating that the evaluation performed by this index did not assess a trustworthy group separability. Only t-SNE had a GDI value above the random baseline (Suppl. Fig. 53a); however, in the figure, we can clearly visualize an erroneous group separability that is not significant (p-value = 0.80) and presents a predominance of the good radar signal group, which is mixed across all other groups. A similar situation was faced by DI, which reported values under (or equal to) the random baseline for the evaluation of NMDS, PCA, MDSbc, and MCE (Suppl. Fig. 51). Moreover, in Suppl. Fig. 51a shows that DI evaluated t-SNE as the DR method with the best group separability; however, its visual representation shows a clear overlap of the groups. In the case of CVDD, this index was able to detect MCE with a significant group separability (Suppl. Fig. 57b), but it evaluated t-SNE as the DR method with the best group separability (Suppl. Fig. 57a). However, in the figure, we can see that t-SNE presents a clear overlap between the different groups.

**TABLE 5.** Values of the different indices obtained for the radar signal dataset.

| Index/Method | MCE | t-SNE | HD | Isomap | PCA | NMDS | MDSbc |
|---|---|---|---|---|---|---|---|
| **PSI-P** | 3.60E-11 | 1.41E-07 | 1.98E-03 | 6.53E-04 | 0.03 | 0.03 | 0.03 |
| **PSI-ROC** | 0.92 | 0.84 | 0.71 | 0.67 | 0.64 | 0.64 | 0.65 |
| **PSI-PR** | 0.93 | 0.82 | 0.50 | 0.42 | 0.40 | 0.40 | 0.34 |
| **PSI-MCC** | 0.70 | 0.54 | 0.26 | 0.19 | 0.12 | 0.12 | 0.02 |
| **DI** | 1.11E-03 | 0.01 | 0.05 | 4.79E-03 | 3.07E-03 | 3.07E-03 | 1.15E-03 |
| **DB*** | 0.54 | 0.45 | 0.21 | 0.29 | 0.30 | 0.22 | 0.19 |
| **GDI** | 0.67 | 0.70 | 0.60 | 0.58 | 0.60 | 0.60 | 0.61 |
| **CH** | 317.11 | 89.37 | 10.26 | 21.40 | 21.52 | 8.67 | 6.13 |
| **SIL** | 0.39 | 0.17 | 0.05 | 0.09 | 0.03 | 0.02 | 0.02 |
| **GSI** | 0.92 | 0.91 | 0.82 | 0.83 | 0.79 | 0.68 | 0.70 |
| **CVDD** | 2.05 | 4.05 | 0.92 | 0.87 | 0.77 | 0.35 | 0.84 |
| **AVG Rank** | 1.73 | 1.82 | 3.73 | 3.82 | 4.82 | 5.55 | 5.73 |

*This table shows the best DR validation values of the different indices for the Radar signal dataset. In the table, the high-dimensional (HD) space was included in order to show the performance of the indices in the original high-dimensional space. Moreover, an average of rank performances (denoted as AVG Rank) of the indices values was calculated, where the lowest AVG Rank value represents the algorithm evaluated as the best DR method (as the one that offers the best group separability of the samples) by a higher number of indices.*

### D. IMAGE PROTEOMICS

In the case of the Image proteomic dataset, we tested whether the DR methods can sort out the ''curse of dimensionality'' by separating the three main groups: control patients (C), patients without neuropathic pain (M), and patients with pain (P) in the low-dimensional space.

In general, the indices evaluated the group separability of the embeddings provided by t-SNE, MDSbc, Isomap, and MCE close to HD (Table 6), corroborating their ability to conserve the original high-dimensional group separability in the reduced dimensional space. Most indices placed t-SNE as the DR method with the best group separability. However, GDI and DI provided unreliable results. GDI (Suppl. Fig. 64) reported values under the random baseline for MCE, MDSbc, NMDS, and PCA. DI (Suppl. Fig. 62) presented the same problem regarding the evaluation of PCA, NMDS, and MDSbc. Moreover, we noticed that in the evaluations provided by CH (Suppl. Fig. 65), all DR methods display a fuzzy structure of the groups involving the classes C and P (both groups are mixed). This situation was shared by PCA and NMDS in all different indices' evaluations (Fig. 8, Suppl. Fig. 59-68). Interestingly, t-SNE showed a perfect group separability according to all PSIs (Fig. 8, Suppl. Fig. 59-61), DB* (Suppl. Fig. 63), and GSI (Suppl. Fig. 67). In the case of the t-SNE embedding evaluated by SIL (Suppl. Fig. 66) as the one with the best group separability, it presents outliers for the classes M and P. On the other hand, and in contrast to all other indices, CVDD evaluated t-SNE as the DR method with the worst group separability; however, its visual representation seems to be correct (Suppl. Fig. 68f). Instead, this index selected PCA as the DR method with the best group separability; however, its embedding provides a poor visual group separability (Suppl. Fig. 68a). Thus, the reliability of this index within this dataset is not plausible.

Regarding the automatic selection of hyperparameters, GSI and the PSIs identified multiple optimal settings for t-SNE, whereas the other indices could not identify more than one. GSI found eight different possible configurations of hyperparameters for t-SNE (Suppl. Fig. 73) with the same best group separability value for this index, whereas the PSIs only detected three possible solutions (Suppl. Fig. 69-72). If we compare these results, we can notice that in some of the configurations evaluated by GSI, either group P or M is highly split into multiple subgroups. In contrast, in all solutions identified by the PSIs, the three groups (C, M, and P) are mostly perfectly segregated; thus, the hyperparameters settings evaluated by the PSIs are more accurate than those evaluated by GSI. This, one more time, demonstrates the utility of our PSIs to automatically determine the optimal hyperparameters for a parameter-dependent algorithm such as t-SNE.

### E. MNIST

One of the limitations of the previous analyses is that we considered datasets composed of no more than four groups; therefore, the embedding in 2D space was sufficient to represent the variability between this limited number of groups. To overcome this limitation, we investigated the performance of the separability indices in a more diversified scenario that involved a dataset composed of 10 groups. Moreover, this time, we did not evaluate the embeddings of the DR methods only in 2D but also in a 3D space. To this end, we analyzed a subset of the MNIST dataset with 300 samples randomly selected from the original 60000 training images. This means

**TABLE 6.** Values of the different indices obtained for the image proteomics dataset.

| Index/Method | t-SNE | MDSbc | Isomap | HD | MCE | PCA | NMDS |
|---|---|---|---|---|---|---|---|
| **PSI-P** | 5.11E-05 | 1.79E-04 | 1.50E-04 | 5.11E-05 | 2.32E-04 | 0.03 | 0.04 |
| **PSI-ROC** | 1.00 | 0.97 | 0.93 | 1.00 | 0.92 | 0.81 | 0.80 |
| **PSI-PR** | 1.00 | 0.97 | 0.95 | 1.00 | 0.93 | 0.87 | 0.84 |
| **PSI-MCC** | 1.00 | 0.86 | 0.75 | 1.00 | 0.66 | 0.63 | 0.61 |
| **DI** | 0.77 | 0.12 | 0.12 | 0.69 | 0.19 | 0.10 | 0.03 |
| **DB*** | 0.62 | 0.55 | 0.52 | 0.30 | 0.52 | 0.39 | 0.36 |
| **GDI** | 1.40 | 0.57 | 0.70 | 0.50 | 0.61 | 0.50 | 0.50 |
| **CH** | 423.82 | 64.19 | 28.37 | 5.31 | 60.45 | 57.51 | 28.37 |
| **SIL** | 0.68 | 0.57 | 0.48 | 0.08 | 0.42 | 0.34 | 0.28 |
| **GSI** | 1.00 | 0.90 | 0.93 | 0.88 | 0.95 | 0.88 | 0.86 |
| **CVDD** | 2.34 | 79.08 | 391.14 | 984.93 | 11.61 | 654.49 | 221.79 |
| **AVG Rank** | 1.55 | 3.00 | 3.27 | 3.55 | 3.64 | 4.73 | 5.91 |

*This table shows the best DR validation values of the different indices for the Image proteomics dataset. In the table, the high-dimensional (HD) space was included in order to show the performance of the indices in the original high-dimensional space. Moreover, an average of rank performances (denoted as AVG Rank) of the indices values was calculated, where the lowest AVG Rank value represents the algorithm evaluated as the best DR method (as the one that offers the best group separability of the samples) by a higher number of indices.*

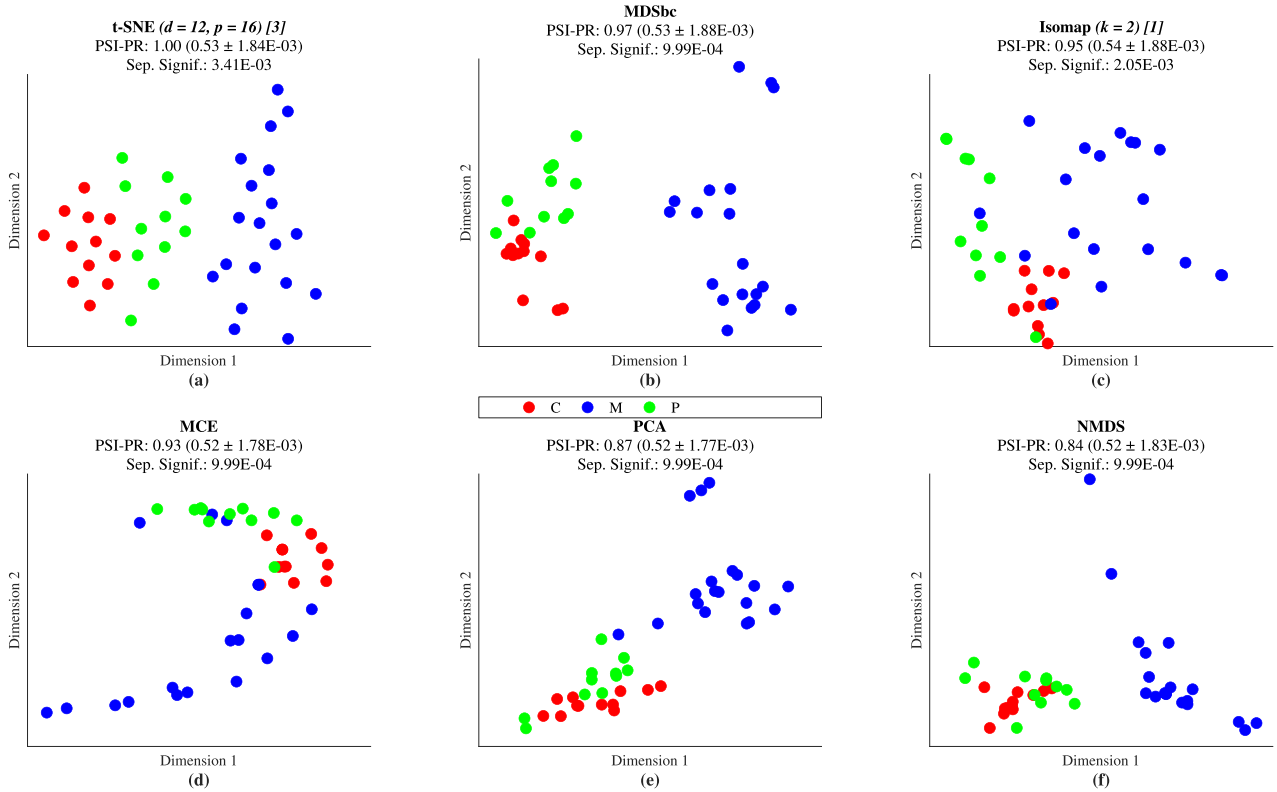



**FIGURE 8.** PSI-PR of the DR methods applied to the Image proteomics dataset. The DR methods were sorted from the best (top left) to the worst (bottom right) in (a)-(f) according to the PSI-PR validation value. In (a) and (c), the optimal hyperparameters are specified in parentheses next to the title of each DR method; moreover, the digit in brackets represents the number of possible solutions with the same validation value. Furthermore, next to the value returned by this index, it is reported in parentheses the random baseline, which is the mean and the standard error of the PSI-PRs computed by randomly re-shuffling the class labels 1000 times. Finally, we added a p-value which indicates the separability significance of each index in comparison to a respective null model computed by random re-shuffling.

30 samples for each number from zero to nine to create 10 different groups. In other words, each group contained 30 samples of images representing the same number from a random handwrite variation. Then, two different embeddings,

one in a two-dimensional space (2D) and another in a three-dimensional space (3D), were performed to compare the group separability given by the DR methods in a different number of dimensions.

Interestingly, all indices evaluated t-SNE as the DR method with the best group separability in both 2D and 3D spaces (Table 7 and Table 8). In fact, in the tables, we can notice that this technique is the only one that outdid HD, confirming that it retains the separability of the original high-dimensional space. These results are in concordance with those of previous studies [74], [75], where it has been confirmed that t-SNE performs exceptionally well in analyzing this dataset. Indeed, t-SNE is able to separate the digit groups better; however, as for the other DR methods, in some cases, certain numbers that are difficult to identify (distorted digits) were embedded in the wrong groups (Fig. 9, Fig. 10, and Suppl. Fig. 74-93).

Again, GDI and DI presented unreliable evaluations under the random baseline, but this time together with CVDD. GDI presented this problem in the evaluation of MDSbc, NMDS, PCA, MCE, and Isomap in both 2D and 3D low-dimensional spaces (Suppl. Fig. 79 and Suppl. Fig. 89, respectively). DI also presented the same problem in the evaluation of PCA, MCE, MDSbc, and NMDS (in both dimensional spaces, Suppl. Fig. 77 and Suppl. Fig. 87). CVDD also presented this issue in the evaluation of MDSbc, MCE, PCA, Isomap, and NMDS (Suppl. Fig. 83 and Suppl. Fig. 93).

As mentioned earlier in the text, the MNIST dataset has the peculiarity that it does not present a hierarchical organization of the samples. Indeed, the graphical shape of the numbers is not organized in a hierarchy. Therefore, MCE should be evaluated with low performance (i.e., a nonsignificant group separability) for this dataset because it is not a manifold embedding but a hierarchical embedding method (for more details, see [51]). Interestingly, PSI-PR (Fig. 9 and Fig. 10) detected this situation and evaluated MCE as the DR technique with the worst group separability in both dimensional spaces (2D and 3D).

In addition, by inspecting the values returned by the PSIs (Table 7 and Table 8), we noticed that in the presence of hundreds of samples and a wider number of groups, the embeddings in 3D were evaluated, as expected, with a higher group separability than the embeddings in 2D. Indeed, the representation of group separability is slightly more evident in a three-dimensional space, in which the different groups of digits appear more separated and less overlapped than in a two-dimensional space. As anticipated in the Introduction, this comparison between separability performances in different dimensions is less straightforward if we employ methods such as the CVIs that do not have a bounded range of evaluation. Therefore, the same value in two different dimensional spaces of different sizes does not necessarily indicate a similar level of separation.

### F. SIMILARITIES ACROSS DIFFERENT INDICES

A final comparison was performed to analyze the similarities between the different indices in each dataset (Suppl. Fig. 101), using PCA. It might seem outlandish; we indeed adopted dimension reduction to evaluate the similarity between techniques designed to quantify the group separability obtained by dimension reduction methods. In Suppl. Fig. 101 and Fig. 11, we have created a quadrilateral whose vertices are the PSIs (PSI-P, PSI-ROC, PSI-PR, and PSI-MCC), named PSI-quadrilateral, with the aim of visually identifying the indices that are comparable to our PSIs (e.g., the ones inside the PSI-quadrilateral).

In the case of the Tripartite-Swiss-Roll dataset, GSI is close to the PSI-quadrilateral, meaning that this index gave similar evaluations to the PSIs. These indices are separability-driven; thus, differently from CVIs, they do not maximize the geometrical compactness of the data samples within a particular group, as confirmed for this dataset. Moreover, in the case of the MNIST (3D) dataset, GDI is also close to the PSI-quadrilateral. In fact, if we compare Fig. 11 and Suppl. Fig. 101, we notice that there are similar low-dimensional embeddings evaluated by these two indices as those with the best group separability. However, our PSIs provided, in contrast to GDI, a significant separability in comparison to a random permutation of the labels; thus, their results are more reliable than those provided by GDI.

Finally, to observe the overall trend across all datasets, the analysis was repeated by merging all indices results obtained for all DR methods in all datasets (Fig. 11). In the figure, PSIs are (once more) grouped. This again confirms the coherence of the different statistical measures implemented as PSIs, which, by following the PS rationale, were specifically designed for the validation of group separability returned by the DR techniques.

In Fig. 11, if we project all indices onto the first dimension (the one with the highest data variability), we can observe that GSI, SIL, and DB$^*$ are the closest indices in relation to the PSIs. SIL and DB have been reported to be the best cluster validity indices in several studies [22], [76]–[78]. Given the above, their proximity to our PSIs confirms that our indices are contenders, even for the best CVIs, and might solve some of their limitations. For instance, DB is not designed to deal with overlapping clusters [79], and SIL is easily skewed by outliers [37]. On the other hand, the proximity between GSI and our PSIs is not surprising because both approaches focus on determining group separability instead of cluster validity (Fig. 1). This explains why they do not neglect the intra-group diversity of the samples, in contrast to the CVIs. A situation that was evidenced in the analysis of the Tripartite-Swiss-Roll dataset, where the PSIs and GSI were the only indices that did not force the samples of each group to collapse at a unique point.

## IV. DISCUSSION

There is no universal way to map a given dataset from a high-dimensional space into a reduced number of dimensions by perfectly preserving all properties of the original structure. Thus, the embedding performance of different DR methods

**TABLE 7.** Values of the different indices obtained for the MNIST dataset in a two-dimensional (2D) space.

| Index/Method | t-SNE | HD | MCE | Isomap | MDSbc | PCA | NMDS |
|---|---|---|---|---|---|---|---|
| **PSI-P** | 1.17E-05 | 7.50E-09 | 0.05 | 0.04 | 0.10 | 0.17 | 0.17 |
| **PSI-ROC** | 0.91 | 0.98 | 0.80 | 0.81 | 0.80 | 0.78 | 0.78 |
| **PSI-PR** | 0.87 | 0.98 | 0.76 | 0.79 | 0.81 | 0.79 | 0.79 |
| **PSI-MCC** | 0.78 | 0.85 | 0.62 | 0.58 | 0.54 | 0.52 | 0.52 |
| **DI** | 0.30 | 0.50 | 0.01 | 0.04 | 0.01 | 0.02 | 4.52E-03 |
| **DB*** | 0.37 | 0.24 | 0.19 | 0.23 | 0.18 | 0.15 | 0.15 |
| **GDI** | 0.74 | 0.66 | 0.56 | 0.50 | 0.66 | 0.66 | 0.66 |
| **CH** | 133.16 | 10.22 | 85.41 | 56.57 | 53.16 | 44.34 | 44.34 |
| **SIL** | 0.33 | 0.07 | -0.06 | -0.06 | -0.04 | -0.07 | -0.07 |
| **GSI** | 0.88 | 0.82 | 0.66 | 0.46 | 0.39 | 0.33 | 0.33 |
| **CVDD** | 2.07 | 0.03 | 0.18 | 1.84E-03 | 8.59 | 1.84E-03 | 1.84E-03 |
| **AVG Rank** | 1.55 | 2.27 | 3.91 | 4.00 | 4.18 | 5.27 | 5.91 |

*This table shows the best DR validation values of the different indices for the MNIST dataset (2D space). In the table, the high-dimensional (HD) space was included in order to show the performance of the indices in the original high-dimensional space. Moreover, an average of rank performances (denoted as AVG Rank) of the indices values was calculated, where the lowest AVG Rank value represents the algorithm evaluated as the best DR method (as the one that offers the best group separability of the samples) by a higher number of indices.*

**TABLE 8.** Values of the different indices obtained for the MNIST dataset in a three-dimensional (3D) space.

| Index/Method | t-SNE | HD | Isomap | MDSbc | MCE | PCA | NMDS |
|---|---|---|---|---|---|---|---|
| **PSI-P** | 4.36E-06 | 7.50E-09 | 2.05E-03 | 0.08 | 0.01 | 0.02 | 0.13 |
| **PSI-ROC** | 0.92 | 0.98 | 0.86 | 0.84 | 0.83 | 0.83 | 0.82 |
| **PSI-PR** | 0.91 | 0.98 | 0.83 | 0.83 | 0.78 | 0.83 | 0.79 |
| **PSI-MCC** | 0.79 | 0.85 | 0.66 | 0.62 | 0.63 | 0.60 | 0.60 |
| **DI** | 0.17 | 0.50 | 0.07 | 0.05 | 0.01 | 0.05 | 0.04 |
| **DB*** | 0.40 | 0.24 | 0.28 | 0.20 | 0.20 | 0.24 | 0.18 |
| **GDI** | 0.76 | 0.66 | 0.57 | 0.66 | 0.56 | 0.66 | 0.66 |
| **CH** | 116.13 | 10.22 | 55.19 | 38.53 | 80.78 | 35.11 | 35.11 |
| **SIL** | 0.38 | 0.07 | 0.04 | -0.01 | -0.06 | -0.01 | -0.01 |
| **GSI** | 0.87 | 0.82 | 0.60 | 0.49 | 0.70 | 0.42 | 0.42 |
| **CVDD** | 3.45 | 0.03 | 0.01 | 37.38 | 0.29 | 0.01 | 0.01 |
| **AVG Rank** | 1.55 | 2.45 | 3.45 | 4.36 | 4.91 | 5.00 | 5.73 |

*This table shows the best DR validation values of the different indices for the MNIST dataset (3D space). In the table, the high-dimensional (HD) space was included in order to show the performance of the indices in the original high-dimensional space. Moreover, an average of rank performances (denoted as AVG Rank) of the indices values was calculated, where the lowest AVG Rank value represents the algorithm evaluated as the best DR method (as the one that offers the best group separability of the samples) by a higher number of indices.*

is often the result of a ''computational trade-off,'' which sacrifices some properties to preserve others. For example, PCA tries to maintain linear structures, classical MDS tries to maintain global geometry, t-SNE tries to preserve local properties and local density of the data, MCE tries to preserve hierarchy, and so on [80]. In this context, we propose PS as a novel rationale for designing separability measures tailored to evaluate and compare the performance of different DR methods or the performance of a single DR method across different dimensions of embedding in relation to their group separability. Based on this rationale, we also propose a new class of indices called projection separability indices (PSIs) to evaluate the group separability performance. Currently, this class includes four indices named PSI-P, PSI-ROC, PSI-PR, and PSI-MCC, which are based on four accepted statistical measures widely adopted in machine learning. However, in the future, other indices could be proposed by simply implementing any other statistical measure (e.g., the Pearson correlation coefficient [81]) according to the PS rationale and its methodology.

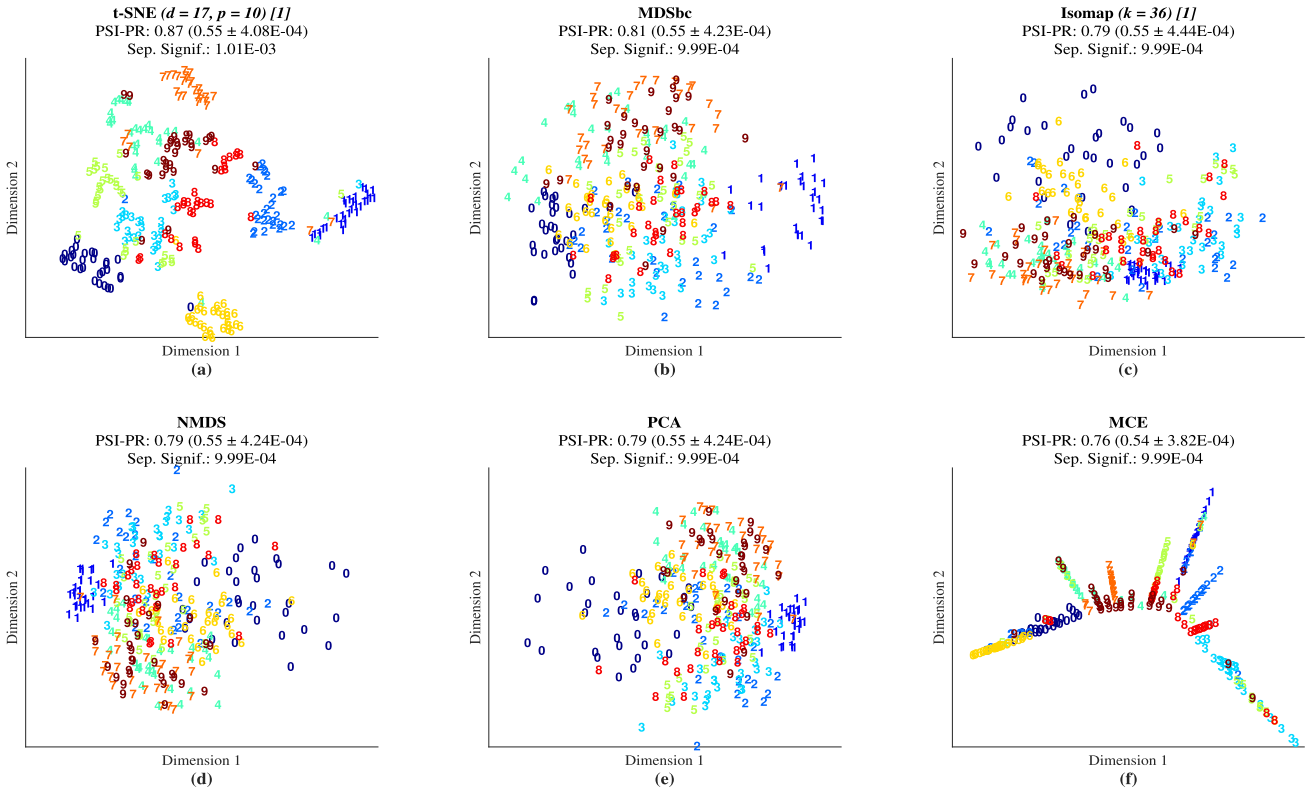



**FIGURE 9.** PSI-PR of the DR methods applied to the MNIST dataset in a two-dimensional (2D) space. The DR methods were sorted from the best (top left) to the worst (bottom right) in (a)-(f) according to the PSI-PR validation value. In (a) and (c), the optimal hyperparameters are specified in parentheses next to the title of each DR method; moreover, the digit in brackets represents the number of possible solutions with the same validation value. Furthermore, next to the value returned by this index, it is reported in parentheses the random baseline, which is the mean and the standard error of the PSI-PRs computed by randomly re-shuffling the class labels 1000 times. Finally, we added a p-value which indicates the separability significance of each index in comparison to a respective null model computed by random re-shuffling.

We compared the effectiveness of the four proposed PSIs with several representative cluster validity indices (CVIs) and a geometrically-driven separability index in: (i) assessing group separability on a synthetic 2D dataset composed of two half-moon clusters in the presence of a nonlinear (and more specifically, curvilinear) pattern with increasing levels of isotropic noise and anisotropic noise (outliers); and (ii) evaluating different DR methods across multiple datasets.

On the synthetic 2D dataset, PSI-ROC, PSI-PR, and PSI-P can detect high group separability, even in the presence of the half-moons' nonlinear (curvilinear) pattern, and they are robust to noise and outliers. PSI-MCC, SIL, and DB* are also stable to noise, but they are partially affected by the presence of nonlinearity. GSI can also detect group separability in the presence of nonlinearity and outliers, but it seems to suffer more than the other indices in the presence of noise. DI, GDI, CH, and CVDD cannot be assessed for their ability to detect separability in the presence of nonlinearity because they are not bounded measures; however, except for GDI, they all seem to suffer, to a certain extent, from increasing levels of noise. Moreover, CH and CVDD seem to suffer significantly from the presence of outliers.

In the evaluation of DR methods, some CVIs (for instance, DI and GDI) reported unreliable group separability results whose values were equal to or even lower than their random baseline (and in most of these cases, the trustworthiness was not significant, i.e., p-value $> 0.01$) in the analysis of the Gastric mucosa microbiome, Radar signal, Image proteomics, and MNIST datasets. In contrast to these indices, our PSIs obtained higher evaluation values than a random permutation of the sample labels for all analyzed cases. This is evidence that the PSIs can better detect and evaluate group separability in comparison to DI and GDI. However, based on the analyses of the Gastric mucosa microbiome dataset, we must mention that PSI-MCC, PSI-ROC, and PSI-P should be used cautiously in a reduced number of samples. For instance, it has been documented that AUC-ROC could be wrongly estimated for a small number of samples, particularly so for sample sizes $\leq$ 100 [82]. Also, it has been shown that for small sample sizes, other methods (e.g., Confusion Entropy or CEN) have a higher discriminant power than MCC [83]. In contrast, AUC-PR performs better when dealing with class-imbalanced data and small sample sizes [84], [85]. Thus, for datasets with a reduced number of samples, we suggest the use of PSI-PR to validate the group separability.

A clear issue with the CVIs emerged during the analysis of the Tripartite-Swiss-Roll dataset, in which all indices except for GSI and our PSIs returned high values in the presence of collapsed groups. In fact, GSI and the PSIs do not maximize the geometrical compactness of the groups

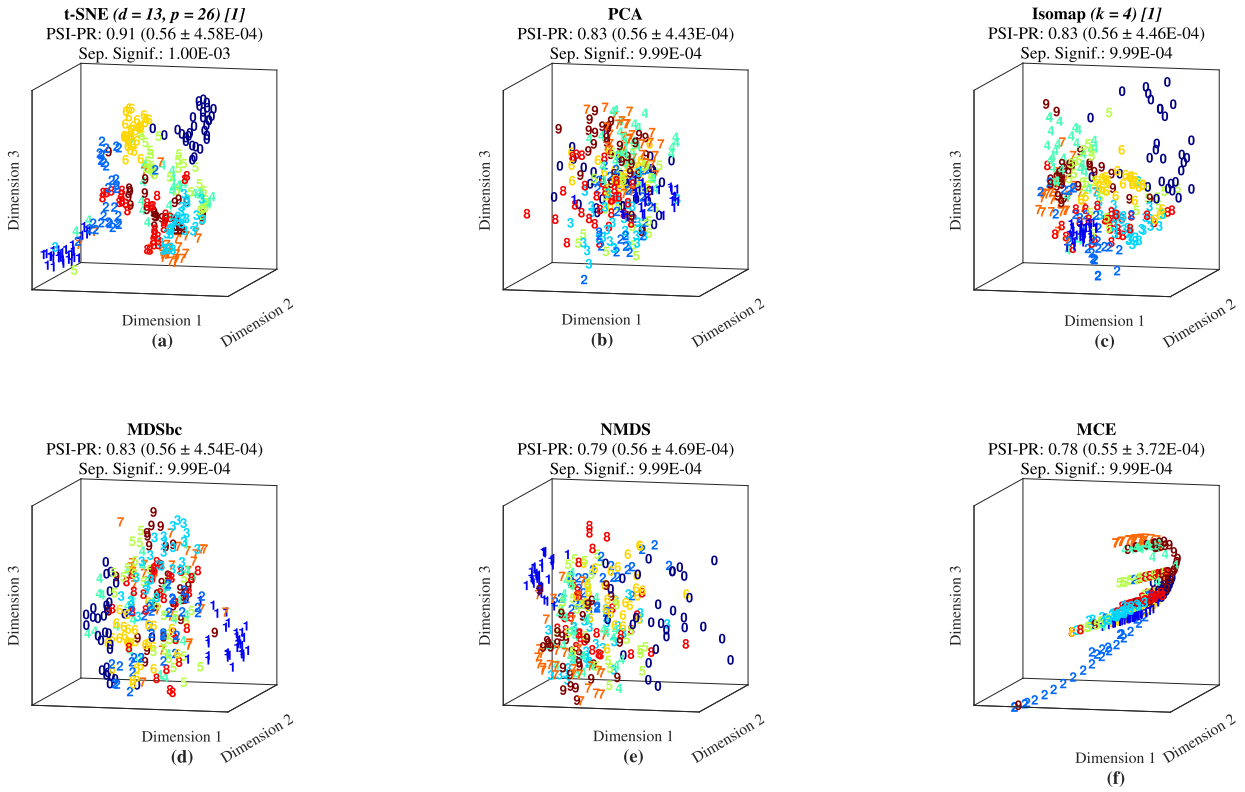



**FIGURE 10.** PSI-PR of the DR methods applied to the MNIST dataset in a three-dimensional (3D) space. The DR methods were sorted from the best (top left) to the worst (bottom right) in (a)-(f) according to the PSI-PR validation value. In (a) and (b), the optimal hyperparameters are specified in parentheses next to the title of each DR method; moreover, the digit in brackets represents the number of possible solutions with the same validation value. Furthermore, next to the value returned by this index, it is reported in parentheses the random baseline, which is the mean and the standard error of the PSI-PRs computed by randomly re-shuffling the class labels 1000 times. Finally, we added a p-value which indicates the separability significance of each index in comparison to a respective null model computed by random re-shuffling.

because they do not reward dimension reduction results in which the samples of each class tend to collapse at one point (each group having a different collapsing point). Thus, evaluating group separability with these indices reduces the potential risk of canceling the visual evidence of intra-group diversity after dimension reduction. Moreover, DI, GDI, SIL, CH, CVDD, and DB* were unable to detect multiple settings of hyperparameters for parameter-dependent algorithms such as Isomap and t-SNE. Instead, GSI and the PSIs were capable of identifying multiple settings of hyperparameters for Isomap, which offered a comparable group separability (Suppl. Fig 11-15), in agreement with the visual perception of the results. In addition, these indices could detect multiple settings of hyperparameters for t-SNE (Suppl. Fig 16-20). Notably, some of the hyperparameters settings for Isomap and t-SNE selected by GSI and the PSIs exhibited more accurate representations of the original nonlinear structure by conserving the initial color gradient and a clear visualization of the groups. However, some of the hyperparameters configurations evaluated by GSI in relation to t-SNE had a questionable group separability (Suppl. Fig. 20a-h). This situation was also encountered in the analysis of the Image proteomic dataset, where GSI selected more hyperparameters configurations than the PSIs for t-SNE. As mentioned early in the text, t-SNE is a parameter-dependent algorithm, with perplexity being one of the hyperparameters to tune. Regarding this parameter, it has been frequently observed that when the perplexity is set to a small value, ''artificial microclusters'' begin forming in t-SNE plots [86]. Indeed, most of the extra hyperparameters solutions provided by GSI while evaluating t-SNE are in the presence of a low perplexity (e.g., Suppl. Fig. 20a-h), meaning that this index is prone to capture these artificial clusters and evaluate them as part of a ''correct'' group separability, which obviously, should not be the case and it might represent a point of failure for this index. Instead, the PSIs validated hyperparameters settings with a more precise group separability, avoiding the inclusion of these artificial microclusters in the evaluation. Therefore, the PSIs are more accurate than GSI for the assessment and validation of optimal hyperparameters settings. Hence, our novel class of indices can also be employed to enhance and automate the tuning process of hyperparameters in dimension reduction algorithms. However, we should also consider that in some cases, there might be more information associated with subgroups in the data that can be lost by forcing the group separability with different hyperparameters. Therefore, parameter-free methods for nonlinear DR, such as MCE, can help to identify hidden patterns from the original data, reducing the risk of neglecting unknown subgroup separations because of hyperparameter bias. In this case, the PSIs can still

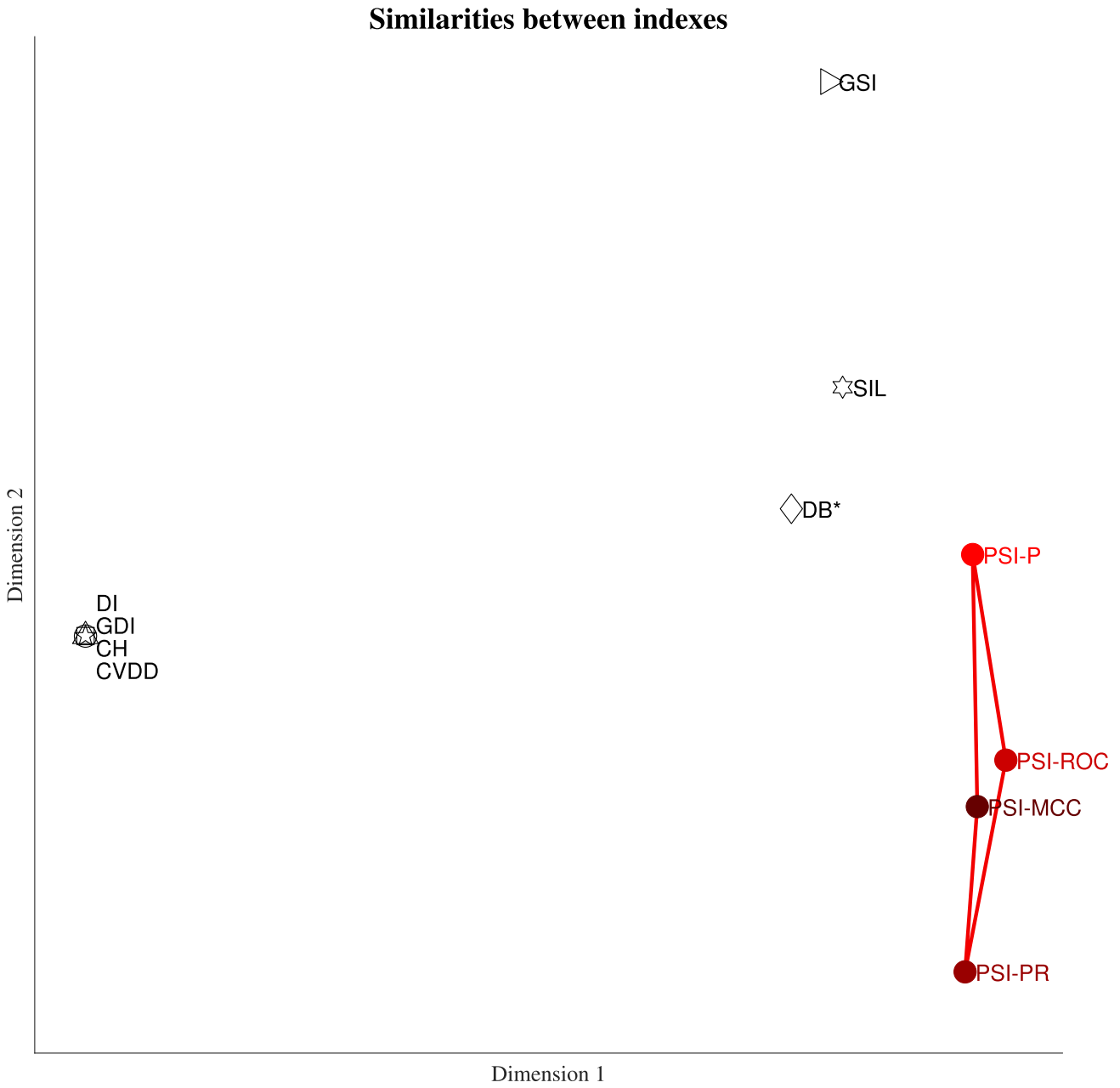



**FIGURE 11.** Similarities of the indices across all the datasets. We create a matrix by merging row vectors, each of which reports the result of a different separability index in all the datasets for all the DR methods. Then, centered PCA is applied to this matrix (after z-scoring the indices' values for each row because each index has a different scale). The first two principal components are plotted to visualize the similarities between the indices. PSIs indices are red marks, CVIs indices are black marks. The PSI-quadrilateral (quadrilateral whose vertices are the PSI indices; PSI-P, PSI-ROC, PSI-PR, and PSI-MCC) is drawn with the aim to show which indices gave comparable results to PSI because they are projected inside the quadrilateral. It is evident that all indices are far from the quadrilateral area; therefore, the proposed PSI indices represent a novel separated class of indices for the evaluation of group separability in a geometrical space.

play a key role in the evaluation of parameter-free algorithms because they can help to detect the dimensions in which the separability of groups emerges, where traditional CVIs are not tailored and are less reliable for comparing results across different dimensions, as explained in the Results section above.

One of the most critical issues in evaluating group separability is to deal with outliers (because they can heavily affect the performance of any index), as shown in Suppl. Fig. 95-100, we can see how the PSIs' values were corrected in each dataset by applying the mathematical formulations proposed in (24). These corrections allow our PSIs to address outliers and provide evidence that we can use statistical measures, indistinctly if they maximize or minimize, to evaluate group separability.

Moreover, we analyzed the similarities between the indices, initially separately for each dataset (Suppl. Fig. 101) and then collectively for all datasets together (Fig. 11). These

analyses revealed that DB*, SIL, and GSI were closer to our PSIs than were the other indices. However, the PSIs were able to overcome the limitations of these indices, such as overlapping, outliers, and incorrect evaluation of microclusters.

In addition, it is enlightening to clarify the reason why the PSIs might offer advantages over the other indices, also considering the theoretical standpoint rather than commenting on computational results only. One of the main theoretical advantages of the PSIs is that they can inherit the boundaries (either upper or lower bounds) from the statistical measures on which they are based, whereas other indices are either boundless or their boundaries do not necessarily address linear separability. In the case of boundless indices (as most of CVIs), they are sensitive to changes in scale, and therefore are not tailored for comparison of results across different dimensions (for instance, the values of these indices applied in two dimensions are not directly comparable with the values of the same indices applied in three dimensions because there is no limit that fixes the comparison). However, indices with boundaries such as GSI and SIL (being a [0,1] range for GSI and a $[-1,1]$ range for SIL) are not exempt from problems. In the case of GSI, as defined in (20), this index looks, for each point, to the first closest neighbor and checks whether their class memberships match. This skews the index to incorrectly consider and evaluate artificial microclusters with optimal separability (as perfectly segregated). In the case of SIL, as noted in (1), for a set of clusters, the value of this index is simply the mean of the silhouette width of each cluster (i.e., how tightly grouped are all the points in the cluster). This definition can cause problems, as the mean is easily skewed by outliers, leading this index to misrepresent how appropriately the data have been clustered, i.e., the effect of a small number of outliers can cause a significant change [37], and thus, an incorrect assessment of the group separability.

Finally, we need to comment on the detection of linear separability, which is a very important concept because the data could be nonlinearly related (see, for instance, the iconic example of the Tripartite-Swiss-Roll dataset) in the original feature space. In this case, linear transformations, such as PCA, fail to provide a low-dimensional reduction representation in which different groups are geometrically separated. However, the same data may be linearized after the application of a nonlinear dimension reduction algorithm, which means that the algorithm for nonlinear dimension reduction can address the problem of data nonlinearity, revealing the presence of linearly separated groups in a 2D or 3D data representation. PSIs are effective tools for addressing whether, after dimension reduction, the data representation reveals a linear separation between the sample groups. This is a direct consequence of the introduction in this study of the innovative theoretical concept of projection separability. More precisely, the fact that a PSI achieves a value close to the significant bound of the statistical measure on which it is based (this is 0 for PSI-P and 1 for all the other PSIs discussed in this study) is a sufficient (but not necessarily) condition to claim linear separability. This means that there might be some solutions of linear separability that a PSI does not detect. However, if a PSI takes a value equal to the significant bound of the statistical measure on which it is based, this is undoubtedly an indication of linear separability. In the case of GSI, the upper bound value of 1 might indicate linear but also nonlinear separability; hence, this index is not reliable for detecting linear separability. In the case of SIL, the upper bound value of 1 indicates only a degenerate case of linear separability that occurs when all the points in a group collapse on its centroid; therefore, this index is not useful for detecting linear separability. None of the other indices has an upper (or lower) bound, and they cannot detect linear separability as the best performance. Table 1, which we introduced at the beginning of the study, also summarizes the new findings discussed in this section, offering an overview of all group separability indices and their properties investigated in this study. Unlikely, none of the reported methods in literature can achieve (in Table 1, the symbol means: Yes, it can achieve) all main properties while evaluating data partitioning: bounded, effective for overlapping groups, arbitrary (e.g., nonspherical) shapes, linearity detection, robustness to nonlinear (curvilinear) pattern, and isotropic/anisotropic noise. In contrast, the proposed projection separability indices (PSIs), which are designed according to the described projection separability (PS) rationale, can address these limitations, and provide a novel and more reliable class of statistical estimators for group separability.

## V. CONCLUSION

There is no universal method to map a given dataset from a high-dimensional space into a reduced number of dimensions by preserving all its original properties. Despite the attempts of different dimension reduction (DR) techniques to preserve all original properties, many of these algorithms may partially fail this task. Thus, the accurate evaluation of such methods remains a challenge. To overcome this, we propose a novel rationale called projection separability (PS), which is specifically tailored to evaluate the performance of DR methods. Based on this rationale, we implemented a new class of statistical-based indices named projection separability indices (PSIs).

The experimental results indicate that the PSIs are better evaluators of the group separability returned by the DR results than other approaches, such as cluster validity indices (CVIs). Furthermore, these results also provide evidence that PSIs are not heavily affected by the limitations of other indices such as overlapping groups, arbitrary shapes, detection of linear separability, nonlinear (curvilinear) pattern, and isotropic/anisotropic noise (Table 1). Thus, we propose the exploitation of the PS rationale (either through the described statistical measures or based on others such as the Pearson correlation coefficient or F-score) as a valid framework for: I) automatic evaluation and identification of the best dimension reduction methods for a certain problem or dataset, II) automatic detection of the best hyperparameters of parameter-dependent DR algorithms, and III) automatic

detection of the best normalization associated with a particular dataset in relation to a given DR method.

## ACKNOWLEDGMENT

The authors thank the High-Performance Computing and Storage Complex (HRSK-II) of the Centre for Information Services and High-performance Computing (ZIH) of the Technische Universität Dresden for the computational power.

## REFERENCES

[1] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 15, no. 6, pp. 1373–1396, Jun. 2003, doi: 10.1162/089976603321780317.

[2] C. V. Cannistraci, T. Ravasi, F. M. Montevecchi, T. Ideker, and M. Alessio, "Nonlinear dimension reduction and clustering by minimum curvilinearity unfold neuropathic pain and tissue embryological classes," *Bioinformatics*, vol. 26, no. 18, pp. i531–i539, Sep. 2010, doi: 10.1093/bioinformatics/btq376.

[3] A. Muscoloni, J. M. Thomas, S. Ciucci, G. Bianconi, and C. V. Cannistraci, "Machine learning meets complex networks via coalescent embedding in the hyperbolic space," *Nature Commun.*, vol. 8, no. 1, p. 1615, Dec. 2017, doi: 10.1038/s41467-017-01825-5.

[4] S. Ciucci, Y. Ge, C. Durán, A. Palladini, V. Jiménez-Jiménez, L. M. Martínez-Sánchez, Y. Wang, S. Sales, A. Shevchenko, S. W. Poser, M. Herbig, O. Otto, A. Androutsellis-Theotokis, J. Guck, M. J. Gerl, and C. V. Cannistraci, "Enlightening discriminative network functional modules behind principal component analysis separation in differential-omic science studies," *Sci. Rep.*, vol. 7, no. 1, p. 43946, Mar. 2017, doi: 10.1038/srep43946.

[5] E. Becht, L. McInnes, J. Healy, C. A. Dutertre, I. W. H. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell, "Dimensionality reduction for visualizing single-cell data using UMAP," *Nat. Biotechnol.*, vol. 37, no. 1, pp. 38–47, Jan. 2019, doi: 10.1038/nbt.4314.

[6] X. Xia, X. Chen, G. Wu, F. Li, Y. Wang, Y. Chen, M. Chen, X. Wang, W. Chen, B. Xian, and W. Chen, "Three-dimensional facial-image analysis to predict heterogeneity of the human ageing rate and the impact of lifestyle," *Nature Metabolism*, vol. 2, no. 9, pp. 946–957, Sep. 2020, doi: 10.1038/s42255-020-00270-x.

[7] P. Bao, L. She, M. McGill, and D. Y. Tsao, "A map of object space in primate inferotemporal cortex," *Nature*, vol. 583, no. 7814, pp. 103–108, Jul. 2020, doi: 10.1038/s41586-020-2350-5.

[8] V. Narula, A. G. Zippo, A. Muscoloni, G. E. M. Biella, and C. V. Cannistraci, "Can local-community-paradigm and epitopological learning enhance our understanding of how local brain connectivity is able to process, learn and memorize chronic pain?" *Appl. Netw. Sci.*, vol. 2, no. 1, pp. 1–28, Dec. 2017, doi: 10.1007/s41109-017-0048-x.

[9] S. Yan, D. Xu, B. Zhang, and H.-J. Zhang, "Graph embedding: A general framework for dimensionality reduction," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 830–837, doi: 10.1109/CVPR.2005.170.

[10] P. Ray, S. S. Reddy, and T. Banerjee, "Various dimension reduction techniques for high dimensional data analysis: A review," *Artif. Intell. Rev.*, vol. 54, no. 5, pp. 3473–3515, Jun. 2021, doi: 10.1007/s10462-020-09928-0.

[11] M. Espadoto, R. M. Martins, A. Kerren, N. S. T. Hirata, and A. C. Telea, "Toward a quantitative survey of dimension reduction techniques," *IEEE Trans. Vis. Comput. Graphics*, vol. 27, no. 3, pp. 2153–2173, Mar. 2021, doi: 10.1109/TVCG.2019.2944182.

[12] G. Alanis-Lobato, C. V. Cannistraci, A. Eriksson, A. Manica, and T. Ravasi, "Highlighting nonlinear patterns in population genetics datasets," *Sci. Rep.*, vol. 5, no. 1, pp. 1–8, Jul. 2015, doi: 10.1038/srep08140.

[13] F. R. S. K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space," *London, Edinburgh, Dublin Philosoph. Mag. J. Sci.*, vol. 2, no. 11, pp. 559–572, 1901, doi: 10.1080/14786440109462720.

[14] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *J. Educ. Psychol.*, vol. 24, no. 6, pp. 417–441, Sep. 1933, doi: 10.1037/h0071325.

[15] W. S. Torgerson, "Multidimensional scaling: I. Theory and method," *Psychometrika*, vol. 17, no. 4, pp. 401–419, Dec. 1952, doi: 10.1007/BF02288916.

[16] K. Bunte, S. Haase, M. Biehl, and T. Villmann, "Stochastic neighbor embedding (SNE) for dimension reduction and visualization using arbitrary divergences," *Neurocomputing*, vol. 90, pp. 23–45, Aug. 2012, doi: 10.1016/j.neucom.2012.02.034.

[17] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, Nov. 2008. Accessed: Oct. 25, 2018. [Online]. Available: http://www.jmlr.org/papers/v9/vandermaaten08a.html.

[18] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction.," *Science*, vol. 290, no. 5500, pp. 23–2319, Dec. 2000, doi: 10.1126/science.290.5500.2319.

[19] Z. Liu, K. Shi, K. Zhang, W. Ou, and L. Wang, "Discriminative sparse embedding based on adaptive graph for dimension reduction," *Eng. Appl. Artif. Intell.*, vol. 94, Sep. 2020, Art. no. 103758, doi: 10.1016/j.engappai.2020.103758.

[20] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," 2018, *arXiv:1802.03426*.

[21] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, Sep. 1999, doi: 10.1145/331499.331504.

[22] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, and I. Perona, "An extensive comparative study of cluster validity indices," *Pattern Recognit.*, vol. 46, no. 1, pp. 243–256, Jan. 2013, doi: 10.1016/j.patcog.2012.07.021.

[23] K. Kryszczuk and P. Hurley, "Estimation of the number of clusters using multiple clustering validity indices," in *Multiple Classifier Systems*, N. El Gayar, J. Kittler, and F. Roli, Eds. Berlin, Germany: Springer, 2010, pp. 114–123.

[24] S. Guan and M. Loew, "An internal cluster validity index using a distance-based separability measure," in *Proc. IEEE 32nd Int. Conf. Tools with Artif. Intell. (ICTAI)*, Nov. 2020, pp. 827–834, doi: 10.1109/ICTAI50040.2020.00131.

[25] F. Raimundo, C. Vallot, and J.-P. Vert, "Tuning parameters of dimensionality reduction methods for single-cell RNA-seq analysis," *Genome Biol.*, vol. 21, no. 1, pp. 1–17, Aug. 2020, doi: 10.1186/s13059-020-02128-7.

[26] R. Xiang, W. Wang, L. Yang, S. Wang, C. Xu, and X. Chen, "A comparison for dimensionality reduction methods of single-cell RNA-seq data," *Frontiers Genet.*, vol. 12, p. 320, Mar. 2021, doi: 10.3389/fgene.2021.646936.

[27] L. Zhang, L. Lin, and J. Li, "CPS analysis: Self-contained validation of biomedical data clustering," *Bioinformatics*, vol. 36, no. 11, pp. 3516–3521, Jun. 2020, doi: 10.1093/bioinformatics/btaa165.

[28] J. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters," *J. Cybern.*, vol. 3, no. 3, pp. 32–57, 1973, doi: 10.1080/01969727308546046.

[29] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979, doi: 10.1109/TPAMI.1979.4766909.

[30] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Commun. Stat., Theory Methods*, vol. 3, no. 1, pp. 1–27, 1974, doi: 10.1080/03610917408548446.

[31] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, no. 1, pp. 53–65, Nov. 1987, doi: 10.1016/0377-0427(87)90125-7.

[32] J. C. Bezdek and N. R. Pal, "Cluster validation with generalized Dunn's indices," in *Proc. 2nd New Zealand Int. Two-Stream Conf. Artif. Neural Netw. Expert Syst.*, Nov. 1995, pp. 190–193, doi: 10.1109/ANNES.1995.499469.

[33] L. Hu and C. Zhong, "An internal validity index based on density-involved distance," *IEEE Access*, vol. 7, pp. 40038–40051, 2019, doi: 10.1109/ACCESS.2019.2906949.

[34] C. Thornton, *Separability is a Learner's Best Friend*. London, U.K.: Springer, 1998, pp. 40–46.

[35] H. Le Capitaine and C. Frelicot, "A cluster-validity index combining an overlap measure and a separation measure based on fuzzy-aggregation operators," *IEEE Trans. Fuzzy Syst.*, vol. 19, no. 3, pp. 580–588, Jun. 2011, doi: 10.1109/TFUZZ.2011.2106216.

[36] A. B. Said, R. Hadjidj, and S. Foufou, "Cluster validity index based on Jeffrey divergence," *Pattern Anal. Appl.*, vol. 20, no. 1, pp. 21–31, Feb. 2017, doi: 10.1007/s10044-015-0453-7.

[37] R. Layton, P. Watters, and R. Dazeley, "Evaluating authorship distance methods using the positive silhouette coefficient," *Natural Lang. Eng.*, vol. 19, no. 4, pp. 517–535, Oct. 2013, doi: 10.1017/S1351324912000241.

[38] S.-H. Lee, Y.-S. Jeong, J.-Y. Kim, and M. K. Jeong, "A new clustering validity index for arbitrary shape of clusters," *Pattern Recognit. Lett.*, vol. 112, pp. 263–269, Sep. 2018, doi: 10.1016/j.patrec.2018.08.005.
[39] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Ann. Math. Statist.*, vol. 18, no. 1, pp. 50–60, 1947, doi: 10.1214/aoms/1177730491.
[40] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, Apr. 1982, doi: 10.1148/radiology.143.1.7063747.
[41] V. Raghavan, P. Bollmann, and G. S. Jung, "A critical investigation of recall and precision as measures of retrieval system performance," *ACM Trans. Inf. Syst.*, vol. 7, no. 3, pp. 205–229, Jul. 1989, doi: 10.1145/65943.65945.
[42] B. W. Matthews, "Comparison of the predicted and observed secondary structure of t4 phage lysozyme," *Biochimica Biophysica Acta-Protein Struct.*, vol. 405, no. 2, pp. 442–451, 1975, doi: 10.1016/0005-2795(75)90109-9.
[43] C. O. S. Sorzano, J. Vargas, and A. P. Montano, "A survey of dimensionality reduction techniques," 2014, *arXiv:1403.2877*.
[44] R. N. Shepard, "The analysis of proximities: Multidimensional scaling with an unknown distance function. I.," *Psychometrika*, vol. 27, no. 2, pp. 125–140, Jun. 1962, doi: 10.1007/BF02289630.
[45] J. B. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, vol. 29, no. 1, pp. 1–27, Mar. 1964, doi: 10.1007/BF02289565.
[46] E. W. Beals, "Bray-curtis ordination: An effective strategy for analysis of multivariate ecological data," *Adv. Ecol. Res.*, vol. 14, pp. 1–55, Jan. 1984, doi: 10.1016/S0065-2504(08)60168-3.
[47] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, and J. Vanderplas, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011.
[48] F. P. Sterbini, A. Palladini, L. Masucci, C. V. Cannistraci, R. Pastorino, G. Ianiro, F. Bugli, C. Martini, W. Ricciardi, A. Gasbarrini, M. Sanguinetti, G. Cammarota, and B. Posteraro, "Effects of proton pump inhibitors on the gastric mucosa-associated microbiota in dyspeptic patients," *Appl. Environ. Microbiol.*, vol. 82, no. 22, pp. 6633–6644, Nov. 2016, doi: 10.1128/AEM.01437-16.
[49] V. G. Sigillito, S. P. Wing, L. V. Hutton, and K. B. Baker, "Classification of radar returns from the ionosphere using neural networks," *Johns Hopkins APL Tech. Dig.*, vol. 10, pp. 262–266, Jul. 1989.
[50] A. D. Shieh, T. B. Hashimoto, and E. M. Airoldi, "Tree preserving embedding," *Proc. Nat. Acad. Sci. USA*, vol. 108, no. 41, pp. 16916–16921, Oct. 2011, doi: 10.1073/pnas.1018393108.
[51] C. V. Cannistraci, G. Alanis-Lobato, and T. Ravasi, "Minimum curvilinearity to enhance topological prediction of protein interactions by network embedding," *Bioinformatics*, vol. 29, no. 13, pp. 199–209, 2013, doi: 10.1093/bioinformatics/btt208.
[52] A. Conti, P. Ricchiuto, S. Iannaccone, B. Sferrazza, A. Cattaneo, A. Bachi, A. Reggiani, M. Beltramo, and M. Alessio, "Pigment epithelium-derived factor is differentially expressed in peripheral neuropathies," *Proteomics*, vol. 5, no. 17, pp. 4558–4567, Sep. 2005, doi: 10.1002/pmic.200402088.
[53] A. Conti, S. Iannaccone, B. Sferrazza, L. De Monte, S. Cappa, D. Franciotta, S. Olivieri, and M. Alessio, "Differential expression of ceruloplasmin isoforms in the cerebrospinal fluid of amyotrophic lateral sclerosis patients," *Proteomics Clin. Appl.*, vol. 2, no. 12, pp. 1628–1637, Dec. 2008, doi: 10.1002/prca.200780081.
[54] K. Das and Z. Nenadic, "An efficient discriminant-based solution for small sample size problem," *Pattern Recognit.*, vol. 42, no. 5, pp. 857–866, May 2009, doi: 10.1016/j.patcog.2008.08.036.
[55] N. Altman and M. Krzywinski, "The curse(s) of dimensionality," *Nature Methods*, vol. 15, no. 6, pp. 399–400, Jun. 2018, doi: 10.1038/s41592-018-0019-x.
[56] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2323, Nov. 1998, doi: 10.1109/5.726791.
[57] N. Tahiri, M. Willems, and V. Makarenkov, "A new fast method for inferring multiple consensus trees using k-medoids," *BMC Evol. Biol.*, vol. 18, no. 1, p. 48, Dec. 2018, doi: 10.1186/s12862-018-1163-8.
[58] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, no. 2, pp. 179–188, Sep. 1936, doi: 10.1111/j.1469-1809.1936.tb02137.x.
[59] W. S. Noble, "What is a support vector machine?" *Nature Biotechnol.*, vol. 24, no. 12, pp. 1565–1567, Dec. 2006, doi: 10.1038/nbt1206-1565.
[60] A. Abdiansah and R. Wardoyo, "Time complexity analysis of support vector machines (SVM) in LibSVM," *Int. J. Comput. Appl.*, vol. 128, no. 3, pp. 28–34, Oct. 2015, doi: 10.5120/ijca2015906480.
[61] D. Cai, X. He, and J. Han, "Training linear discriminant analysis in linear time," in *Proc. IEEE 24th Int. Conf. Data Eng.*, Apr. 2008, pp. 209–217, doi: 10.1109/ICDE.2008.4497429.
[62] I. W. Tsang, J. T. Kwok, P.-M. Cheung, and N. Cristianini, "Core vector machines: Fast SVM training on very large data sets," *J. Mach. Learn. Res.*, vol. 6, no. 4, pp. 363–392, 2005.
[63] J. Schaffer, "What not to multiply without necessity," *Australas. J. Philosophy*, vol. 93, no. 4, pp. 644–664, Oct. 2015, doi: 10.1080/00048402.2014.992447.
[64] M. Blum, R. W. Floyd, V. Pratt, R. L. Rivest, and R. E. Tarjan, "Time bounds for selection," *J. Comput. Syst. Sci.*, vol. 7, no. 4, pp. 448–461, Aug. 1973.
[65] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and F-score, with implication for evaluation," in *Advances in Information Retrieval*, D. E. Losada and J. M. Fernández-Luna, Eds. Berlin, Germany: Springer, 2005, pp. 345–359.
[66] N. P. Pérez, M. A. Guevara López, A. Silva, and I. Ramos, "Improving the Mann–Whitney statistical test for feature selection: An approach in breast cancer diagnosis on mammography," *Artif. Intell. Med.*, vol. 63, no. 1, pp. 19–31, Jan. 2015, doi: 10.1016/j.artmed.2014.12.004.
[67] R. C. Milton, "An extended table of critical values for the Mann–Whitney (Wilcoxon) two-sample statistic," *J. Amer. Stat. Assoc.*, vol. 59, no. 307, pp. 925–934, Sep. 1964, doi: 10.1080/01621459.1964.10480740.
[68] N. Nagarajan and U. Keich, "Reliability and efficiency of algorithms for computing the significance of the Mann–Whitney test," *Comput. Statist.*, vol. 24, no. 4, pp. 605–622, Mar. 2009, doi: 10.1007/s00180-009-0148-x.
[69] M. Natole, Y. Ying, and S. Lyu, "Stochastic AUC optimization algorithms with linear convergence," *Frontiers Appl. Math. Statist.*, vol. 5, p. 30, Jun. 2019, doi: 10.3389/fams.2019.00030.
[70] B. Zhou, Y. Ying, and S. Skiena, "Online AUC optimization for sparse high-dimensional datasets," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2020, pp. 881–890, doi: 10.1109/ICDM50108.2020.00097.
[71] B. Szörényi, S. Cohen, and S. Mannor, "Non-parametric Online AUC maximization," in *Machine Learning and Knowledge Discovery in Databases*, M. Ceci, J. Hollmén, L. Todorovski, C. Vens, and S. Džeroski, Eds. Cham, Switzerland: Springer, 2017, pp. 575–590.
[72] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *Proc. 23rd Int. Conf. Mach. Learn. (ICML)*, 2006, pp. 233–240.
[73] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *J. Roy. Statist. Soc., B*, vol. 57, pp. 289–300, Jan. 1995, doi: 10.2307/2346101.
[74] L. Van Der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2625, 2008.
[75] L. van der Maaten, "Accelerating t-SNE using tree-based algorithms," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3221–3245, Oct. 2014, doi: 10.5555/2627435.
[76] S. Chaimontree, K. Atkinson, and F. Coenen, "Best clustering configuration metrics: Towards multiagent based clustering," in *Advanced Data Mining and Applications*, L. Cao, Y. Feng, and J. Zhong, Eds. Berlin, Germany: Springer, 2010, pp. 48–59.
[77] S. Petrovic, "A comparison between the silhouette index and the Davies-Bouldin index in labelling IDS clusters," in *Proc. 11th Nord. Work. Secur. IT-Syst.*, 2006, pp. 53–64. [Online]. Available: https://xp-dev.com/svn/b_frydrych../silhuetteIndexRegulaStopu.pdf
[78] M. Kim and R. S. Ramakrishna, "New indices for cluster validity assessment," *Pattern Recognit. Lett.*, vol. 26, no. 15, pp. 2353–2363, Nov. 2005, doi: 10.1016/j.patrec.2005.04.007.
[79] S. Saitta, B. Raphael, and I. F. C. Smith, "A bounded index for cluster validity," in *Machine Learning and Data Mining in Pattern Recognition*, P. Perner, Ed. Berlin, Germany: Springer, 2007, pp. 174–187.
[80] L. H. Nguyen and S. Holmes, "Ten quick tips for effective dimensionality reduction," *PLOS Comput. Biol.*, vol. 15, no. 6, Jun. 2019, Art. no. e1006907, doi: 10.1371/journal.pcbi.1006907.
[81] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Springer Topics in Signal Processing*, vol. 2. Berlin, Germany: Springer, 2009, pp. 1–4.
[82] B. Hanczar, J. Hua, C. Sima, J. Weinstein, M. Bittner, and E. R. Dougherty, "Small-sample precision of ROC-related estimates," *Bioinformatics*, vol. 26, no. 6, pp. 822–830, Feb. 2010, doi: 10.1093/bioinformatics/btq037.

[83] G. Jurman, S. Riccadonna, and C. Furlanello, "A comparison of MCC and CEN error measures in multi-class prediction," *PLoS One*, vol. 7, no. 8, p. 41882, Aug. 2012, doi: 10.1371/journal.pone.0041882.

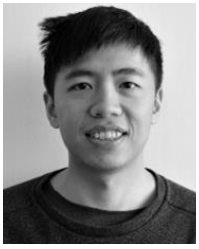
[84] G.-H. Fu, F. Xu, B.-Y. Zhang, and L.-Z. Yi, "Stable variable selection of class-imbalanced data with precision-recall criterion," *Chemometric Intell. Lab. Syst.*, vol. 171, pp. 241–250, Dec. 2017, doi: 10.1016/j.chemolab.2017.10.015.

[85] Y. Ge, P. Rosendahl, C. Duran, N. Topfner, S. Ciucci, J. Guck, and C. V. Cannistraci, "Cell mechanics based computational classification of red blood cells via machine intelligence applied to morpho-rheological markers," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 4, pp. 1405–1415, Jul. 2021, doi: 10.1109/TCBB.2019.2945762.

[86] M. Wattenberg, F. Viégas, and I. Johnson, "How to use t-SNE effectively," *Distill*, vol. 1, no. 10, p. e2, Oct. 2016, doi: 10.23915/distill.00002.

**ALDO ACEVEDO** received the Diploma in Engineering degree in bioinformatics from the University of Talca, Chile, in 2015. He worked as a Full-Stack Developer with the University of Talca, from 2015 to 2016. He is currently pursuing the Ph.D. degree in computer science with the Biomedical Cybernetics Group led by Dr. Carlo Vittorio Cannistraci, Biotechnology Center, Technische Universität Dresden, Germany. His research interests include software engineering, artificial intelligence, and bioinformatics.

**CLAUDIO DURÁN** received the Diploma in Engineering degree in bioinformatics from the University of Talca, Chile, in 2016, and the Ph.D. degree in computer science from Technische Universität Dresden, Germany, in 2021, under the supervision of Dr. Carlo Vittorio Cannistraci at the Biomedical Cybernetics Group. Currently, he is working as a Data Analyst in the lipidomics field. His research interests include machine learning, network science, and systems biomedicine.



**MING-JU KUO** received the B.Sc. degree in life science and geosciences from the National Taiwan University, Taiwan, in 2015, and the master's degree in bioengineering from the Biotechnology Center, Technische Universität Dresden, Germany, under the supervision of Dr. Carlo Vittorio Cannistraci at the Biomedical Cybernetics Group. He was a Scientific Assistant with Technische Universität Dresden. His interests include molecular biology and computational biology.

**SARA CIUCCI** received the B.Sc. degree in mathematics from the University of Padova, Italy, in 2010, the M.Sc. degree in mathematics from the University of Trento, Italy, in 2014, and the Ph.D. degree in physics from Technische Universität Dresden, Germany, in 2018, under the supervision of Dr. Carlo Vittorio Cannistraci at the Biomedical Cybernetics Group. She has been a Postdoctoral Researcher with Technische Universität Dresden, since March 2019. Her research interests include machine learning, network science, and systems biomedicine.

**MICHAEL SCHROEDER** received the degree in computer science from RWTH Aachen and the Ph.D. degree from the University of Hannover. Prior to joining TU Dresden in 2003, he was a Lecturer and a Senior Lecturer with the City, University of London. He is currently a Professor in bioinformatics with TU Dresden. He has published more than 200 scientific papers and holds two granted patents. His Hirsch index on Google Scholar is greater than 40. He obtained over ten million euros in research funding from EU, BMBF, BMWi, and DFG. He supervised over 20 Ph.D. students, of which eight obtained a distinction. His eight former group members are now a professors or a group leaders. He served on the Advisory Board of the Gene Ontology. He co-founded pharm.ai, which focuses on structure-based drug target prediction and transinsight GmbH, which developed GoPubMed, a semantic biomedical search engine. His research interests include machine learning algorithms applied to structural and sequence data to improve the diagnosis and treatment of infectious diseases and cancer.

**CARLO VITTORIO CANNISTRACI** was born in Milazzo, Sicily, Italy, in 1976. He received the M.S. degree in biomedical engineering from the Polytechnic of Milano, Italy, in 2005, and the Ph.D. degree in biomedical engineering from the Inter-Polytechnic School of Doctorate, Italy, in 2010. From 2009 to 2010, he was Visiting Scholar with the Integrative Systems Biology Laboratory of Dr. Trey Ideker, University of California San Diego (UCSD), USA. From 2010 to 2013, he was a Postdoctoral Researcher and a Research Scientist in machine intelligence and complex network science for personalized biomedicine with the King Abdullah University of Science and Technology (KAUST), Saudi Arabia. From 2014 to 2020, he was a Group Leader and the Head of the Biomedical Cybernetics Laboratory, Biotechnological Center (BIOTEC). He was affiliated with the Department of Physics and the Department of Computer Science, Technische Universität Dresden, Germany. Since 2021, he has been a Chair Professor and a Chief Scientist with the Center for Complex Network Intelligence, Tsinghua Laboratory of Brain and Intelligence, Beijing, China. He is currently an Adjunct Professor with the Department of Computer Science and the Department of Biomedical Engineering, Tsinghua University. He is also affiliated with the MPI Center for Systems Biology, Dresden, Germany. He is a Theoretical Engineer. His research interests include subjects at the interface between complex data and information, physics of complex systems, complex networks, and machine intelligence, with particular interest for applications in life and socioeconomical science. He is the author of three book chapters and more than 40 articles. TU Dresden honored him with the Young Investigator Award 2016 in physics for his recent work on the local-community-paradigm theory and link prediction in monopartite and bipartite complex networks. In 2018, *Nature Communications* featured his article titled "Machine learning meets complex networks via coalescent embedding in the hyperbolic space" in the selected interdisciplinary collection of recent research on complex systems. In 2019, he was awarded the Shanghai 1000 Plan and the Zhou Yahui Chair Professorship at Tsinghua University. In 2020, he was awarded the National High-Level Talent Program of the Ministry of Science and Technology of China.