Advances in Science and Technology Research Journal

*Advances in Science and Technology Research Journal*, 2026, 20(7), 54–70
https://doi.org/10.12913/22998624/218660
ISSN 2299-8624, License CC-BY 4.0

Received: 2026.02.04
Accepted: 2026.05.15
Published: 2026.06.01

# Comparative evaluation of persistence diagram vectorisation methods in classification tasks

Dominika Sulowska[1]

[1] Department of Computer Science, Lublin University of Technology, ul. Nadbystrzycka 36B, 20-618 Lublin, Poland
E-mail: d.sulowska@pollub.pl

## ABSTRACT

Topological data analysis (TDA) enables the analysis of the geometric structure of data using tools from algebraic topology. A central technique in TDA is persistent homology, the results of which are represented by persistence diagrams (PDs) describing the lifespan of topological features. Since PDs lack a natural vector-space representation, their direct use in machine learning (ML) classifiers is challenging. Therefore, several vectorisation methods have been proposed, including persistence image (PI), persistence landscapes (PL), Betti curves (BC), and persistence silhouettes (PS). This study presents a comparative analysis of these vectorisation methods in classification tasks involving both synthetic and real-world datasets, using three classifiers: Logistic regression (LR), XGBoost (XGB), and multilayer perceptron (MLP). Hyperparameter tuning and cross-validation were applied, and model performance was evaluated using accuracy, precision, recall, and F1-score. The results showed that PI and PL consistently achieve the highest classification performance across different data types and classifiers. For synthetic datasets, these methods reached scores above 0.98, while for the ECG dataset, they outperformed alternative approaches by up to 30%. In contrast, all methods exhibited limited effectiveness on the MNIST dataset due to high geometric complexity and noise in pixel-based point cloud representations. For the ModelNet10 dataset, PI clearly outperformed other techniques, achieving scores of approximately 0.75. Overall, the results indicate that PI provides a robust and versatile topological representation for classification tasks, while PL stands out for its stability and interpretability in complex data analysis.

**Keywords:** topological data analysis, persistent homology, persistence diagram, vectorisation methods, machine learning, classification.

## INTRODUCTION

Topological Data Analysis is a rapidly developing research area that combines concepts from applied mathematics and machine learning. The core idea of TDA is the use of tools from algebraic topology to describe the global structure of data. In contrast to classical feature extraction methods, the topological approach enables the identification of qualitative features, such as the number of connected components, cycles, or higher-dimensional structures [1,2].

Among the key tools of TDA, persistent homology plays a central role, as it is used to track the evolution of topological features as a function of the analysis scale [3,4]. The outcome of such an analysis is a persistence diagram, which provides a graphical representation of the lifespan of individual features, indicating their birth and death times. Although PD constitutes a rich source of information, its direct use in machine learning algorithms is often challenging due to the lack of a fixed number of points and the absence of a natural vector-space representation.

For this reason, a variety of methods for vectorising PDs have been developed in recent years. These methods transform PDs into fixed-length vectors that can be used as inputs to conventional classification models. The most widely used vectorisation techniques include persistence

landscapes [5], Betti curves [6], persistence images [7], and silhouettes [8]. These approaches differ in the way topological information is aggregated, which affects their ability to preserve essential features of persistence diagrams.

The aim of this study is a comparative analysis of selected persistence diagram vectorisation methods in the context of classification tasks. The extent to which different vector representations preserve topological information relevant to class separability is investigated. The analysis is conducted using both synthetic and real-world datasets, allowing the effectiveness of the methods to be assessed under idealised and realistic conditions.

This study focused on the evaluation of four vectorisation techniques in combination with different types of classifiers, including linear models, non-linear models, and neural networks. This approach enables not only a comparison of the vectorisation methods themselves, but also an analysis of their interaction with models of varying expressive power. The obtained results aimed to provide practical guidance on the use of topological representations in classification tasks involving data with complex structure.

This paper is organised as follows. Section 2 reviews the works related to the research topic. Subsequently, Section 3 introduces the theoretical background and issues related to TDA. Section 4 presents the material and methods employed in the study. The results obtained from the comparative analysis of PD vectorisation methods are discussed in Section 5. Finally, Section 6 provides conclusions and a summary of the conducted research, along with a brief outline of planned future work.

## RELATED WORKS

In the literature, Topological Data Analysis is increasingly described as a novel approach to the analysis of data characterised by complex, non-linear structure. The TDA methods enable the investigation of global data properties by interpreting data as topological objects, which allows the identification of features that are not accessible to classical statistical methods. The foundations of this field were formulated in seminal works introducing the concept of data analysis based on representations in the form of topological spaces [9,10].

A key tool of TDA is persistent homology, which enables multiscale data analysis by tracking the evolution of topological features throughout a filtration process [3]. The persistence of individual topological features (such as connected components, cycles, and higher-dimensional structures) is interpreted as a measure of their significance [4]. The results of the PH analysis are represented in the form of persistence diagrams or equivalent barcode representations, which describe the birth and death times of individual topological features [11].

One of the major advantages of topological descriptors is their stability with respect to small perturbations of the data, which makes them particularly well-suited for the ML applications [12]. Numerous studies have demonstrated the effectiveness of PH in a wide range of domains, including image analysis [13,14], time-series signal analysis [15,16], biomedical data [17,18], graph-structured data [19], and high-dimensional datasets [20]. At the same time, it should be emphasised that in most of these works, TDA is not employed as a standalone decision-making algorithm, but rather serves as a feature extraction tool the outputs of which are subsequently provided to the ML algorithms [9].

Despite its many advantages, the direct integration of persistence diagrams with the ML models remains challenging. This difficulty arises primarily since persistence diagrams lack a natural vector-space structure and consist of point sets with variable cardinality [21]. Although direct comparison methods for persistence diagrams do exist, such as bottleneck and Wasserstein distances [22,2], the literature highlights that these approaches are associated with high computational complexity, limited scalability, and poor compatibility with the commonly used ML models [23].

In response to these limitations, a range of methods for transforming persistence diagrams into fixed-dimensional vectors or functions has been proposed. One of the earliest vectorisation approaches is persistence landscapes, in which PDs are represented as families of real-valued functions, thereby enabling the application of classical statistical tools and machine learning methods [5]. These representations have been successfully applied to tasks such as image classification, biological data analysis, and the recognition of geometric structures [24, 25].

Another widely adopted vectorisation technique is persistence images, where PDs are transformed into images of fixed resolution through the discretisation of the diagram space and aggregation of information about the persistence of topological features. Empirical studies

have shown that this approach often outperforms persistence landscapes in terms of classification accuracy, particularly in the analysis of signals and dynamical systems [7]. Persistence images have been applied, among others, to ECG signal analysis, three-dimensional shape classification, and image data analysis [17].

In addition to the aforementioned methods, simpler PD representations, such as Betti curves and persistence silhouettes have also been proposed [6,8]. These approaches are characterised by lower computational complexity and improved interpretability. However, it should be noted that such simplifications often entail a loss of information regarding the precise spatial distribution of topological features within the PD, which may negatively affect classification performance in more complex tasks. Despite the intensive development of PD vectorisation methods, the existing literature still exhibits notable gaps. Most studies focus on single datasets or individual representation methods [7,26]. Relatively few works simultaneously analyse the effect of different vectorisation techniques, different homology dimensions, and different classes of machine learning models within a unified experimental framework [27–29]. Moreover, systematic analyses of result stability with respect to random data splits and of the interactions between data representation and classifier expressive power remain scarce. A particular lack of studies is observed in comparative evaluations of vectorisation methods combined with linear, nonlinear, and deep learning models.

This work bridged the above-mentioned research gap by providing a comprehensive comparative analysis of PD vectorisation methods in classification tasks, encompassing both synthetic and real-world datasets and considering multiple classes of the ML models.

## THEORETICAL BACKGROUND

This chapter presents the theoretical foundations of TDA, which are crucial to understanding the methodology used in this study. It discusses the concepts of Vietoris-Rips filtration, persistent homology, persistence diagrams and selected methods for their vectorisation.

### Vietoris-Rips filtration

The Vietoris-Rips filtration is a particular case of a filtration of topological spaces in which successive sets are simplicial complexes constructed from a given point cloud and a distance parameter $r$ [11].

Formally, let $X = \{x_1, x_2, \ldots, x_N\}$ be a finite set of points in a metric space $(R^d, \|\cdot\|)$. The Vietoris-Rips complex is then defined as the set of simplices for a fixed r > 0 (Equation 1).

$$VR(X,r) = \{\sigma \subseteq X : \mathrm{diam}(\sigma) \leq 2r\} \quad (1)$$

where: diam(σ) denotes the diameter of the set σ (Equation 2).

$$\mathrm{diam}(\sigma) = \max\{\|x - y\| : x, y \in \sigma\} \quad (2)$$

This means that a $k$-simplex belongs to the complex $VR(X, r)$ if and only if all its vertices are within a distance not exceeding $2r$.

By varying the distance parameter $r$, one obtains an increasing sequence of simplicial complexes (Equation 3).

$$VR(X,r_0) \subseteq VR(X,r_1) \subseteq \subseteq VR(X,r_2) \subseteq \cdots \subseteq VR(X,r_m) \quad (3)$$

for $0 < r_0 < r_1 < r_2 < ... < r_m$, which constitutes a filtration of the space.

As $r$ increases, the complexes become progressively denser, allowing the evolution of topological features to be tracked through the analysis of homology groups $H_k$ computed at successive stages of the filtration.

Figure 1 illustrates an example of the Vietoris-Rips filtration process for a non-uniform point cloud, that is represented in a two-dimensional Cartesian coordinate system, where both axes correspond to spatial coordinated expressed in arbitrary units. The initial point cloud (Figure 1a) consists of two circles, several well-separated clusters, and background noise. For a small radius $r$, a limited number of short edges appear. At the same time, most points remain as separate components (Figure 1b). As $r$ increases, points begin to merge into local clusters and edges partially connect to form circular structures (Figure 1c). Further increases in $r$ lead to the formation of one-dimensional cycles (Figure 1d), followed by the appearance of the first 2-simplices (Figure 1e), and ultimately to the formation of numerous triangles and the complete filling of existing cycles (Figure 1f).

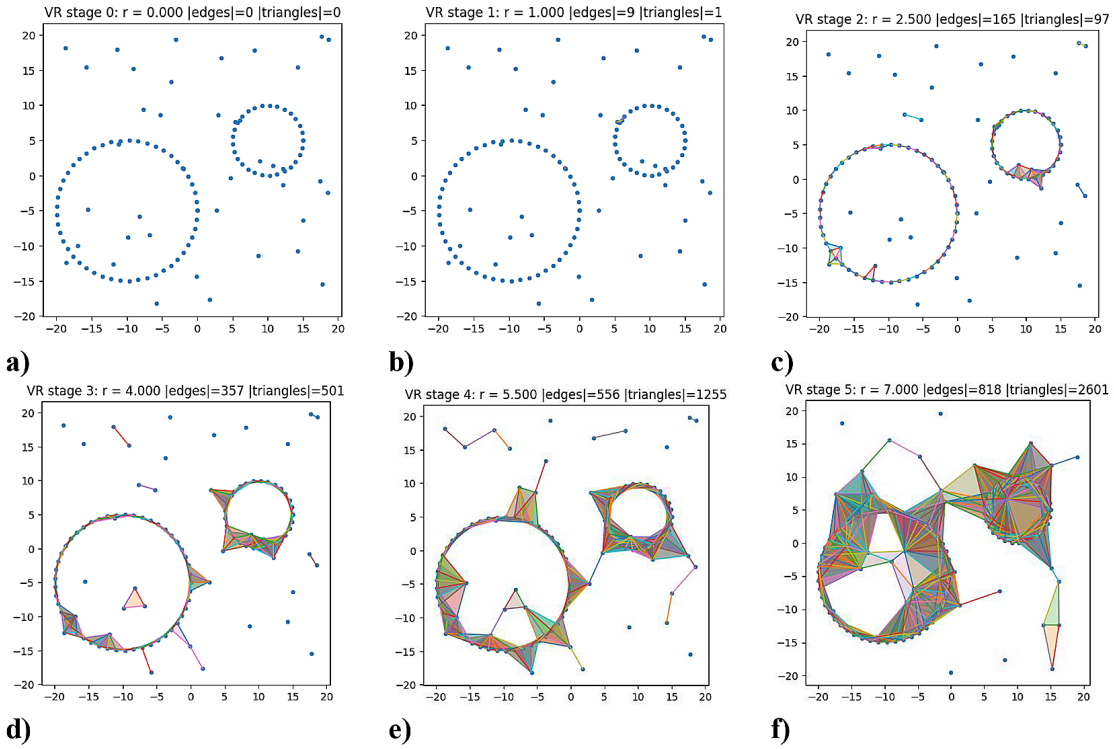



**Figure 1.** Vietoris-Rips filtration: (a) point cloud; (b-f) successive stages of the filtration

## Homology

Homology is used for the quantitative description of the structure of a topological space by identifying its fundamental features, such as the number of connected components, holes, or cavities, based on simplicial complexes [4]. Each algebraic group $H_k$ characterises the number of topological features of dimension $k$, where:

- $H_0$ denotes the number of connected components of the space,
- $H_1$ corresponds to the number of non-contractible loops (holes in $\mathbb{R}^2$, tunnels in $\mathbb{R}^3$),
- $H_2$ represents the number of voids or cavities in three dimensions,
- higher groups $H_k$ describe features of higher dimensions.

Formally, for a simplicial complex, one considers chain spaces $C_k$, which consist of combinations of $k$-simplices. The boundary operator $\partial_k : C_k \to C_{k-1}$ specifies how $k$-simplices are connected to their boundaries. The homology group is defined as the set of cycles that are not boundaries of higher-dimensional simplices (Equation 4).

$$H_k = Z_k / B_k \tag{4}$$

where: $Z_k = \ker(\partial_k)$ is the space of cycles and $B_k = \mathrm{im}(\partial_{k+1})$ is the space of boundaries.

In the case of the Vietoris-Rips filtration, analysing changes in the homology groups $H_k$ for successive values of the distance parameter $r$ makes it possible to track the emergence and disappearance of topological structures in the data.

## Persistence diagrams

For each $k$-dimensional homology class, it is possible to determine its birth time $b$ and death time $d$. This leads to the construction of a $k$-dimensional persistence diagram (Equation 5) [2,4].

$$D^k = \{(b_i, d_i)\}_{i=1}^{M} \subset \mathbb{R}^2, b_i < d_i \tag{5}$$

The distance of a point from the diagonal $\Delta$ $\{(x, x) : x \in \mathbb{R}\}$ is interpreted as the persistence of the corresponding feature (Equation 6).

$$l_i = d_i - b_i \tag{6}$$

Figure 2 presents a point cloud together with the persistence diagram constructed from the Vietoris-Rips filtration. In Figure 2a, the axes represent Cartesian spatial coordinates of points in a two-dimensional Euclidean space expressed in arbitrary units. In Figure 2b, the horizontal axis represents the birth time, and the vertical axis represents the death time of topological features, both expressed in the filtration parameter units

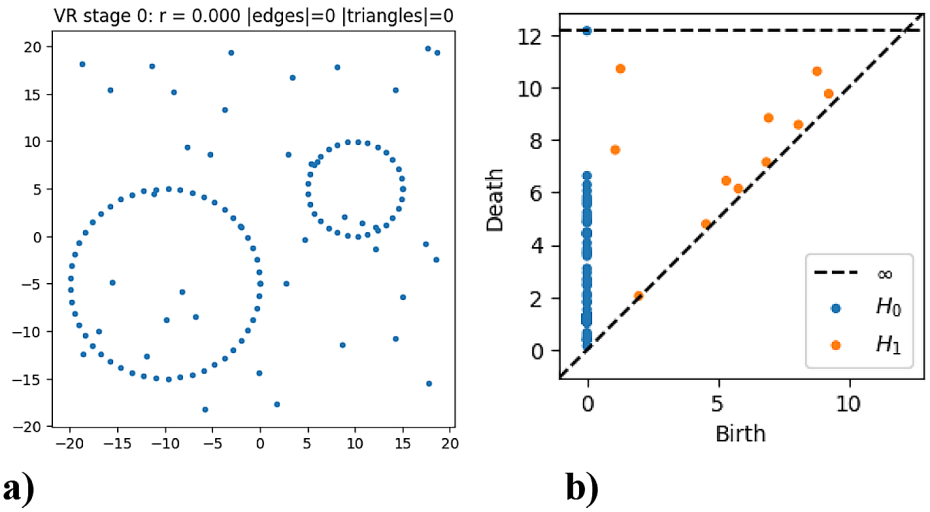



**Figure 2.** Comparison of a point cloud (a) and its corresponding persistence diagram (b)

corresponding to the distance threshold used in the Vietoris-Rips filtration.

## Vectorisation of persistence diagrams

Persistence diagrams provide rich information about the topological structure of data and the lifetimes of the features present in this structure. However, their direct use in ML classifiers is difficult or even impossible due to the variable number of points in PDs. This issue is addressed by applying various vectorisation methods.

Vectorisation of PDs consists of transforming them into fixed-length vectors that preserve the key information about the persistence and spatial distribution of topological features. Several such methods have been proposed in the literature. In this work, four approaches were analysed: Persistence Landscape, Betti Curves, Persistence Image, and Persistence Silhouette.

### *Persistence landscape*

PL is one of the most used techniques for PD vectorisation. The method transforms each homological interval into a one-dimensional function. For each point $(b_i, d_i)$ in a PD and for each time $t$, the function is defined in Equation 7.

$$\lambda_{(b,d)}(t) = \max(0, \min(t - b, d - t) \quad (7)$$

This function attains its maximum at the midpoint of the interval and decreases linearly to zero towards the birth and death times, yielding a tent (hat) function shape. The functions $\lambda_{(b_i, d_i)}(t)$ obtained for all points in the persistence diagram are superimposed and then sorted in descending order at each time $t$. This results in a family of persistence landscape layers (Equation 8).

$$\Lambda_k(t) = \text{the } k\text{-th largest value in } \{\lambda_{(b_i,d_i)}(t)\} \quad (8)$$

The first layer $\Lambda_1(t)$ corresponds to the most persistent topological features, while subsequent layers represent progressively less significant ones. After discretising the functions on a fixed grid, a feature vector of length $K \cdot N$ is obtained, where $K$ denotes the number of layers and $N$ the number of grid samples [5]. An exemplary PL for the data in Figure 2 is presented in Figure 3a.

### *Betti curves*

BCs constitute one of the simplest approaches to PD vectorisation. The method consists in computing the number of active topological features for a given filtration parameter $\varepsilon$. Active features are defined as all points $(b_i, d_i)$ where $b_i \leq \varepsilon \leq d_i$.

The result is a function $\beta(\varepsilon)$ that describes changes in the number of topological features throughout the filtration. After discretising the range of $\varepsilon$ into a fixed number of intervals $N$, the function $\beta(\varepsilon)$ is a vector of length $N$ [6]. Exemplary BCs for the data shown in Figure 2 are shown in Figure 3b.

### *Persistence image*

The first step in PI-based vectorisation is the transformation of points $(b_i, d_i)$ into $(b_i, pi)$ where $p_i$ denotes the persistence of the feature, defined as $p_i = d_i - b_i$. Each point is then represented by a Gaussian function (Equation 9).

$$\rho_i(x,y) = w(p_i)\ \exp\left(-\frac{(x-b_i)^2+(y-p_i)^2}{2\sigma^2}\right) \quad (9)$$

where: $wp_i(p_i)$ is a weighting function that increases the influence of more persistent topological features, and σ controls the degree of smoothing.

By summing all functions $\rho_i$ and discretising the resulting image on a regular grid, a matrix of values is obtained, which can be flattened into a feature vector [7]. An exemplary PI for the data in Figure 2 is shown in Figure 3c.

### *Persistence silhouette*

PS can be viewed as an averaged form of the persistence landscape. The method is defined as a weighted average of the functions $\lambda_k(t)$ (Equation 10).

$$\phi(t) = \frac{\sum_k w_k \lambda_k(t)}{\sum_k w_k} \quad (10)$$

where: the weights $w_k$ may depend on the persistence of the corresponding features.

The resulting function $\phi(t)$ is one-dimensional and can be sampled at a fixed number of points, yielding a low-dimensional feature vector [8]. An exemplary PS for the data in Figure 2 is presented in Figure 3d.

## MATERIALS AND METHODS

This section describes the datasets used in the experimental study and the techniques used to transform them into point clouds. The classification models are also presented, along with a description of the evaluation methods and metrics used to assess the quality of various PDs vectorisation techniques.

### Datasets

The study employed a collection of datasets comprising both synthetic and real-world data, representing different types of topological structures and measurement modalities. Depending on the characteristics of a given dataset, the input data were transformed into point clouds using different strategies. The topological analysis involved the computation of homology in dimensions $H_0$, $H_1$, and $H_2$; the choice of homology dimensions was dictated by the nature of the obtained representation rather than by intrinsic properties of raw data.

### *Synthetic dataset in* $\mathbb{R}^2$

The two-dimensional synthetic dataset consists of 700 samples uniformly distributed across

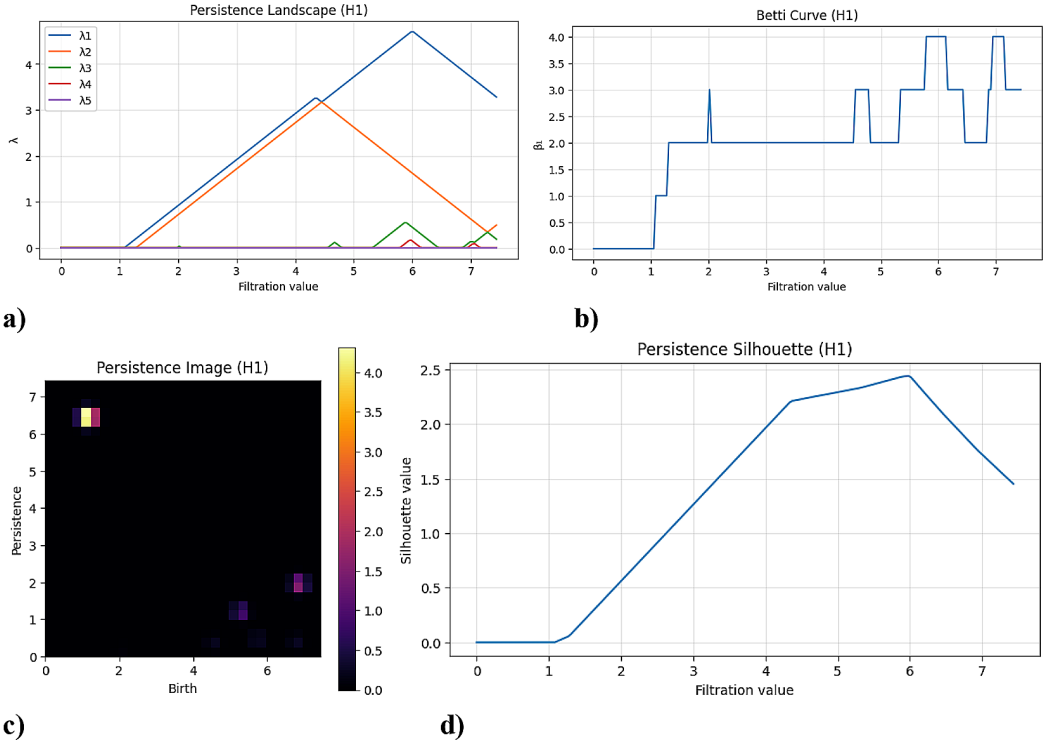



**Figure 3.** Visualisation of different PD (Figure 2) vectorisation methods: (a) PL, (b) BC, (c) PI, (d) PS

seven classes, each representing a distinct point cloud configuration with a known geometric and topological structure. The analysis includes circular structures, figure-eight shapes, Gaussian clusters, spirals, square contours, parallel line arrangements, and two-moons configurations. The selected structures were designed to ensure diverse topological characteristics, ranging from point clouds without prominent one-dimensional cycles (Gaussian clusters, spirals, parallel lines) to configurations exhibiting dominant $H_1$ homological features (circles, figure-eight shapes, squares).

Each point cloud was generated using different parametric models or stochastic sampling schemes. For instance, circular structures were constructed based on trigonometric parametrisations, while spiral configurations were obtained through radial parametrisation. Individual point clouds consisted of 100–200 points, depending on the geometric complexity of the structure. To avoid trivial topological cases, all points sets were perturbed using low-level Gaussian noise with controlled intensity.

The synthetic dataset was constructed to evaluate the sensitivity of the considered vectorisation methods to variations in shape, connectivity patterns, and the presence or absence of one-dimensional topological features.

### *Synthetic dataset in* $\mathbb{R}^3$

To evaluate the effectiveness of vectorisation methods for PD in the context of $H_2$ homology, an analysis was conducted on synthetic data represented as point cloud embedded in three-dimensional Euclidean space. The dataset consists of 600 samples, evenly distributed across six classes representing spheres, toruses, helices, cubes parallel planes, and Gaussian blobs. The selected structures provide diverse topological properties, ranging from simple arrangement without prominent cycles (blobs, cubes, planes) to configurations exhibiting significant $H_1$ and $H_2$ features (spheres, toruses, helices).

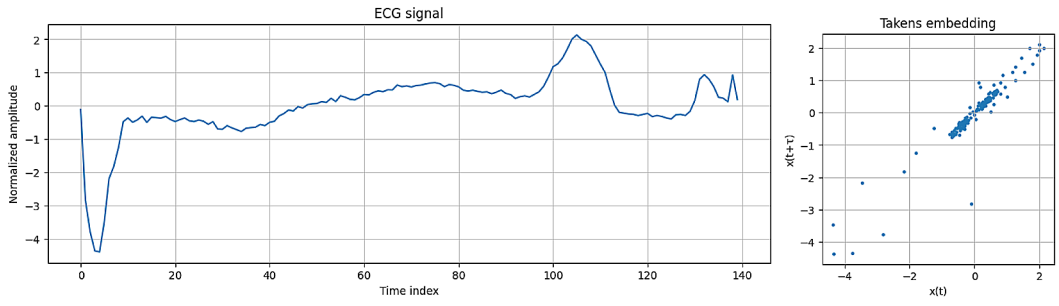
Individual point clouds were generated using parametric models or stochastic sampling schemes; for example, a torus was created using a double-angular parametrisation. Each point cloud contained between 400 and 1000 point, depending on the complexity of the structure. Similar to the synthetic dataset in $\mathbb{R}^2$, all point clouds were perturbed with low-intensity Gaussian noise with controlled magnitude.

### *ECG5000 dataset*

The ECG5000 dataset from the UCR/UEA repository consists of one-dimensional electrocardiographic signals [31]. After normalising each sample using z-score standardisation, the data were transformed into phase space using time-delay embedding in accordance with Takens' method [32]. An embedding dimension of 3 and a delay $\tau = 1$ enabled the reconstruction of signal dynamics as trajectories in three-dimensional space. It should be emphasised that this space is auxiliary in nature and does not correspond to the physical dimensionality of the original data. The resulting point clouds exhibited cycling patterns arising from the periodic nature of the ECG signals. An example transformation of a sample from the ECG dataset is shown in Figure 4.

### *MNIST dataset*

The MNIST dataset from the scikit-datasets environment comprises 70,000 images of handwritten digits, each stored as a 28 × 28 matrix with a single intensity channel [33]. To prepare the data for topological analysis, the images were transformed into point clouds by extracting active pixels, i.e., pixels with non-zero intensity, and




**a)** **b)**

**Figure 4.** Example ECG5000 sample: (a) original signal, (b) representation as a point cloud

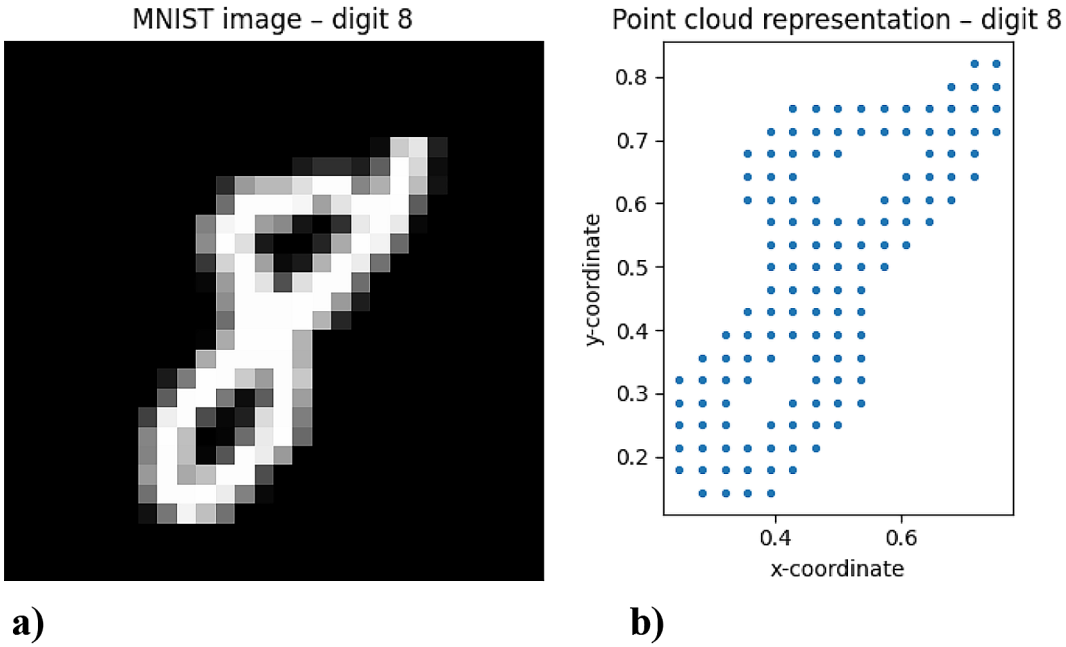



**Figure 5.** Example MNIST image: (a) original image, (b) representation as a point cloud

mapping their coordinates into a two-dimensional space. The resulting point-based representations reflect the geometric structure of the digit, enabling the analysis of $H_1$ homology, in particular, the presence of loops and closed shapes in individual digits. An exemplary image transformation is presented in Figure 5.

### *ModelNet10 dataset*

The ModelNet10 dataset is a widely used benchmark for three-dimensional shape recognition tasks and consists of models of everyday objects represented as triangular meshes and grouped into semantic classes [34]. In this study, a subset containing 100 samples from each of five classes (bathtub, bed, chair, desk, and dresser) was used. Each model was converted into a point cloud with a fixed number of points and subsequently normalised by translating the centre of mass to the origin and scaling the cloud to fit within the unit sphere. As a result of this normalisation, the coordinate values are dimensionless and expressed in a standardised Cartesian coordinate system (Figure 6). This procedure reduced the influence of differences in object scale and position on the topological analysis. The use of the ModelNet10 dataset enabled the evaluation

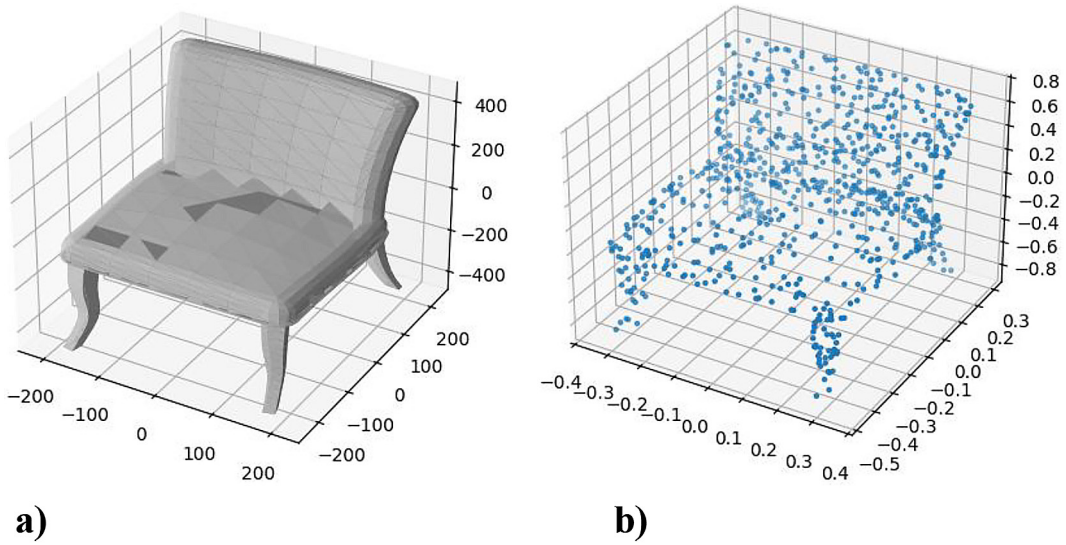



**Figure 6.** Example ModelNet10 sample: (a) original data, (b) representation as a point cloud

of persistence diagram vectorisation methods under realistic conditions characterised by complex and irregular geometric structures.

### Selection of classifiers

To assess the effectiveness of individual PD vectorisation methods, three distinct classes of classification models were employed, responding to different machine learning paradigms. Each model was trained independently for each dataset and each feature vectorisation technique.

The first group consisted of a linear classifier implemented using Logistic Regression [35], which served as a baseline for evaluating the degree of non-linearity captured by the representations produced by different vectorisation methods.

The non-linear model was based on gradient boosting (XGBoost) [36], a tree-based classifier employing iterative boosting. This model is particularly well-suited to the data with a complex structure and enables the capture of non-linear relationships between topological features.

The final group comprised a deep learning model, namely a Multilayer Perceptron [38], which allowed an assessment of the extent to which vectorisation methods provide representations suitable for models with high expressive power.

Hyperparameter tuning was performed separately for each PD vectorisation method (PL, BC PI, PS) using ten-fold cross-validation implemented via GridSearchCV procedure [39]. The dataset was randomly partitioned into ten approximately equal subset. In each iteration, nine subsets were used for training and one subset for validation, corresponding to an effective 90%/10% split. The partitioning procedure was stratified to preserve class proportions across folds. This process was repeated ten times so that each subset served as the validation set exactly once. The optimal hyperparameter configuration was selected by maximising the mean validation accuracy.

Parameter grids were tailored to each algorithm. All models were trained using the default optimisation mechanism provided by the scikit-learn and XGB libraries.

### Model evaluation metrics

Model performance was assessed using the metrics derived from the confusion matrix. For $c$-class classification problem, the confusion matrix has size $c \times c$, where rows correspond to true classes and columns to predicted classes. The diagonal elements $n_{ii}, i = 1, 2, ..., c$ represent the number of correctly classified samples belonging to class $i$, while off-diagonal elements correspond to misclassifications. The sum of all entries equals the total number of analysed samples $N$.

The confusion matrix is interpreted using the following quantities for class $i$:

- True positive (TP) – the diagonal element in row $i$ and column $i$,
- False positive (FP) – the sum of elements in column $i$, excluding the diagonal element,
- False negative (FN) – the sum of elements in row $i$, excluding the diagonal element,
- True negative (TN) – the sum of all remaining elements in the matrix.

On the basis of these quantities, the following evaluation metrics are defined for class $k$:

- Accuracy – proportion of correctly classified samples among all samples (Equation 11),

$$\text{Accuracy}_k = \frac{TP_k + TN_k}{TP_k + FP_k + FN_k + TN_k} \tag{11}$$

- Precision – proportion of correctly classified samples of class $k$ among all samples predicted as class $k$ (Equation 12),

$$\text{Precision}_k = \frac{TP_k}{TP_k + FP_k} \tag{12}$$

- Recall – proportion of samples of class $k$ correctly identified by the model (Equation 13),

$$\text{Recall}_k = \frac{TP_k}{TP_k + FN_k} \tag{13}$$

- F1-score – harmonic mean of precision and recall (Equation 14).

$$\text{F1}_k = 2\left(\frac{1}{\text{Precision}_k} + \frac{1}{\text{Recall}_k}\right)^{-1} \tag{14}$$

### Experimental pipeline architecture

Figure 7 presents the complete data processing pipeline employed in this study. The designed workflow is universal in nature and can be applied to both synthetic and real-world data, regardless of their dimensionality.

## RESULTS AND DISCUSSION

This section presents and discusses the results obtained from the experiments conducted. The performance of individual PD vectorisation

methods across different classifiers was compared using metrics such as accuracy, precision, recall and F1-score.

## Results for two-dimensional synthetic data

Table 1 presents the classification performance results for a synthetic two-dimensional dataset comprising various geometric and topological structures. Among the analysed vectorisation methods, PI achieved the highest performance, with all evaluation metrics attaining values close to 0.98, regardless of the used classifier (LR, XGB, MLP). For the LR and XGB classifiers, PL and PS yielded comparable results, with metric values of approximately 0.965. In the case of the MLP classifier, PL achieved results only slightly lower than PI (around 0.975), whereas PS reached values of approximately 0.945. Across all classifiers, BC performed the weakest, with a classification quality of around 0.93.

It is worth noting that PL exhibited the highest stability, with standard deviations approximately two times lower than those of the other methods. These results indicate that for data containing pronounced one-dimensional cycles, PI and PL constitute the most informative representations of PDs.

## Results for three-dimensional synthetic data

The classification results for three-dimensional synthetic data containing structures with non-zero $H_2$ homology are presented in Table 2. The highest performance was observed for the PI method, which achieved metric values of approximately 0.99 independently of the classifier applied. Slightly lower, yet still very high, results were obtained for PL (around 0.985). The BC method yielded intermediate performance (approximately 0.96), while PS performed the weakest, with metric values ranging between 0.93 and 0.94.

All classification models achieved very high accuracy, indicating strong class separability in the space of topological features for the

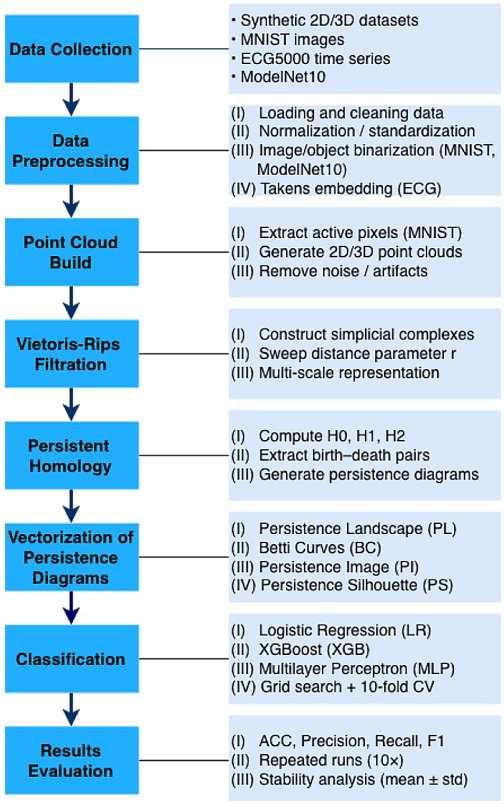



**Figure 7.** Flowchart of the experimental pipeline

**Table 1.** Classification results for 2D synthetic data

| Classifier | Method | Accuracy | | Precision | | Recall | | F1 | |
|---|---|---|---|---|---|---|---|---|---|
| | | mean | ± std | mean | ± std | mean | ± std | mean | ± std |
| LR | PI | **0.9814** | **0.0120** | **0.9821** | **0.0117** | **0.9814** | **0.0120** | **0.9813** | **0.0120** |
| | BC | 0.9329 | 0.0156 | 0.9351 | 0.0159 | 0.9329 | 0.0156 | 0.9321 | 0.0155 |
| | PL | 0.9686 | 0.0064 | 0.9700 | 0.0061 | 0.9686 | 0.0064 | 0.9686 | 0.0065 |
| | PS | 0.9643 | 0.0124 | 0.9669 | 0.0115 | 0.9643 | 0.0124 | 0.9642 | 0.0118 |
| XGB | PI | **0.9757** | **0.0130** | **0.9774** | **0.0119** | **0.9757** | **0.0130** | **0.9757** | **0.0130** |
| | BC | 0.9371 | 0.0106 | 0.9405 | 0.0091 | 0.9371 | 0.0106 | 0.9372 | 0.0113 |
| | PL | 0.9657 | 0.0060 | 0.9698 | 0.0053 | 0.9657 | 0.0060 | 0.9656 | 0.0062 |
| | PS | 0.9529 | 0.0211 | 0.9548 | 0.0173 | 0.9529 | 0.0211 | 0.9525 | 0.0212 |
| MLP | PI | **0.9824** | **0.0134** | **0.9835** | **0.0127** | **0.9826** | **0.0134** | **0.9315** | **0.0134** |
| | BC | 0.9271 | 0.0178 | 0.9290 | 0.0181 | 0.9271 | 0.0178 | 0.9265 | 0.0188 |
| | PL | 0.9786 | 0.0081 | 0.9812 | 0.0072 | 0.9786 | 0.0081 | 0.9788 | 0.0080 |
| | PS | 0.9443 | 0.0113 | 0.9470 | 0.0096 | 0.9443 | 0.0113 | 0.9441 | 0.0115 |

**Table 2.** Classification results for 3D synthetic data

| Classifier | Method | Accuracy | | Precision | | Recall | | F1 | |
|---|---|---|---|---|---|---|---|---|---|
| | | mean | ± std | mean | ± std | mean | ± std | mean | ± std |
| LR | PI | **0.9899** | **0.0187** | **0.9889** | **0.0187** | **0.9889** | **0.0186** | **0.9885** | **0.0185** |
| | BC | 0.9658 | 0.0093 | 0.9663 | 0.0083 | 0.9658 | 0.0093 | 0.9658 | 0.0094 |
| | PL | 0.9833 | 0.0272 | 0.9859 | 0.0226 | 0.9833 | 0.0272 | 0.9832 | 0.0275 |
| | PS | 0.9333 | 0.0401 | 0.9391 | 0.0375 | 0.9333 | 0.0401 | 0.9331 | 0.0399 |
| XGB | PI | **0.9875** | **0.0186** | **0.9889** | **0.0166** | **0.9875** | **0.0186** | **0.9871** | **0.0193** |
| | BC | 0.9650 | 0.0174 | 0.9685 | 0.0147 | 0.9650 | 0.0174 | 0.9648 | 0.0176 |
| | PL | 0.9825 | 0.0401 | 0.9802 | 0.0298 | 0.9825 | 0.0401 | 0.9817 | 0.0413 |
| | PS | 0.9408 | 0.0432 | 0.9468 | 0.0324 | 0.9408 | 0.0432 | 0.9402 | 0.0445 |
| MLP | PI | **0.9875** | **0.0115** | **0.9889** | **0.0101** | **0.9875** | **0.0115** | **0.9881** | **0.0112** |
| | BC | 0.9617 | 0.0114 | 0.9626 | 0.0101 | 0.9617 | 0.0114 | 0.9616 | 0.0115 |
| | PL | 0.9872 | 0.0114 | 0.9889 | 0.0101 | 0.9872 | 0.0114 | 0.9875 | 0.0115 |
| | PS | 0.9417 | 0.0373 | 0.9448 | 0.0364 | 0.9417 | 0.0373 | 0.9418 | 0.0371 |

three-dimensional data. Once again, PI and PL proved to be the most informative representations, whereas PS exhibited the lowest performance, which may suggest a limited ability of this technique to capture more complex topological structures.

## Results for the ECG5000 dataset

Table 3 presents the classification results for the ECG signals after their transformation into phase space using Takens' embedding. In contrast to the synthetic datasets, a clear differentiation in performance between classification models can be observed. The LR classifier achieved metric values not exceeding 0.85, whereas both non-linear models (XGB, MLP) obtained results above 0.95. For all considered classifiers, PI and PL exhibited the highest and very similar performance, with differences between them not exceeding 0.005. For LR, PS yielded intermediate results (around 0.7), while the lowest classification quality was observed for BC (approximately 0.6). For the XGB and MLP models, BC and PS achieved comparable results at the level of approximately 0.95–0.96. These findings suggest that phase-space reconstruction generates clear and stable cyclic structures, which are most effectively captured by PI and PL, irrespective of the classification model employed.

## Results for the MNIST dataset

Table 4 presents the classification results obtained for the MNIST dataset after transforming

**Table 3.** Classification results for the ECG5000 dataset

| Classifier | Method | Accuracy | | Precision | | Recall | | F1 | |
|---|---|---|---|---|---|---|---|---|---|
| | | mean | ± std | mean | ± std | mean | ± std | mean | ± std |
| LR | PI | 0.8359 | 0.0088 | 0.8353 | 0.0090 | 0.8359 | 0.0088 | 0.8340 | 0.0088 |
| | BC | 0.6197 | 0.0081 | 0.6169 | 0.0092 | 0.6196 | 0.0081 | 0.6173 | 0.0089 |
| | PL | **0.8417** | **0.0090** | **0.8399** | **0.0091** | **0.8417** | **0.0090** | **0.8397** | **0.0091** |
| | PS | 0.7150 | 0.0072 | 0.7065 | 0.0077 | 0.7150 | 0.0072 | 0.7085 | 0.0078 |
| XGB | PI | **0.9770** | **0.0035** | **0.9770** | **0.0035** | **0.9770** | **0.0035** | **0.9769** | **0.0035** |
| | BC | 0.9602 | 0.0042 | 0.9603 | 0.0040 | 0.9602 | 0.0042 | 0.9602 | 0.0041 |
| | PL | 0.9741 | 0.0024 | 0.9740 | 0.0024 | 0.9741 | 0.0024 | 0.9740 | 0.0024 |
| | PS | 0.9672 | 0.0024 | 0.9671 | 0.0024 | 0.9672 | 0.0024 | 0.9670 | 0.0024 |
| MLP | PI | **0.9765** | **0.0038** | **0.9765** | **0.0038** | **0.9765** | **0.0038** | **0.9763** | **0.0039** |
| | BC | 0.9629 | 0.0045 | 0.9628 | 0.0046 | 0.9629 | 0.0045 | 0.9625 | 0.0046 |
| | PL | 0.9710 | 0.0036 | 0.9710 | 0.0035 | 0.9710 | 0.0036 | 0.9707 | 0.0036 |
| | PS | 0.9516 | 0.0049 | 0.9520 | 0.0048 | 0.9515 | 0.0049 | 0.9510 | 0.0051 |

images into point cloud representations. Compared to the other datasets, MNIST proves to be considerably more challenging for the TDA-based analysis. All analysed models achieved metric values in the range of 0.35–0.49.

For the LR classifier, the best results were achieved using PI (approximately 0.47), while PL yielded intermediate values of around 0.43. The lowest performance was recorded for BC and PS, with classification quality ranging between 0.35 and 0.38. For the XGB classifier, PI and PL again achieved the highest effectiveness (around 0.48), whereas BC and PS reached noticeably lower values of approximately 0.42. The weakest results were observed for the MLP classifier, where PI achieved around 0.42, PL around 0.41, PS around 0.39, and BC around 0.35. Similarly to the other datasets, PI and PL obtained the highest performance, while BC consistently produced the weakest results.

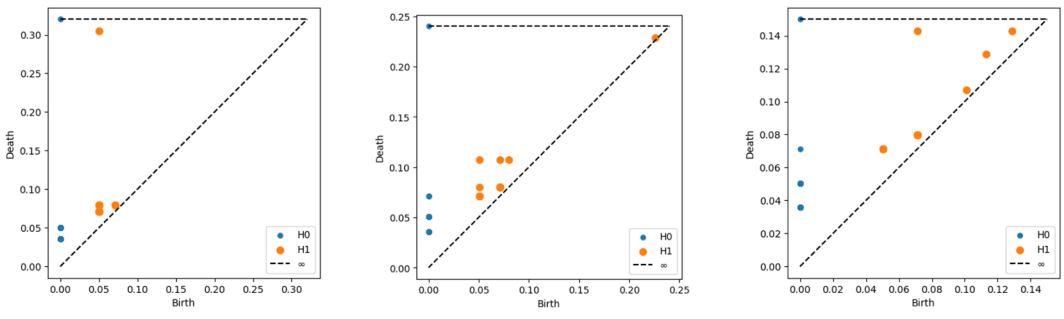
These results can be primarily attributed to the limitations inherent in the adopted representation strategy. The conversion of grayscale images into point clouds based solely on active pixels fundamentally alters the structure of the data by eliminating intensity information as well as local spatial dependencies between neighbouring pixels. Since such properties play a critical role in image recognition tasks, the resulting PDs may not preserve sufficiently informative characteristics required for effective class discrimination.

From a topological standpoint, PDs derived from pixel-based point clouds tent to exhibit many low-persistence features concentrated near the diagonal (Figure 8). These features are frequently associated with discretisation artefacts, minor geometric perturbations, or sampling noise, rather than meaningful structural patterns. Consequently, the diagrams become dominated by short-lived topological features, which reduces their discriminative capacity and increases representation variability.

Additionally, handwritten digits do not form clean and stable topological structures comparable to analytically generated geometric shapes.

**Table 4.** Classification results for the MNIST dataset

| Classifier | Method | Accuracy | | Precision | | Recall | | F1 | |
|---|---|---|---|---|---|---|---|---|---|
| | | mean | ± std | mean | ± std | mean | ± std | mean | ± std |
| LR | PI | **0.4685** | **0.0228** | **0.4667** | **0.0236** | **0.4685** | **0.0228** | **0.4623** | **0.0215** |
| | BC | 0.3580 | 0.0247 | 0.3401 | 0.0290 | 0.3580 | 0.0247 | 0.3443 | 0.0274 |
| | PL | 0.4340 | 0.0197 | 0.4333 | 0.0222 | 0.4340 | 0.0197 | 0.4314 | 0.0206 |
| | PS | 0.3850 | 0.0116 | 0.3912 | 0.0089 | 0.3850 | 0.0116 | 0.3769 | 0.0077 |
| XGB | PI | 0.4775 | 0.0307 | 0.4790 | 0.0370 | 0.4775 | 0.0307 | 0.4673 | 0.0311 |
| | BC | 0.4265 | 0.0201 | 0.4163 | 0.0213 | 0.4265 | 0.0201 | 0.4176 | 0.0202 |
| | PL | **0.4800** | **0.0252** | **0.4838** | **0.0288** | **0.4800** | **0.0252** | **0.4773** | **0.0249** |
| | PS | 0.4185 | 0.0178 | 0.4250 | 0.0147 | 0.4185 | 0.0178 | 0.4134 | 0.0141 |
| MLP | PI | **0.4205** | **0.0351** | **0.4187** | **0.0305** | **0.4205** | **0.0351** | **0.4170** | **0.0328** |
| | BC | 0.3515 | 0.0171 | 0.3418 | 0.0200 | 0.3515 | 0.0171 | 0.3443 | 0.0186 |
| | PL | 0.4090 | 0.0193 | 0.4094 | 0.0119 | 0.4090 | 0.0193 | 0.4068 | 0.0151 |
| | PS | 0.3920 | 0.0207 | 0.3915 | 0.0178 | 0.3920 | 0.0207 | 0.3861 | 0.0187 |




**a)** **b)** **c)**

**Figure 8.** Example persistent diagrams obtained for: (a) digit 0, (b) digit 6, (c) digit 9

Apparent loop-like patterns observed in certain digits are often fragmented, irregular, and highly sensitive to the variations in handwriting style, thresholding, and sampling procedures. Small geometric differences may therefore induce disproportionately large changes in PDs, further limiting classification performance.

A similar limitation of purely topological representations in image classification problems has been reported in previous studies, where persistence-based features often require integration with geometric or learned representations to achieve competitive performance [20,21,23].

Alternative representation strategies may therefore offer improved suitability for topological analysis. The approaches based on cubical complexes or intensity-based filtrations enable persistence diagrams to be computed directly from image data while preserving richer structural information. Similarly, hybrid frameworks integrating deep learning models with topological descriptors have demonstrated promising potential in capturing complementary geometric and topological features.

## Results for the ModelNet10 dataset

Table 5 presents the classification results obtained for the ModelNet10 dataset. Regardless of the classifier used, the highest performance was achieved by the PI method, with metric values ranging from 0.7 to 0.8. The BC and PL vectorisation techniques yielded nearly identical results across all analysed cases, with evaluation metrics oscillating between 0.6 and 0.7. For the PS method, markedly lower results were observed, not exceeding 0.6. In contrast to the previously analysed datasets, a single vectorisation method clearly dominates the other in this case, namely PI.

## Statistical significance analysis

Given the extensive experimental framework, statistical significance analysis was restricted to the ECG5000 dataset to avoid excessive multiple comparisons while ensuring representative statistical validation. Statistical significance was assessed using post-hoc Wilcoxon signed-rank tests for paired samples. The analysis was based on the results obtained from ten-fold cross-validation. The tests were conducted using the XGB classifier, selected as representative non-linear model exhibiting high and stable classification performance. This choice allows differences between vectorisation methods to be evaluated while minimising model-specific effects. The ECG5000 dataset was selected as it represents a realistic classification scenario characterised by non-trivial topological structure and low cross-validation variability, making it particularly suitable for statistical analysis.

The results of the Wilcoxon tests (Table 6) indicate that PI representation achieves statistically significant improvements compared to the remaining vectorisation methods ($p < 0.025$). No statistically significant differences were observed between the PL and PS representations ($p \approx 0.8$). In contrast, the BC method yielded significantly lower classification performance ($p < 0.045$).

**Table 5.** Classification results for the ModelNet10 dataset

| Classifier | Method | Accuracy | | Precision | | Recall | | F1 | |
|---|---|---|---|---|---|---|---|---|---|
| | | mean | ± std | mean | ± std | mean | ± std | mean | ± std |
| LR | PI | **0.7000** | **0.0000** | **0.6810** | **0.0000** | **0.7000** | **0.0000** | **0.6723** | **0.0000** |
| | BC | 0.6600 | 0.0000 | 0.6676 | 0.0000 | 0.6600 | 0.0000 | 0.6442 | 0.0000 |
| | PL | 0.6600 | 0.0000 | 0.6676 | 0.0000 | 0.6600 | 0.0000 | 0.6442 | 0.0000 |
| | PS | 0.5000 | 0.0000 | 0.5133 | 0.0000 | 0.5000 | 0.0000 | 0.5053 | 0.0000 |
| XGB | PI | **0.7400** | **0.0000** | **0.7983** | **0.0000** | **0.7200** | **0.0000** | **0.7060** | **0.0000** |
| | BC | 0.7000 | 0.0000 | 0.7533 | 0.0000 | 0.7000 | 0.0000 | 0.7022 | 0.0000 |
| | PL | 0.7000 | 0.0000 | 0.7533 | 0.0000 | 0.7000 | 0.0000 | 0.7022 | 0.0000 |
| | PS | 0.5800 | 0.0000 | 0.5922 | 0.0000 | 0.5800 | 0.0000 | 0.5553 | 0.0000 |
| MLP | PI | **0.7000** | **0.0267** | **0.7455** | **0.0527** | **0.7000** | **0.0267** | **0.6801** | **0.0335** |
| | BC | 0.6280 | 0.0413 | 0.6114 | 0.0743 | 0.6280 | 0.0413 | 0.5994 | 0.0469 |
| | PL | 0.6360 | 0.0430 | 0.6166 | 0.0659 | 0.6360 | 0.0430 | 0.6139 | 0.0546 |
| | PS | 0.5440 | 0.0295 | 0.5400 | 0.0362 | 0.5440 | 0.0295 | 0.5289 | 0.0333 |

**Table 6.** Pairwise Wilcoxon signed-rank test results for persistence diagram vectorisation methods (ECG5000)

| Parameter | PI | BC | PL | PS |
|---|---|---|---|---|
| PI | - | 0.001953 | 0.023438 | 0.023413 |
| BC | 0.001953 | - | 0.044688 | 0.031250 |
| PL | 0.023438 | 0.044688 | - | 0.804688 |
| PS | 0.023413 | 0.031250 | 0.804688 | - |

The obtained results remain consistent with the trends reported in previous studies on the PD vectorisation methods. Representations such as PI and PL have frequently been identified as among the most effective and stable approaches across various classification tasks. Similar observations regarding their robustness and discriminative capabilities have been reported in earlier works [7,21,23]. While direct quantitative comparisons are limited due to differences in datasets, feature extraction pipelines, and classifier configurations, the overall performance patterns observed in this study align with the existing body of research on topological representations in ML.

## Computational complexity considerations

In addition to classification performance, the practical of PD vectorisation methods depends on their computational and memory requirements. The overall processing pipeline consists of two stages: PH computation and PD vectorisation. Since PH was computed using the same procedure for all experiments, its computational cost does not affect the relative comparison of vectorisation methods and is therefore not analysed separately.

Let $m$ denote the number of points in a PD and $G$ the discretisation grid size (or R × R) resolution for PI). Among the analysed approaches, BC and PS exhibit the lowest computational complexity, typically scaling linearly with the number of diagram points and grid resolution (approximately $O(mG)$, with memory requirements proportional to $O(G)$. PL introduces additional sorting operations required to extract landscape layers, leading to complexity on the order of $O(G \cdot m \log m)$ and memory usage proportional to the number of retained layers. PI are computationally more demanding due to two-dimensional grid discretisation, with complexity approximately $O(mR^2)$ and memory requirements of $O(R^2)$.

## Practical applications and limitations

The obtained results indicate that PD vectorisation methods may have practical applicability in classification tasks, particularly when the considered classes differ clearly in a topological sense. This is confirmed by the results obtained for synthetic datasets, where accuracy for PI and PL exceeded 0.98. Such performance suggests potential usefulness in the analysis of two- and three-dimensional shapes, where global structural properties play a significant role, for example in spatial object classification, biological data analysis, or the study of porous structures.

The results obtained for the ECG5000 dataset (accuracy < 0.95 for PI combined with nonlinear models) demonstrate that topological representations are especially suitable for time-series analysis after phase-space reconstruction. In practical terms, this may correspond to the applications in anomaly detection in time series, biomedical signal analysis, or the monitoring of dynamic systems, where cyclic patterns constitute diagnostically relevant features.

In contrast, the moderate results obtained for the ModelNet10 dataset (accuracy ≈ 0.70–0.75) suggests that topological information represents an important, yet insufficient, source of features in three-dimensional object classification tasks. In practical applications, this implies that TDA may serve as a complementary component to classical geometric descriptors, rather than a standalone representation.

The most significant limitations of the TDA-based approach are observed in the case of the MNIST dataset (accuracy ≈ 0.35–0.48). Transforming images into point cloud representations leads to the loss of information regarding pixel dependencies and local spatial relationships, which results in PDs dominated by many short-lived features. In practical terms, this indicated that purely topological representations are insufficient in the tasks where local structural dependencies, rather than global topology, play decisive role.

An additional limitation concerns the computational cost of PH for large and dense point clouds, which may hinder applications in real-time or large-scale systems. Furthermore, vectorisation methods require the selection of several hyperparameters, which may increase the complexity of system design and optimisation.

## CONCLUSIONS

The objective of this study was achieved. The experimental results confirm that the use of vectorisation methods for PDs enables topological information to be effectively incorporated as input to classical machine learning models. It was demonstrated that the choice of vectorisation method has a significant impact not only on classification performance, but also on model stability and robustness with respect to random data splits.

The analysis shows that, regardless of the classifier employed or the type of data considered, the highest performance is consistently observed for PI and PL. BC and PS, as simpler representations of PDs, exhibit noticeably lower effectiveness. The main conclusions drawn from the conducted experiments can be summarised as follows.

1. Classifiers achieved very high metric values (above 0.9) for synthetic datasets, indicating strong class separability in the space of topological features. The best results, exceeding 0.98, were obtained for PI and PL.
2. For real-world data, such as the ECG5000 dataset, PI and PL consistently achieved the highest performance irrespective of the classifier used. The non-linear models (XGB, MLP) outperformed the linear LR classifier by up to 30%, highlighting the importance of non-linear relationships in the topological feature space.
3. For the MNIST dataset, the results ranged between approximately 0.35 and 0.48 across all combinations of vectorisation methods and classifiers. This behaviour highlights the challenges associated with representing complex pixel-based image data using purely topological descriptors derived from point cloud transformations. In particular, the loss of intensity information and local spatial relationships may limit the discriminative power of persistence-based features in such tasks.
4. In the classification of the ModelNet10 dataset, the PI method achieved significantly higher performance (approximately 0.70–0.75) compared to the remaining types of PD representations. This confirms the strong capability of this method to aggregate topological information in 3D object classification tasks.
5. The PI and PL methods exhibited the highest stability with respect to random data splits, further supporting their suitability for practical applications.
6. The conducted complexity analysis indicated that the considered vectorisation methods differ in their computational and memory requirements. While BC and PS are computationally lightweight and suitable for large-scale settings, PI and PL provide more expressive representations at the cost of increased computational memory requirements. Consequently, the choice of vectorisation technique should depend not only on predictive accuracy, but also on available computational resources and the scale of the analysed data.
7. PD vectorisation methods are most effective in the tasks where class separability is encoded in global structure. For the data characterised by complex local dependencies, they should be integrated with geometric or learned representations within hybrid modelling frameworks.

On the basis of the obtained results, it can be concluded that PI may be regarded as a universal PD vectorisation method for classification tasks across various data types and model classes. In contrast, PL demonstrates high stability and interpretability, albeit with slightly lower performance.

Future research directions may include a systematic analysis of the robustness of individual vectorisation methods under different noise models and varying noise intensities. Although PDs are supported by theoretical stability guarantees, controlled empirical investigation may provide valuable insights into the practical behaviour of vectorised representations in noisy environments. Another important avenue for further study involves scalability analysis, particularly in the context of large-scale and high-density datasets. Given the computational complexity associated with PH, future work may explore performance characteristics, approximation strategies, and computational trade-offs for larger and more diverse data collections, especially in three-dimensional settings.

Additionally, further investigation may examine the interaction between topological representations and deep learning models, including the

impact of network architecture, depth, and model capacity on classification effectiveness. Such analyses may contribute to a better understanding of how model expressiveness influences the utilisation of topological features. Finally, a promising direction is the integration of topological descriptors with a classical statistical, geometric, or learned representations, potentially enabling complementary information fusion and improved predictive performance.

## REFERENCES


1. Carlsson G. Topology and data. Bulletin of the American Mathematical Society. 2009; 46(2): 255–308. https://doi.org/10.1090/S0273-0979-09-01249-X
2. Edelsbrunner H., Harer J. Computational topology: an introduction. American Mathematical Society. 2010. https://doi.org/10.1007/978-3-540-33259-6_7
3. Edelsbrunner H., Letscher D., Zomorodian A. Topological persistence and simplification. Discrete & computational geometry. 2002; 28: 511–533. https://doi.org/10.1007/s00454-002-2885-2
4. Zomorodian A., Carlsson G. Computing persistent homology. Discrete & Computational Geometry. 2005; 33(2): 249–274. https://doi.org/10.1007/s00454-004-1146-y
5. Bubenik P. Statistical topological data analysis using persistence landscapes. Journal of Machine Learning Research. 2015; 16(1): 77–102.
6. Chung Y. M., Lawson A. Persistence curves: a canonical framework for summarizing persistence diagrams. Advances in Computational Mathematics. 2022; 48(1): 6. https://doi.org/10.1007/s10444-021-09893-4
7. Adams H., Chepushtanova S., et al. Persistence images: a stable vector representation of persistent homology. Journal of Machine Learning Research. 2017; 18(8): 1–35.
8. Chazal F., Fasy B. T., et al. Stochastic convergence of persistence landscapes and silhouettes. In: Proceedings of the 30th Annual Symposium on Computational Geometry. Kyoto, Japan, June; 2014. https://doi.org/10.1145/2582112.2582128
9. Chazal F., Michel B. An introduction to topological data analysis: fundamental and practical aspects for data scientists. Frontiers in artificial intelligence. 2021; 4: 667963. https://doi.org/10.3389/frai.2021.667963
10. Wasserman L. Topological data analysis. Annual review of statistics and its application. 2018; 5(1): 501–532. https://doi.org/10.1146/annurev-statistics-031017-100045
11. Bauer U. Ripser: efficient computation of Vietoris–Rips persistence barcodes. Journal of Applied and Computational Topology. 2021; 5(3): 391–423. https://doi.org/10.1007/s41468-021-00071-5
12. Cohen-Steiner D., Edelsbrunner H., Harer J. Stability of persistence diagrams. In Proceedings of the 21st Annual Symposium on Computational Geometry. Pisa, Italy, June; 2005. https://doi.org/10.1007/s00454-006-1276-5
13. Yang L., Oyen D., Wohlberg B. Image classification using topological features automatically extracted from graph representation of images. In: Proceedings of the International Workshop on Mining and Learning with Graphs (MLG). Anchorage, AK, USA, August; 2019.
14. Peng Y., Wang H., Sonka M., Chen D. PHG-Net: persistent homology guided medical image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA, June; 2024. https://doi.org/10.48550/arXiv.2311.17243
15. Elyaagoubi A.B., Chung M.K., Ombao H. Topological data analysis for multivariate time series data. Entropy. 2023; 25(11): 1509. https://doi.org/10.3390/e25111509
16. Salazar Martinez D.A., Razmarashooli A., Chua Y.K., et al. Fast topological data analysis features for nonstationary time series. In: Proceedings of the Dynamic Data Driven Applications Systems (DDDAS) Conference. Philadelphia, PA, USA; 2024. https://doi.org/10.1007/978-3-031-94895-4_38
17. Zhu Y., Singh Y., Younis K., Bao S., Huo Y. Persistence image from 3D medical image: superpixel and optimized Gaussian coefficient. In: Proceedings of SPIE – The International Society for Optical Engineering. San Diego, CA, USA; 2025. https://doi.org/10.1117/12.3047078
18. Bukkuri A., Andor N., Darcy, I.K. Applications of topological data analysis in oncology. Frontiers in artificial intelligence. 2021; 4: 659037. https://doi.org/10.3389/frai.2021.659037
19. Sun C. Exploration of mapper-a method for topological data analysis. In: Proceedings of the 2020 International Conference of Information Science, Parallel and Distributed Systems (ISPDS). Xi'an, China, August; 2020. https://doi.org/10.1109/ISPDS51347.2020.00036
20. Sekuloski P., Ristovska V.D. Image classification using deep neural networks and persistent homology. In: Proceedings of the International Conference on ICT Innovations. Skopje, North Macedonia, September; 2023. https://doi.org/10.1007/978-3-031-54321-0_11
21. Turkes R., Montufar G. F., Otter N. On the effectiveness of persistent homology. Advances in Neural Information Processing Systems. 2022; 35: 35432–35448. https://doi.org/10.48550/arXiv.2206.10551
22. Berwald J., Gottlieb J., Munch E. Computing Wasserstein distance for persistence diagrams on a

quantum computer. In: Proceedings of the Workshop on Topological Data Analysis and Visualization. 2018. https://doi.org/10.48550/arXiv.1809.06433

23. Hofer C., Kwitt R., Niethammer M., Uhl A. Deep learning with topological signatures. Advances in neural information processing systems. 2017; 30. https://doi.org/10.48550/arXiv.1707.04041
24. Liu Y.S., Wang L., Cen J., et al. Persistence landscape based topological data analysis for personalized arrhythmia classification. https://doi.org/10.20944/preprints201908.0320.v2
25. Matuk J., Kurtek S., Bharath K. Topo-geometric analysis of variability in point clouds using persistence landscapes. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2024; 1–13. https://doi.org/10.1109/TPAMI.2024.3451328
26. Barnes D., Polanco L., Perea J. A comparative study of machine learning methods for persistence diagrams. Frontiers in Artificial Intelligence. 2021; 4: 681174. https://doi.org/10.3389/frai.2021.681174
27. Cao Y., Leung P., Monod A. k-means clustering for persistent homology. Advances in Data Analysis and Classification. 2025; 19(1): 95–119. https://doi.org/10.1007/s11634-023-00578-y
28. Ali D., Asaad A., Jiménez M., et al. A Survey of vectorization methods in topological data analysis. IEEE transactions on pattern analysis and machine intelligence. 2023; 45(12): 14069–14080. https://doi.org/10.1109/TPAMI.2023.3308391
29. Perea J., Munch E., Khasawneh F. Approximating continuous functions on persistence diagrams using template functions. Foundations of Computational Mathematics. 2022; 23(4): 1215–1272. https://doi.org/10.1007/s10208-022-09567-7
30. Dau A., Bagnall A., Kamgar K., et al. The UCR time series archive. IEEE/CAA Journal of Automatica Sinica. 2019; 6(6): 1293–1305. https://doi.org/10.1109/JAS.2019.1911747
31. Noakes L. The Takens embedding theorem. International Journal of Bifurcation and Chaos. 1991; 1(4): 867–872.
32. LeCun Y. The MNIST database of handwritten digits. 1998.
33. Wu Z., Song S., Khosla A., et al. 3D ShapeNets: A deep representation for volumetric shapes. In: Proceesings of the IEEE Conference on Computer Vision and Pattern Recongition (CVPR). Boston, MA, USA, June; 2015. https://doi.org/10.1109/CVPR.2015.7298801
34. Hosmer Jr D.W., Lemeshow S., Sturdivant R.X. Applied logistic regression. John Wiley & Sons. 2013.
35. Chen T. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, CA, USA, August; 2016.
36. Jaiswal S. Multilayer perceptrons in machine learning: a comprehensive guide. Data Camp. 2024.
37. Bergstra J., Bengio Y. Random search for hyper-parameter optimization. Journal of Machine Learning Research. 2012; 13: 281–305.