

# Topological Machine Learning for Mixed Numeric and Categorical Data

Chengyuan Wu<sup>\*,†</sup> and Carol Anne Hargreaves<sup>\*,‡</sup>

*\*Data Analytics Consulting Centre, Department of Statistics and Applied Probability  
National University of Singapore, Singapore 117546, Singapore*

<sup>†</sup>*stawuc@nus.edu.sg*

<sup>‡</sup>*stacah@nus.edu.sg*

Received (Day Month Year)

Revised (Day Month Year)

Accepted (Day Month Year)

Topological data analysis is a relatively new branch of machine learning that excels in studying high-dimensional data, and is theoretically known to be robust against noise. Meanwhile, data objects with mixed numeric and categorical attributes are ubiquitous in real-world applications. However, topological methods are usually applied to point cloud data, and to the best of our knowledge there is no available framework for the classification of mixed data using topological methods. In this paper, we propose a novel topological machine learning method for mixed data classification. In the proposed method, we use theory from topological data analysis such as persistent homology, persistence diagrams and Wasserstein distance to study mixed data. The performance of the proposed method is demonstrated by experiments on a real-world heart disease dataset. Experimental results show that our topological method outperforms several state-of-the-art algorithms in the prediction of heart disease.

*Keywords:* Topological data analysis; machine learning; artificial intelligence; mixed data; heart disease.

## 1. Introduction

Topological data analysis (TDA) is a relatively new subject that is gaining popularity in many fields, such as network analysis,<sup>1,2</sup> biomolecular chemistry,<sup>3,4</sup> and drug design.<sup>5,6</sup> Topological data analysis is often referred to as studying the “shape” of data, in order to deduce fundamental characteristics of the data. The primary tool used in TDA is persistent homology,<sup>7,8</sup> though there are also other tools such as Mapper,<sup>9,10</sup> discrete Morse theory,<sup>11,12,13</sup> as well as other techniques from algebraic topology.<sup>14,15,16,17</sup> It is generally acknowledged that topological data analysis is effective at analyzing high-dimensional noisy data.<sup>18,19</sup> We also remark that topological methods have also recently gained prominence in physics, with the 2016 Nobel Prize in Physics being awarded for theoretical discoveries of topological phase transitions and topological phases of matter.<sup>20</sup>

In real-world applications, data sets often have both numeric and categorical attributes. The coexistence of numeric and categorical variables often makes machine

<sup>†</sup>Corresponding author.

learning methods designed for single-type data inapplicable to mixed-type data.<sup>21,22</sup> Traditionally, TDA is usually applied to point cloud data or spatial data.<sup>23</sup> The strengths of TDA include the property of being coordinate-free<sup>24,19</sup> (independent of the coordinate system chosen), as well as being translation-invariant and rotation-invariant.<sup>25,26</sup> A drawback of these strengths is that it may be hard for TDA to effectively analyze data that is sensitive to choice of coordinates, translation, and/or rotation. Examples of such data include data with heterogeneous features, where each coordinate represents a fundamentally different feature (e.g. light, temperature, humidity).<sup>27</sup> To the best of our knowledge, there is currently no readily available framework for the classification of mixed numeric and categorical data using TDA. In view of the ubiquity of mixed-type data and the rising popularity of TDA, it is of interest to develop a topological machine learning method for mixed data.

In this paper, we propose a novel topological machine learning method for mixed data (TopMix). In our method, the categorical variables are first converted to binary variables via one-hot encoding. All predictor variables are subsequently standardized, and a basic *symmetry breaking*<sup>27</sup> technique is applied to the data for TDA to better deal with heterogeneous features. Subsequently, each data point is converted into a point cloud via multiple projection maps. We then generate persistence diagrams from the point cloud data, and calculate the Wasserstein distance between the persistence diagrams. Lastly, we use the  $k$ -nearest neighbors algorithm ( $k$ -NN) for supervised machine learning (classification). The basic workflow of our paper is summarized in Figure 1.

We remark that the technique of converting each data point into a point cloud via multiple projection maps is specialized for the setting of mixed numeric and categorical data. Hence, it is a new innovation that is not present in the authors' previous paper<sup>27</sup>, which focuses on the setting of multivariate time series data.

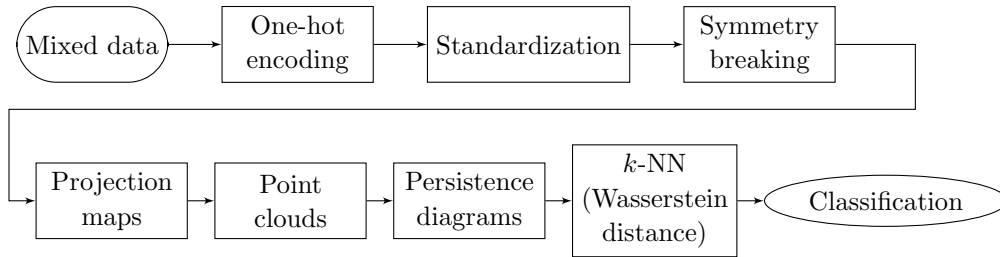


Fig. 1. Basic workflow of Topological Machine Learning for Mixed Numeric and Categorical Data.

For applications, we apply our method to heart disease prediction. Heart disease is the leading cause of death in the industrialized world.<sup>28</sup> For instance, in 2002, 696,947 people in the United States died of heart disease, compared with 557,271 deaths from cancer.<sup>29</sup> We use a dataset originating from the seminal paper by R. Detrano et al.<sup>30</sup> In the dataset, there are 14 attributes including numeric and cate-

gorical variables. The goal is to predict whether a patient has heart disease ( $> 50\%$  luminal narrowing of any major epicardial vessel) or not. We show that topological methods are effective in predicting heart disease using mixed data. Our topological method outperforms several state-of-the-art algorithms in the classification of heart disease.

The rest of the paper is organized as follows. We first review some related work in Section 1.1. This is followed by a brief introduction to the background information on TDA in Section 2. In Section 3, we present our topological machine learning method for mixed numeric and categorical data. In Section 4, we report the experimental results, which demonstrate the viability of the proposed method. Finally, we draw conclusions in Section 5.

### 1.1. *Related Work*

In the paper by X. Ni et al.,<sup>31</sup> the authors proposed a clustering method for mixed data based on a tree-structured graphical model. Their tree-structured model factorizes into a product of pairwise interactions. Furthermore, the authors leverage theory from TDA to adaptively merge trivial peaks of the density function into larger ones in order to achieve meaningful clusterings. Persistent homology theory is used to automatically determine the number of clusters in the data. An earlier seminal paper by Chazal et al.<sup>32</sup> introduced the novel idea of using topological persistence to guide the merging of clusters. Their algorithm provides additional feedback in the form of a persistence diagram, which the authors prove to reflect the prominences of the modes of the density. The algorithm requires rough estimates of the density at the data points, and knowledge of approximate pairwise distances between them, and hence is applicable in any metric space. Their method can be theoretically proven to output the correct number of clusters under certain mild sampling conditions.

In recent years, topological techniques have been effectively combined with machine learning or statistical methods. In the paper by C. Hofer et al.,<sup>33</sup> the authors introduced a technique that enables the input of topological signatures to deep neural networks for learning a task-optimal representation during training. An advantage of their method is that it learns the representation instead of mapping topological signatures to a pre-defined representation. P. Bubenik defined the persistence landscape,<sup>34</sup> which is a novel topological summary for data. Since this summary lies in a vector space, it is possible to combine it with tools from statistics and machine learning. A number of standard statistical tests can be used for statistical inference using persistence landscapes, for example the two-sample  $Z$ -test and Hotelling's  $T^2$  test.

C. Wu and C. A. Hargreaves<sup>27</sup> developed a framework for analyzing multivariate time series using TDA. The methodology includes converting the multivariate time series to point cloud data, calculating Wasserstein distances between the persistence diagrams, and using the  $k$ -NN algorithm for classification. For applications, the

authors focus on room occupancy detection based on 5 time-dependent variables (temperature, humidity, light, CO<sub>2</sub> and humidity ratio).

In the paper by J. Ji et al.,<sup>22</sup> the authors proposed a new initialization method for mixed data clustering. Prior to their paper, most of the initialization approaches are dedicated to partitional clustering algorithms which process either categorical or numerical data only. In the paper, the authors introduced a new definition of density to assess the cohesiveness of data objects with mixed numeric and categorical attributes.

A. Ahmad and L. Dey<sup>35</sup> presented a clustering algorithm that works well for data with mixed numeric and categorical features. The authors proposed a new cost function and distance measure based on co-occurrence of values. In their scheme,  $\delta(p, q)$  which denotes the distance between a pair of distinct values  $p$  and  $q$  of an attribute, is computed as a function of their co-occurrence with other attribute values. The contribution of a categorical attribute is inherent in the distance measure itself and need not be user defined.

The paper by J. Nahar et al.<sup>36</sup> investigates various computational intelligence techniques in the detection of heart disease. In the paper, the Cleveland dataset<sup>30</sup> from the UCI Machine Learning Repository<sup>37</sup> is used. In particular, the authors highlight the potential of a medical knowledge driven feature selection process for heart disease diagnosis. Experiments show that the medical knowledge based feature selection method has shown promise for use in heart disease diagnostics.

R. Das, I. Turkoglu and A. Sengur<sup>38</sup> explored the effective diagnosis of heart disease through neural network ensembles. Ensemble based methods can enable an increase in performance by combining several individual neural networks to train on the same task. The authors utilize SAS base software 9.1.3 in their methodology, and achieved good results using three independent neural network models in the ensemble model.

## 2. Background

We give a brief overview of the key concepts in TDA and persistent homology, and refer the reader to the appropriate references for more details. A classical text for algebraic topology is the book by A. Hatcher.<sup>39</sup> The survey article by H. Edelsbrunner and J. Harer,<sup>40</sup> as well as the review paper by R. Ghrist,<sup>23</sup> provide a superb introduction to persistent homology. In addition, the paper by A. Zomorodian and G. Carlsson<sup>8</sup> gives a comprehensive overview of persistent homology from a mathematical and computational perspective.

### 2.1. *Simplicial complexes*

Simplicial complexes are one of the main objects of study in algebraic topology. A simplicial complex can be regarded as a set composed of vertices, edges, triangles, and higher dimensional simplices.

More formally, a *simplicial complex*  $K$  is a collection of sets such that  $\sigma \in K$  and  $\tau \subseteq \sigma$  implies  $\tau \in K$ . The sets  $\sigma \in K$  are called the *simplices* of the simplicial complex  $K$ . We call the singleton sets  $\{v\}$  the *vertices* of  $K$ . The dimension of a simplex  $\sigma \in K$  is defined to be  $\dim(\sigma) = |\sigma| - 1$ , and we call a simplex of dimension  $k$  a *k-simplex*. Simplices of dimension 0, 1, 2, 3 represent a *vertex*, *edge*, *triangle* and *tetrahedron* respectively, as shown in Figure 2.

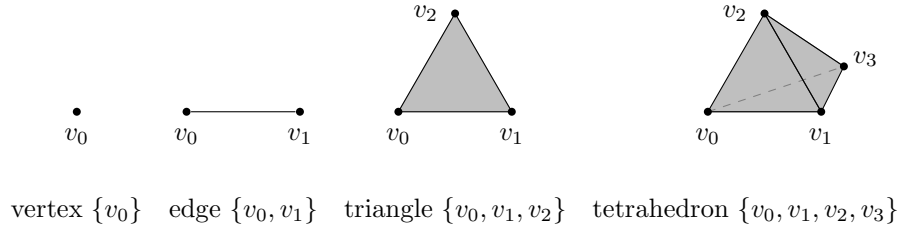


Fig. 2. A 0-simplex (vertex), 1-simplex (edge), 2-simplex (triangle) and 3-simplex (tetrahedron).

A type of simplicial complex frequently used in TDA is the *Vietoris-Rips complex* (or *Rips complex* for short).

**Definition 2.1.** Let  $\{x_i\}$  be a set of points in the Euclidean space  $\mathbb{R}^n$ . The Rips complex  $\mathcal{R}_\epsilon$  is the simplicial complex whose  $k$ -simplices consist of each subset of  $k + 1$  points  $\{x_j\}_{j=0}^k$  which are pairwise within distance  $\epsilon$ .

**Definition 2.2.** Let  $K$  be a simplicial complex. Suppose  $L$  is a simplicial complex such that every face of  $L$  belongs to  $K$ , that is,  $L \subseteq K$ . We say that  $L$  is a *simplicial subcomplex* of  $K$ .

We also introduce the notion of a *filtration* of a simplicial complex  $K$ , which is a nested sequence of complexes  $\emptyset = K^0 \subseteq K^1 \subseteq \dots \subseteq K^m = K$ . We say that  $K$  is a *filtered complex*.

## 2.2. Homology

The  $k$ th *chain group*  $C_k$  of a simplicial complex  $K$  is defined to be the free abelian group with basis to be the set of oriented  $k$ -simplices. The boundary operator  $\partial_k : C_k \rightarrow C_{k-1}$  is defined on an oriented simplex  $\sigma = [v_0, v_1, \dots, v_k]$  by

$$\partial_k(\sigma) = \sum_{i=0}^k (-1)^i [v_0, \dots, \hat{v}_i, \dots, v_k],$$

where  $\hat{v}_i$  denotes the deletion of the vertex  $v_i$ .

Subsequently, the  $k$ th *homology group* is defined as the quotient  $H_k = Z_k/B_k$ , where  $Z_k = \ker \partial_k$  and  $B_k = \text{Im } \partial_{k+1}$  are the *cycle group* and the *boundary group* respectively. The rank of the  $k$ th homology group  $\beta_k = \text{rank}(H_k)$  can be said to count the number of  $k$ -dimensional “holes” in  $K$ , as illustrated in Figure 3.

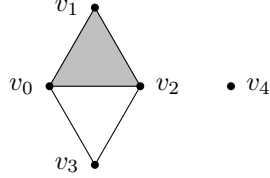


Fig. 3. For the above simplicial complex, we have  $\beta_0 = 2$  (2 connected components),  $\beta_1 = 1$  (1 “circular” hole which corresponds to the unshaded region) and  $\beta_2 = 0$  (no “voids”).

### 2.3. Persistent homology

Given a filtered complex  $K$ , we may define the corresponding boundary operators  $\partial_k^i$  and groups  $C_k^i$ ,  $Z_k^i$ ,  $B_k^i$  and  $H_k^i$  for the  $i$ th complex  $K^i$ . The  $p$ -persistent  $k$ th homology group of  $K^i$  is defined as

$$H_k^{i,p} = Z_k^i / (B_k^{i+p} \cap Z_k^i).$$

The filtered complex  $K$  is usually obtained by the construction of Rips complexes over a range of distances  $\epsilon$ . Persistent homology detects those topological features which persist over a parameter range, revealing meaningful structures in the data.

## 3. Topological Machine Learning for Mixed Numeric and Categorical Data

In this section, we describe our approach of using topological machine learning methods to analyze mixed data. A basic summary of the workflow can be found in Figure 1.

### 3.1. Notation

We first introduce a standard notation for mixed data, following the paper by Z. Huang and M. K. Ng,<sup>41</sup> as well as J. Ji et al.<sup>22</sup> Let  $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$  denote a dataset of  $n$  data objects. Each object  $X_i$  has  $m$  attributes  $A_1, A_2, \dots, A_m$ . We represent each  $X_i$  as a  $m$ -tuple  $(x_{i,1}, x_{i,2}, \dots, x_{i,m})$ . Each attribute  $A_j$  is associated with a domain of values, denoted by  $\text{Dom}(A_j)$ , which is either numeric (real numbers) or categorical (finite, unordered set). A categorical domain is generally represented by  $\text{Dom}(A_j) = \{a_{j,1}, a_{j,2}, \dots, a_{j,s}\}$ , where  $s$  is the number of possible categorical values for the categorical attribute  $A_j$ . Each data object  $X_i$  can be logically represented as a conjunction of attribute-value pairs:

$$[A_1 = x_{i,1}] \wedge [A_2 = x_{i,2}] \wedge \dots \wedge [A_m = x_{i,m}].$$

### 3.2. Methodology

#### 3.2.1. One-hot encoding

Firstly, we apply one-hot encoding to the mixed data, converting each categorical variable with  $s$  possible values to  $s$  binary variables. That is, each categorical attribute  $A_j$  with  $\text{Dom}(A_j) = \{a_{j,1}, a_{j,2}, \dots, a_{j,s}\}$  is replaced with  $s$  binary attributes  $B_1, B_2, \dots, B_s$ , with  $\text{Dom}(B_i) = \{0, 1\}$  for  $1 \leq i \leq s$ .

#### 3.2.2. Standardization

We standardize all variables (including binary variables) to have zero mean and unit variance. This is to ensure that all variables are on the same scale, preventing a feature with larger scale from dominating other features. We remark that the standardization of binary variables is also done in the algorithms KNNImpute (for categorical data)<sup>42</sup> and Lasso.<sup>43</sup>

#### 3.2.3. Symmetry breaking

Symmetry breaking refers to adding a fixed constant vector to each data object, with the purpose of enabling TDA methods to better distinguish point clouds that may just differ by translation or rotation. Symmetry breaking was introduced in the context of studying multivariate time series using topological methods.<sup>27</sup> Basically, symmetry breaking attempts to “disable” the translational / rotational invariance property of TDA for data that do not require it.

**Definition 3.1.** Let  $X = (x_1, x_2, \dots, x_m)$  be a data object represented as a  $m$ -tuple in  $\mathbb{R}^m$ . Let  $\mathbf{v} = (c_1, c_2, \dots, c_m)$  be a fixed vector in  $\mathbb{R}^m$ . We define the new data object  $X'$  obtained by *symmetry breaking* (of  $X$ ) to be  $X' = X + \mathbf{v}$ .

An example of the fixed vector is  $\mathbf{v} = (5, 6, 7, \dots, m + 4)$ . We will be using this fixed vector in the paper. In Section 3.3, we will illustrate how symmetry breaking can be useful in analyzing data with heterogeneous features, as well as explain our heuristic choice of fixed vector  $\mathbf{v}$ .

#### 3.2.4. Projection maps

After applying symmetry breaking, the new data object

$$X' = (x_1 + c_1, x_2 + c_2, \dots, x_m + c_m)$$

is a single point in  $\mathbb{R}^m$ . However, a single point has trivial homology and trivial persistent homology, hence we will need a point cloud (set of multiple data points in Euclidean space) in order for topological methods to work. In contrast, in the authors' previous paper on multivariate time series<sup>27</sup>, the point cloud resulting from a time window of length  $w > 1$  already consists of multiple data points. Hence, there

was no need for the technique of projection maps in the setting of multivariate time series data.

**Definition 3.2.** The  $i$ th projection map  $p_i : \mathbb{R}^m \rightarrow \mathbb{R}^m$  is defined by

$$p_i(y_1, y_2, \dots, y_i, \dots, y_m) = (y_1, y_2, \dots, 0, \dots, y_m).$$

The projection map  $p_i$  changes the  $i$ th coordinate of a vector to 0. This is equivalent to projecting the vector onto the hyperplane  $H = \{(x_1, x_2, \dots, x_m) \in \mathbb{R}^m \mid x_i = 0\}$ . The projection map defined above is idempotent, namely  $p_i \circ p_i = p_i$ .

### 3.2.5. Point clouds

We define the point cloud  $S(X')$  associated to the data object  $X' \in \mathbb{R}^m$  to be

$$S(X') = \{X'\} \cup \{p_1(X'), p_2(X'), \dots, p_m(X')\}.$$

That is,  $S(X')$  consists of  $m + 1$  points, namely the point  $X'$  as well as the  $m$  projected points  $p_1(X'), p_2(X'), \dots, p_m(X')$ . We show an example for the case  $m = 3$  in Figure 4.

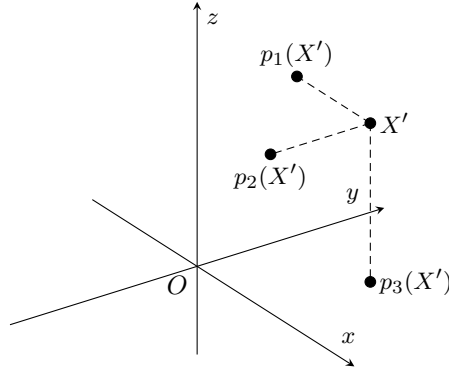


Fig. 4. For  $m = 3$ , the point cloud  $S(X')$  consists of  $X'$ , as well as the 3 projected points  $p_1(X')$ ,  $p_2(X')$  and  $p_3(X')$ . The 3 projected points are projections of  $X'$  onto the  $yz$ -,  $xz$ -, and  $xy$ -planes respectively.

The point cloud  $S(X')$  contains intrinsic information about the data object  $X'$  in the form of distances between the points in  $S(X')$ . For instance, if  $X' = (y_1, y_2, \dots, y_m)$ , we can calculate the following Euclidean distances between  $X'$  and its projections:

$$d(X', p_i(X')) = |y_i|, \quad (1)$$

$$d(p_i(X'), p_j(X')) = \sqrt{y_i^2 + y_j^2}, \quad \text{for } i \neq j. \quad (2)$$



### 3.2.6. Persistence diagrams

A persistence diagram<sup>44</sup> is a multiset of points in the space  $\Delta := \{(b, d) \in \mathbb{R}^2 \mid b, d \geq 0, b \leq d\}$ . Each point  $(b, d)$  represents a persistent generator (of a given dimension), where  $b$  denotes the birth of the generator and  $d$  its death. In brief, the persistence diagram is a visual representation of the persistent homology of a point cloud. The persistence diagram is independent of choice of generators and hence is unique.<sup>45</sup> A notable result is the stability of persistence diagrams with respect to Hausdorff distance, bottleneck distance,<sup>44</sup> as well as Wasserstein distance.<sup>45</sup> Such stability results give TDA the benefit of being robust to noise.

For the consideration of readability, we include a concrete example that illustrates the relationship between the persistent homology of a point cloud and its persistence diagram. Consider the point cloud  $S = \{(0, 0), (1, 0)\}$  consisting of two points (0-simplices). At the start of the filtration process, there are two separate connected components (namely the two 0-simplices in  $S$ ), hence this corresponds to two points on the persistence diagram (Figure 5) with birth time 0. At the filtration stage of  $\epsilon = 1$ , the Rips complex  $\mathcal{R}_\epsilon$  now consists of only one single connected component (namely the 1-simplex consisting of the two points in  $S$  and the edge joining them). This corresponds to a death time of 1, which explains the point  $(0, 1)$  on the persistence diagram. This 1-simplex theoretically persists to infinity (death time of infinity), but for practical purposes in the code we have to set a maximum value of  $\epsilon$  for the Rips filtration (in this case `maxscale = 5`). Hence, this explains the point  $(0, 5)$  on the persistence diagram.

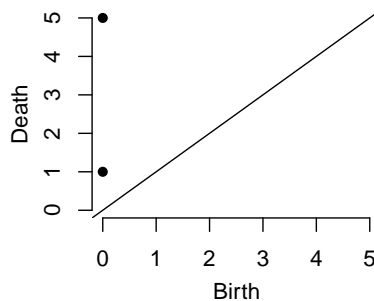


Fig. 5. The persistence diagram (dimension 0) for the point cloud  $S$ .

### 3.2.7. $k$ -NN (Wasserstein distance)

The Wasserstein distance<sup>46,45,47</sup> is commonly used to compare between two persistence diagrams.

**Definition 3.3.** The  $p$ -th Wasserstein distance between two persistence diagrams  $D_1, D_2$  (of the same dimension) is defined to be

$$W_p(D_1, D_2) = \left( \inf_{\varphi: D_1 \rightarrow D_2} \sum_{x \in D_1} \|x - \varphi(x)\|_\infty^p \right)^{1/p},$$

where the infimum is taken over all bijections  $\varphi$  between  $D_1$  and  $D_2$ .

As  $p$  tends to infinity, the Wasserstein distance  $W_p$  approaches the bottleneck distance  $W_\infty$ . The bottleneck distance captures the most perturbed topological feature (or the extreme behavior) of a point cloud, and can lead to noisier results than the Wasserstein distance.<sup>48</sup>

For this paper, we will use the Wasserstein distance with  $p = 1$ , also known as the 1-Wasserstein distance or “earth mover’s distance”. The 1-Wasserstein distance is widely utilized in computer science,<sup>49,50</sup> including a recent usage in generative adversarial networks.<sup>51</sup>

Subsequently, to carry out classification (supervised machine learning), we use the  $k$ -nearest neighbors algorithm ( $k$ -NN) based on the Wasserstein distance. For each point cloud  $S(X')$  (corresponding to a data object  $X'$ ) in the test set, we will determine its  $k$ -nearest neighbors  $\{S(Y_1), S(Y_2), \dots, S(Y_k)\}$  in the training set, with respect to the Wasserstein distance. Finally, we classify  $X$  based on the majority class of the elements in the set  $\{Y_1, Y_2, \dots, Y_k\}$ .

### 3.3. *Elaboration on symmetry breaking*

In this section, we illustrate how symmetry breaking, together with projection maps, can be helpful in analyzing data with heterogeneous features.

Consider two data objects  $X = (1, 2)$  and  $Y(2, 1)$ . Their associated point clouds are  $S(X) = \{(1, 2), (0, 2), (1, 0)\}$  and  $S(Y) = \{(2, 1), (0, 1), (2, 0)\}$ . We note that the pairwise distances between points in  $S(X)$  are exactly the same as the respective pairwise distances between points in  $S(Y)$ , namely 1, 2 and  $\sqrt{5}$ . This would mean that topological methods will not be able to distinguish between  $S(X)$  and  $S(Y)$ . The basic principle is that topological methods does not distinguish between point clouds that are related by “symmetry” (e.g. differ by rotation, translation, reflection).

Now, consider  $\mathbf{v} = (5, 6)$  such that we have

$$X' = X + \mathbf{v} = (6, 8)$$

and

$$Y' = Y + \mathbf{v} = (7, 7).$$

Then, the associated point clouds become  $S(X') = \{(6, 8), (0, 8), (6, 0)\}$  and  $S(Y') = \{(7, 7), (0, 7), (7, 0)\}$ . The pairwise distances between points in  $S(X')$  are 6, 8 and 10, while the pairwise distances between points in  $S(Y')$  are 7, 7 and  $7\sqrt{2}$ . Due to the difference in distances, TDA will be able to tell apart the point clouds  $S(X')$  and  $S(Y')$ , which is the desired outcome.

Next, we will explain our heuristic choice of fixed vector  $\mathbf{v} = (5, 6, 7, \dots, m+4)$  as mentioned in Section 3.2.3. The main reason is to try to make the components  $y_i$  in the data object  $X' = (y_1, y_2, \dots, y_m)$  all positive (or mostly positive). By observing Equations 1 and 2 in Section 3.2.5, we see that the distances  $|y_i|$  and  $\sqrt{y_i^2 + y_j^2}$  are not sensitive to signs (positive/negative) of the components  $y_i$ . For instance, there would be difficulty in distinguishing between say,  $X' = (1, 2)$  and  $Y' = (-1, 2)$ . Thus, we can see that our proposed method works better if components in the data objects are all positive (or mostly positive).

After the standardization step, each component  $x_i$  of the data object  $X = (x_1, x_2, \dots, x_m)$  comes from a distribution with mean 0 and standard deviation 1. Hence, by adding the fixed vector  $\mathbf{v} = (5, 6, 7, \dots, m+4)$  (note that all components of  $\mathbf{v}$  are 5 and above) to  $X$ , we have taken reasonable steps to try to make components of  $X' = X + \mathbf{v}$  mostly positive, since only components that are more than 5 standard deviations below the mean (in  $X$ ) would remain negative in  $X'$ . We remark that the above choice of fixed vector  $\mathbf{v}$  is not unique (there could be other choices of  $\mathbf{v}$  that work as well).

#### 4. Experimental Results

To evaluate the effectiveness of our proposed method, we use a real-world mixed dataset on heart disease<sup>30</sup> taken from the UCI Machine Learning Repository.<sup>37</sup> We focus on the Cleveland dataset, which comprises of data from patients referred for coronary angiography at the Cleveland Clinic. The algorithms were mostly implemented in Python, with the exception of computing persistence diagrams and Wasserstein distances using the R package TDA.<sup>52</sup> The codes in the paper are made publicly available on GitHub: <https://github.com/wuchengyuan88/topology-mixed-data>.

The 14 attributes of the heart disease dataset along with their data types and a brief description are presented in Table 1. For this dataset, heart disease is defined as greater than 50% luminal narrowing of any major epicardial vessel.

After one-hot encoding, there are a total of 25 predictive attributes. Hence, each patient is represented as a data object  $X$  in  $\mathbb{R}^m$ , where  $m = 25$ . Each patient will then be represented by a point cloud  $S(X')$  consisting of  $25 + 1 = 26$  points in  $\mathbb{R}^{25}$ .

For the construction of persistence diagrams, we use the `ripsDiag` function in the R package TDA. We show examples of two persistence diagrams from different classes in Figure 6. Qualitatively, we can visually observe some differences, for instance the persistence diagram for the patient of class 1 (heart disease) contains a denser cluster of points in the region corresponding to low death times of ap-

Table 1. Attributes of the heart disease dataset (6 numeric and 7 categorical predictive attributes, and 1 categorical target attribute).

	Attribute	Data type	Brief description
1	age	numeric	Age in years
2	sex	categorical	Gender of patient (1 = male; 0 = female)
3	cp	categorical	Chest pain type (1 = typical angina; 2 = atypical angina; 3 = non-anginal pain; 4 = asymptomatic)
4	trestbps	numeric	Resting blood pressure in mmHg
5	chol	numeric	Serum cholesterol in mg/dl
6	fbs	categorical	Fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
7	restecg	categorical	Resting electrocardiographic results (0 = normal; 1 = having ST-T wave abnormality; 2 = left ventricular hypertrophy)
8	thalach	numeric	Maximum heart rate achieved
9	exang	categorical	Exercise induced angina (1 = yes; 0 = no)
10	oldpeak	numeric	ST depression induced by exercise relative to rest
11	slope	categorical	Slope of the peak exercise ST segment (1 = upsloping; 2 = flat; 3 = downsloping)
12	ca	numeric	Number of major vessels (0-3) colored by fluoroscopy
13	thal	categorical	Heart status (3 = normal; 6 = fixed defect; 7 = reversible defect)
14	num (target attribute)	categorical	Presence of heart disease (0 = healthy; 1 = heart disease)

proximately 5 to 8. Quantitatively, the difference between persistence diagrams is measured by the Wasserstein distance, using the `wasserstein` function from the R package TDA. For this paper, distances between persistence diagrams are computed using 0 dimensional features, as we experimentally observe that 1 dimensional and higher features rarely appear in the persistence diagrams for our dataset.

We split our initial dataset (consisting of 297 patients in the Cleveland heart disease dataset) randomly into training, validation and test sets in a 60:20:20 ratio. A further summary of the split data sets can be found in Table 2.

Table 2. Description of split data sets.

Data set	Number of patients	Data class distribution (%)	
		0 (healthy)	1 (heart disease)
Training set	179	54.19	45.81
Validation set	59	52.54	47.46
Test set	59	54.24	45.76

To choose a suitable value for the parameter  $k$  in the  $k$ -NN algorithm, we experiment with various values of  $k$  on the validation set. The  $k$  nearest neighbors will be selected from the training set based on the Wasserstein distance. We show the accuracy, sensitivity (true positive rate) and specificity (true negative rate) for various values of  $k$  in Table 3. We select  $k = 5$  as it corresponds to the highest accuracy, as well as relatively high sensitivity and specificity (above 70%).

With the chosen value of  $k = 5$ , we show the results for the test set in Table 4. We achieve a high level of accuracy, sensitivity (recall of positive class) and specificity

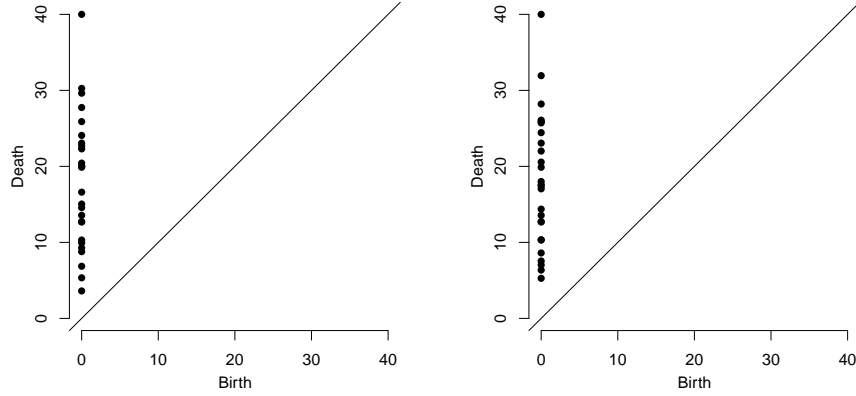


Fig. 6. The persistence diagram on the left belongs to a patient of class 0 (healthy), while that on the right belongs to a patient of class 1 (heart disease). The points refer to homological features in dimension 0.

Table 3. Accuracy, sensitivity and specificity for different values of  $k$  on the validation set.

Value of $k$	1	2	3	4	5	6	7	8	9	10
Accuracy (%)	69.49	74.58	76.27	77.97	<b>81.36</b>	79.66	77.97	76.27	77.97	79.66
Sensitivity (%)	64.29	85.71	78.57	82.14	<b>75.00</b>	78.57	71.43	71.43	67.86	75.00
Specificity (%)	74.19	64.52	74.19	74.19	<b>87.10</b>	80.65	83.87	80.65	87.10	83.87

(recall of negative class) on the test set.

Table 4. Results for test set (using  $k = 5$ ).

Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)		$F_1$ score (%)	
			(class 0)	(class 1)	(class 0)	(class 1)
<b>89.83</b>	88.89	90.62	90.62	88.89	90.62	88.89

Following best practices in machine learning, we also report the results for 10-fold cross-validation. The optimal value of  $k$  (for the  $k$ -NN algorithm) in the case of 10-fold cross-validation is found to be  $k = 16$ . We list the results in Table 5.

Table 5. Results for 10-fold cross-validation (using  $k = 16$ ).

Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)		$F_1$ score (%)	
			(class 0)	(class 1)	(class 0)	(class 1)
82.52	79.51	85.54	82.89	82.10	83.90	80.37

For reference, the accuracy of state-of-the-art algorithms reported in the literature typically ranges from around 60% to 90%.<sup>53,38,54,36,55,56</sup> We compare our

test set results with some previous results reported in literature. We remark that some of the accuracy results in the literature are based on 10-fold cross-validation on the total data, while other results are based on train-test splits. The 10-fold cross-validation methodology has the advantage of reduced bias as every data point gets to be tested exactly once and is used in training 9 times. However, some authors also argued that selecting the best training parameters on a validation set and reporting prediction on a test set (which is how we obtained our test accuracy) is more authentic than simply performing a 10-fold cross-validation on a training set.<sup>36</sup>

Table 6 gives the classification accuracies of our method and other previous approaches. Our Topological Machine Learning for Mixed Data method (TopMix) outperforms a number of other algorithms, including several state-of-the-art algorithms.

We remark that the Cleveland dataset in the UCI Machine Learning Repository consists of 303 original instances (including 297 complete instances and 6 instances with missing attributes). We only use the 297 complete instances (approximately 98% of the full dataset) for confirming the efficiency of our method. For the results listed in Table 6, it is not clearly specified whether the authors used the reduced dataset of 297 instances or the full dataset. Hence, we also reprogram some of the methods in the list and show their accuracy for a fairer comparison. We use the Scikit-learn package in Python and reprogram 5 methods (SVM, Logistic regression, Decision tree, Naive Bayes, and Multi-layer Perceptron) using the reduced dataset. The results are recorded in Table 6 as well.

## 5. Conclusions

Data objects with mixed numeric and categorical attributes are common in real-world applications. However, many algorithms are not compatible with mixed data and can only work on single-type data, that is, either numeric or categorical data. On the other hand, TDA is a rapidly emerging machine learning method that has benefits of robustness to noise and effectiveness in high dimensions. However, traditionally TDA is applied to point cloud data, not mixed data.

In this paper, we proposed a novel topological machine learning method to classify mixed numeric and categorical data. In our method, we utilize theory from TDA such as persistent homology, persistence diagrams and Wasserstein distance in order to study mixed data. In doing so, we expand the repertoire of TDA to include mixed data.

We test our proposed method on a heart disease dataset from the UCI machine learning repository. The experimental results demonstrate that the proposed method is effective at predicting heart disease, and also outperforms several state-of-the-art algorithms.

In conclusion, our paper represents a first step towards using TDA to classify mixed numeric and categorical data and can be viewed as a proof of concept that

Table 6. Classification accuracies obtained with our proposed Topological Machine Learning for Mixed Data method (TopMix) and other classifiers from literature. We have also reprogrammed some of the methods ourselves using classifiers from Scikit-learn, labelled as Scikit-learn (2020) under the “Source” column.

Source	Method	Accuracy (%)
ToolDiag	IB1-4	50.00
WEKA, RA	InductH	58.50
ToolDiag, RA	RBF	60.00
WEKA, RA	FOIL	64.00
ToolDiag, RA	MLP+BP	65.60
Scikit-learn (2020)	Decision tree	67.80
WEKA, RA	T2	68.10
S. Pouriyeh et al. (2017)	SCRL	69.96
WEKA, RA	1R	71.40
WEKA, RA	IB1c	74.00
Scikit-learn (2020)	Naive Bayes	76.27
J. Nahar et al. (2013)	J48	76.57
WEKA, RA	K*	76.70
J. Nahar et al. (2013)	IBK	76.90
R. Detrano	Logistic regression	77.00
S. Pouriyeh et al. (2017)	Decision tree	77.55
J. Nahar et al. (2013)	AdaBoostM1+CFS	77.94
J. Nahar et al. (2013)	SMO+MFS	77.95
N. Cheung (2001)	BNNF	80.96
N. Cheung (2001)	BNND	81.11
N. Cheung (2001)	C4.5	81.11
N. Cheung (2001)	Naive Bayes	81.48
J. Nahar et al. (2013)	PART	81.52
S. Pouriyeh et al. (2017)	SVM	84.15
Scikit-learn (2020)	Multi-layer Perceptron	84.75
J. Nahar et al. (2013)	PART+MFS	86.77
Polat et al. (2006)	Fuzzy-AIRS-Knn based system	87.00
Scikit-learn (2020)	Logistic regression	88.14
Scikit-learn (2020)	SVM	88.14
R. Das et al. (2009)	Neural networks ensemble	89.01
<b>TopMix (Test accuracy)</b>	<b>Topological machine learning</b>	<b>89.83</b>

methods from TDA are effective in the domain of mixed data, as well as in heart disease prediction.

## Acknowledgments

The authors wish to thank the referees most warmly for numerous suggestions that have improved the exposition of this paper.

## References

1. C. Carstens and K. Horadam, Persistent homology of collaboration networks, *Mathematical problems in engineering* **2013** (2013).
2. D. Horak, S. Maletić and M. Rajković, Persistent homology of complex networks, *Journal of Statistical Mechanics: Theory and Experiment* **2009**(03) (2009) p. P03034.
3. K. Xia, Z. Li and L. Mu, Multiscale persistent functions for biomolecular structure characterization, *Bulletin of Mathematical Biology* **80**(1) (2018) 1–31.
4. K. Xia, Z. Zhao and G.-W. Wei, Multiresolution persistent homology for excessively large biomolecular datasets, *The Journal of Chemical Physics* **143**(13) (2015) p. 10B603\_1.
5. Z. Cang and G.-W. Wei, Integration of element specific persistent homology and ma-

- chine learning for protein-ligand binding affinity prediction, *International journal for numerical methods in biomedical engineering* **34**(2) (2018) p. e2914.
6. K. Wu, Z. Zhao, R. Wang and G.-W. Wei, TopP-S: Persistent homology-based multi-task deep neural networks for simultaneous predictions of partition coefficient and aqueous solubility, *Journal of computational chemistry* **39**(20) (2018) 1444–1454.
  7. H. Edelsbrunner and D. Morozov, Persistent homology: theory and practice, tech. rep., Ernest Orlando Lawrence Berkeley National Laboratory, Berkeley, CA (US) (2012).
  8. A. Zomorodian and G. Carlsson, Computing persistent homology, *Discrete & Computational Geometry* **33**(2) (2005) 249–274.
  9. M. Nicolau, A. J. Levine and G. Carlsson, Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival, *Proceedings of the National Academy of Sciences* **108**(17) (2011) 7265–7270.
  10. G. Singh, F. Mémoli and G. E. Carlsson, Topological methods for the analysis of high dimensional data sets and 3D object recognition., in *SPBG2007*, pp. 91–100.
  11. R. Forman, A user’s guide to discrete Morse theory, *Sém. Lothar. Combin* **48** (2002) p. 35pp.
  12. J. Reininghaus, D. Günther, I. Hotz, S. Prohaska and H.-C. Hege, TADD: A computational framework for data analysis using discrete Morse theory, in *International Congress on Mathematical Software Springer2010*, pp. 198–208.
  13. C. Wu, S. Ren, J. Wu and K. Xia, Discrete Morse theory for weighted simplicial complexes, *Topology and its Applications* **270** (2020) p. 107038.
  14. J. Hansen and R. Ghrist, Toward a spectral theory of cellular sheaves, *Journal of Applied and Computational Topology* **3**(4) (2019) 315–358.
  15. D. Letscher, On persistent homotopy, knotted complexes and the Alexander module, in *Proceedings of the 3rd Innovations in Theoretical computer Science Conference2012*, pp. 428–441.
  16. C. Wu, S. Ren, J. Wu and K. Xia, Weighted fundamental group, *Bulletin of the Malaysian Mathematical Sciences Society* (2020) 1–24.
  17. C. Wu, S. Ren, J. Wu and K. Xia, Magnus representation of genome sequences, *Journal of theoretical biology* **480** (2019) 104–111.
  18. H. Adams, T. Emerson, M. Kirby, R. Neville, C. Peterson, P. Shipman, S. Chepushanova, E. Hanson, F. Motta and L. Ziegelmeier, Persistence images: a stable vector representation of persistent homology, *Journal of Machine Learning Research* **18**(8) (2017) 1–35.
  19. M. Offroy and L. Duponchel, Topological data analysis: A promising big data exploration tool in biology, analytical chemistry and physical chemistry, *Analytica chimica acta* **910** (2016) 1–11.
  20. F. D. M. Haldane, Nobel lecture: Topological quantum matter, *Reviews of Modern Physics* **89**(4) (2017) p. 040502.
  21. C.-C. Hsu, Y.-P. Huang and K.-W. Chang, Extended naive Bayes classifier for mixed data, *Expert Systems with Applications* **35**(3) (2008) 1080–1083.
  22. J. Ji, W. Pang, Y. Zheng, Z. Wang and Z. Ma, An initialization method for clustering mixed numeric and categorical data based on the density and distance, *International Journal of Pattern Recognition and Artificial Intelligence* **29**(07) (2015) p. 1550024.
  23. R. Ghrist, Barcodes: the persistent topology of data, *Bulletin of the American Mathematical Society* **45**(1) (2008) 61–75.
  24. P. Y. Lum, G. Singh, A. Lehman, T. Ishkanov, M. Vejdemo-Johansson, M. Alagappan, J. Carlsson and G. Carlsson, Extracting insights from the shape of complex data using topology, *Scientific reports* **3** (2013) p. 1236.
  25. T. Bonis, M. Ovsjanikov, S. Oudot and F. Chazal, Persistence-based pooling for shape



- pose recognition, in *International Workshop on Computational Topology in Image Context* Springer2016, pp. 19–29.
26. F. A. Khasawneh and E. Munch, Chatter detection in turning using persistent homology, *Mechanical Systems and Signal Processing* **70** (2016) 527–541.
  27. C. Wu and C. A. Hargreaves, Topological machine learning for multivariate time series, *arXiv preprint arXiv:1911.12082* (2019).
  28. M. S. Lauer, E. H. Blackstone, J. B. Young and E. J. Topol, Cause of death in clinical research: time for a reassessment?, *Journal of the American College of Cardiology* **34**(3) (1999) 618–620.
  29. R. Twombly, Cancer surpasses heart disease as leading cause of death for all but the very elderly, *Journal of the National Cancer Institute* **97**(5) (2005) 330–331.
  30. R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J.-J. Schmid, S. Sandhu, K. H. Guppy, S. Lee and V. Froelicher, International application of a new probability algorithm for the diagnosis of coronary artery disease, *The American journal of cardiology* **64**(5) (1989) 304–310.
  31. X. Ni, N. Quadrianto, Y. Wang and C. Chen, Composing tree graphical models with persistent homology features for clustering mixed-type data, in *Proceedings of the 34th International Conference on Machine Learning-Volume 70* JMLR. org2017, pp. 2622–2631.
  32. F. Chazal, L. J. Guibas, S. Y. Oudot and P. Skraba, Persistence-based clustering in Riemannian manifolds, *Journal of the ACM (JACM)* **60**(6) (2013) 1–38.
  33. C. Hofer, R. Kwitt, M. Niethammer and A. Uhl, Deep learning with topological signatures, in *Advances in Neural Information Processing Systems*2017, pp. 1634–1644.
  34. P. Bubenik, Statistical topological data analysis using persistence landscapes, *The Journal of Machine Learning Research* **16**(1) (2015) 77–102.
  35. A. Ahmad and L. Dey, A k-mean clustering algorithm for mixed numeric and categorical data, *Data & Knowledge Engineering* **63**(2) (2007) 503–527.
  36. J. Nahar, T. Imam, K. S. Tickle and Y.-P. P. Chen, Computational intelligence for heart disease diagnosis: A medical knowledge driven approach, *Expert Systems with Applications* **40**(1) (2013) 96–104.
  37. D. Dua and C. Graff, UCI machine learning repository (2017).
  38. R. Das, I. Turkoglu and A. Sengur, Effective diagnosis of heart disease through neural networks ensembles, *Expert systems with applications* **36**(4) (2009) 7675–7680.
  39. A. Hatcher, Algebraic topology. 2002, *Cambridge UP, Cambridge* **606**(9) (2002).
  40. H. Edelsbrunner and J. Harer, Persistent homology—a survey, *Contemporary mathematics* **453** (2008) 257–282.
  41. Z. Huang and M. K. Ng, A fuzzy k-modes algorithm for clustering categorical data, *IEEE transactions on Fuzzy Systems* **7**(4) (1999) 446–452.
  42. D. J. Stekhoven and P. Bühlmann, MissForestnon-parametric missing value imputation for mixed-type data, *Bioinformatics* **28**(1) (2012) 112–118.
  43. R. Tibshirani, The lasso method for variable selection in the Cox model, *Statistics in medicine* **16**(4) (1997) 385–395.
  44. D. Cohen-Steiner, H. Edelsbrunner and J. Harer, Stability of persistence diagrams, *Discrete & Computational Geometry* **37**(1) (2007) 103–120.
  45. D. Cohen-Steiner, H. Edelsbrunner, J. Harer and Y. Mileyko, Lipschitz functions have Lp-stable persistence, *Foundations of computational mathematics* **10**(2) (2010) 127–139.
  46. J. J. Berwald, J. M. Gottlieb and E. Munch, Computing Wasserstein distance for persistence diagrams on a quantum computer, *arXiv preprint arXiv:1809.06433* (2018).
  47. Y. Mileyko, S. Mukherjee and J. Harer, Probability measures on the space of persis-

- tence diagrams, *Inverse Problems* **27**(12) (2011) p. 124007.
48. M. Hajij, B. Wang, C. Scheidegger and P. Rosen, Visual detection of structural changes in time-varying graphs using persistent homology, in *2018 IEEE Pacific Visualization Symposium (PacificVis)* IEEE2018, pp. 125–134.
  49. J. Rabin, J. Delon and Y. Gousseau, A statistical approach to the matching of local features, *SIAM Journal on Imaging Sciences* **2**(3) (2009) 931–958.
  50. Y. Rubner, C. Tomasi and L. J. Guibas, The earth mover’s distance as a metric for image retrieval, *International journal of computer vision* **40**(2) (2000) 99–121.
  51. M. Arjovsky, S. Chintala and L. Bottou, Wasserstein generative adversarial networks, in *International Conference on Machine Learning*2017, pp. 214–223.
  52. B. T. Fasy, J. Kim, F. Lecci and C. Maria, Introduction to the R package TDA, *arXiv preprint arXiv:1411.1830* (2014).
  53. N. Cheung, Machine learning techniques for medical analysis, *School of Information Technology and Electrical Engineering* (2001).
  54. C. B. C. Latha and S. C. Jeeva, Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques, *Informatics in Medicine Unlocked* **16** (2019) p. 100203.
  55. K. Polat, S. Şahan and S. Güneş, Automatic detection of heart disease using an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism and k-nn (nearest neighbour) based weighting preprocessing, *Expert Systems with Applications* **32**(2) (2007) 625–631.
  56. S. Pouriye, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia and J. Gutierrez, A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease, in *2017 IEEE Symposium on Computers and Communications (ISCC)* IEEE2017, pp. 204–207.