Article

Persistence Landscape based Topological Data Analysis for Personalized Arrhythmia Classification

Yan Yan [1,2,3], Kamen Ivanov [1,2,3], Jian Cen [4], Qiuhua Liu [1,3] and Lei Wang [1,3,*]

[1] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[2] University of Chinese Academy of Sciences
[3] CAS Key Laboratory of Health Informatics, Chinese Academy of Sciences
[4] The Mindray Medical International Limited
[*] Correspondence: wang.lei@siat.ac.cn

Abstract: Data has shapes, and shapes may provide insights to data modeling and information extraction. Topological data analysis (TDA) paves new avenues in the evaluation of biomedical data, where algebraic-topological tools are used for knowledge discovery. In the present work, we apply TDA for personalized electrocardiographic signal classification toward arrhythmia analysis. First, to facilitate the TDA, signal samples are converted into point clouds using phase space reconstruction. Topological techniques are then used to extract the persistence landscapes from the point clouds as features used in the subsequent arrhythmia classification. The proposed persistence landscape based feature learning method is robust to the size of the training set. With only 20% of the full training dataset, it achieves a 100% accuracy for normal heartbeats, 97.13% for ventricular beats, 94.27% for supra-ventricular beats, and 94.27% for fusion beats. Thus the method can be trained for a single individual, allowing for personalized analysis systems. With the present study, we show that TDA could be a useful tool for biomedical signal analysis, with potential application in the personalized data processing.

Keywords: Electrocardiography Analysis; Persistence Landscape; Signal Analysis; Topological Data Analysis; Signal Classification; Biomedical Signal Analysis; persistent homology

1. Introduction

Electrocardiographic signals reflect cardiac electrical activity over time. The abnormal pattern of ECG signal is associated with cardiac arrhythmia as a result of severe cardiac risks like stroke, heart failure, and sudden death [1]. Arrhythmia is one of the main tasks that could be induced by aging or different diseases like diabetes, hypertension, and obesity. Recognizing and detection of arrhythmia of the ECG sensor data is one of the main tasks of healthcare applications such as smart home [2], rehabilitation [3], and mobile health[4,5].

Physiologically, the heartbeats are illustrations complex system involving turbulence and spatiotemporal wave propagation. The heart's physiological dynamical can be reflected by electrocardiographic (ECG) signals captured by the different kind of sensors. The signals are observation information of a dynamical system [6]. So the arrhythmia represented by abnormal ECG pattern can be considered as an abnormal state of the dynamical system. In the recognition and classification tasks, the phase-space reconstruction method [7,8] could provide essential features besides the traditional representation of statistical features.

As a typical method for phase-space reconstruction, the time-delay embedding had already been used to study dynamical systems. With time-delay embedding, the underlying dynamical systems can be recovered from observed time-series data, such as in industrial applications [9,10] and human body systems [11–13]. The idea that the human heart as a dynamical system and the application of the time-delay embedding technique for phase-space reconstruction from ECG signals was first proposed in [6]. The parameters of the human body system such as Lyapunov Exponents[14], Correlation Dimensions[15] had been used for the study of ECG signal dysfunction state. Other Nonlinear Dynamical System Analysis Techniques such as Detrended Fluctuation Analysis (DFA) [16], Recurrence Quantification Analysis[17,18], and Poincaré Plot [19] are also used in the ECG analysis applications. In this study, we focus on the study of phase-space reconstruction jointly with a recently fast developing technique of topological data analysis (TDA). The jointly nonlinear analysis methods could provide essential features for arrhythmia analysis.

Recently, a developing and growing trends in TDA brought the motivation for TDA features in ECG signal analysis [20]. We explored how the topological information of data based on geometry and topology inspire ECG analysis tasks. The TDA techniques had become a useful representation extraction tool for different complex data analysis applications [21–24]. In the field of signal analysis, the TDA showed potentials for classification and detection [20,25,26], other than the traditional statistical-based methods. TDA provided an alternative viewpoint for signal analysis and enriched the nonlinear dynamical system based signal processing research.

Arrhythmia patterns in ECG signals are the observation of states diversification in the heart system. Recent works use nonlinear parameters like entropy or Lyapunov parameter to describe the abnormal state. TDA tools provide an alternative way from the geometrical and topological structure of point clouds as an extension for nonlinear analysis. In [27] time-delay embedding and persistent homology theory had been used in wheeze detection for breath system. [6] use time-delay embedding to convert the ECG signal into a 2-D point cloud for cardiac system analysis. In [28] a term *topological signal representation* was proposed using time-delay embedding and TDA. Besides, Safarbali [29] proposed a statistical analysis using the time-of-life representation in persistent homology of TDA, toward the atrial fibrillation nonlinear dynamic analysis, and Dindin [30] used the Betti curves as an alternative features of the deep learning representations, using in a recognition system with a cascaded modular neural network.

TDA constitutes an alternative of traditional statistical methods used in the recognition applications. From a nonlinear dynamical system viewpoint, the system's information can contribute to the machine learning tasks, which depends on quite different theory compare to the statistical way. How to incorporate the system information into the machine learning statistical computation framework, a recent topological signature [31–33] termed persistence landscape answers. This work illustrates how to use the topological signatures in the arrhythmia classification system. In this spirit, the main contribution of this work include:

1. We proposed a topological data analysis driven framework for electrocardiographic analysis and arrhythmia classification (Section 2), and demonstrate that TDA is a practical solution for robust detection of ventricular ectopic beats and supra-ventricular ectopic beats. We validated it in the long term single lead ECG dataset.
2. The framework (Section 3.1) includes four stages: the ECG segments firstly embedded into the space as point clouds (Section 3.2); then the persistent homology was used to get the topological signatures (Section 3.3); with which the persistence landscapes based features achieved (Section 3.4); then performed using the random forests classifier for arrhythmia classification.
3. We illustrated that, with time delay embedding, the shape illustrations from different arrhythmia types are distinguishable (Section 4.2). In the meantime, Barcodes, persistence diagrams, and persistence landscapes are also compared (Section 4.4).
4. When applying the proposed method on a balanced dataset, the classification performance of normal heartbeat class are 100% recognized, ventricular beats for 97.13%, supra-ventricular beats for 94.27% and fusion beats for 94.27%. Also, the performance obtained with a small training set shows good performance as well. (Section 4.5)

doi:10.20944/preprints201908.0320.v1

2. Background

2.1. Clinical Background: Arrhythmia Analysis

Arrhythmia refers to an irregular heartbeat is an essential event to be captured and analyzed in modern healthcare systems. In most heart monitoring systems, the automatic detection and classification of the heartbeat signal are one of the most fundamental aspects for the conservation of life and personalized medicine in today's growing population. As the Association for the Advancement of Medical Instrumentation (AAMI, Arlington, VA) recommended, the heartbeats could be classified into one of the five ECG patterns: N (normal beats originating in the sinus node), S (supra-ventricular ectopic beats), V (ventricular ectopic beats), F (fusion beats), and Q (unclassifiable beats, ignored in this study) for heart monitoring. The feature extracted from the clinical ECG systems is an essential component in the modern arrhythmia systems. Some of the primary examples like morphological features with shapes, amplitudes, and durations of ECG signal [34,35]; frequency domain features obtained with wavelet transformation explained and used recently [36]; statistical features derived with higher-order statistics [37]; and also the new deep learning-based feature extraction methods [38,39]. In this work, we consider the arrhythmia phenomenon as a representation of heart system abnormality. With the nonlinear analysis tools from a dynamical system angle, we found an entirely different tool in arrhythmia analysis.

2.2. The Heart as Dynamical Systems

The nonlinear analysis and chaos theory used in studying the biological system had inspired some of the works in heart system analysis [40]. The features based on nonlinear dynamical systems like the correlation dimension, Lyapunov exponents, and entropy-based parameters explored in the previous studies. A useful review of the work of nonlinear analysis can refer to [41,42]. Typically, the nonlinear dynamic system methods are used in the atrial fibrillation analysis or long term ECG heart rate variability analysis. The phase space reconstruction method used in the nonlinear analysis comes from the dynamical system theory. The phase space is an abstract space, which could graphically represent the possible states of a dynamical system [43]. From the Takens's theorem, there should be an actual number of variables which govern all possible behaviors of the dynamical system, which is not possible to get all the variables in the real-world situation [44]. However, if one variable of the system (i.e. a dimension which could be measured), like x is accessible, then the full dynamics of the system could be reconstructed from the observed x, with time delay embedding [45]. The embedded trajectory and points in the phase space could be considered as a kind of descriptor of the original signal. Using TDA to analysis the phase space reconstruction from a different pattern of ECG time series could be used in arrhythmia analysis.

2.3. Topological Data Analysis

TDA comes from the intuition that the shape of data matters. The primary technique used in this work is persistent homology, an adaptation of homology to point cloud data, which was first proposed by [46]. The topological summaries like Barcode [47] and persistence diagram [48,49] had been proposed as useful representations from the data point cloud. The kernel-based method is another option for TDA use in data analysis [50]. For the ease of use in statistical analysis, the persistence bag-of-words [51] and persistence landscape [31] had been used in data analysis. Recently, more and more work had been proposed using this topology inspired method because of its robustness and theoretical integrity. In time series analysis using TDA tools, [20] provided a useful resource for different types of time series analysis. Other notable work of time series analysis can be referred from [52–54].

In this work, the ECG signals are embedded into the corresponding abstract phase space, different record patterns with different shapes because of the physiological system's state difference could be discriminate by TDA features based learning models.

3. Methodology

The TDA methods extract information from the topological and geometrical properties of the data point clouds. In this study, we first build data point clouds for each time series with the phase state reconstruction based method of time delay embedding. With the extracted points we build the simplicial complexes, each with a radius parameter. Then, the persistent homologies for each data point cloud are acquired with the complexes. From the complexes, we can get the topological summaries, namely, Barcode or persistence diagram. Further, we use persistence landscape as extracted features for the classification task. With the persistence landscape representations from each time series, the random forest classifier was used for classification.

3.1. Proposed Framework

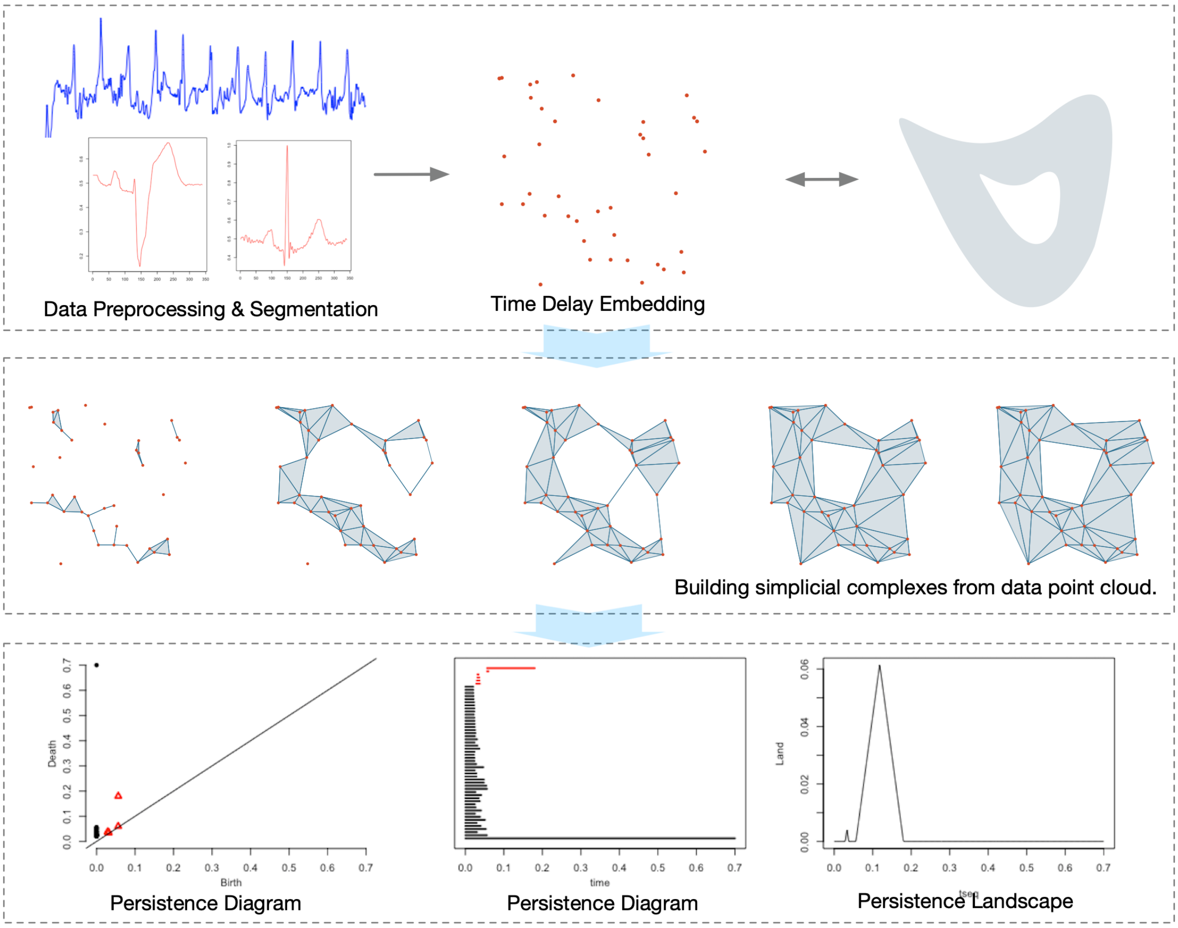

Figure 1. Proposed framework: firstly, the time series are pre-processed and segment into beats, then the heart beats can be embedded into data point clouds with the time delay embedding method, the point clouds have been assumed to be in some topological space; secondly, simplicial complex can be build upon the point clouds with rising radius parameters, the process composed to a series of simplicial complex namely, persistent homology; thirdly, topological summaries of Barcode, persistence diagrams are achieved, with them the statistical representation persistence landscape can be extracted

The overview of the method can be found in Figure 1, and the detailed description of the framework can be found in following sections.

3.2. Time Delay Embedding

The electrocardiography time series is a kind of nonlinear time-series, which can be considered as time-ordered observation data from some dynamical systems. The purpose of time-series analysis can be considered as the study of dynamics behind the observed time-ordered data [55]. The phase-space reconstruction is proposed for dynamics analysis, the reconstructed space may not be identical to the real system, but with an appropriate reconstruction, some of the properties of the system revealed because of the topological equivalent [56]. The delay reconstruction for ECG had been first analyzed in [6] for multiple heartbeats. In this work, we consider the difference between the phase space reconstructed from different types of a single heartbeat. With TDA, we could extract features from the phase space, which is an alternative way compared to the statistical methods.

Mathematically, suppose the time-series signal sequence $f(n), n \in \mathbb{Z}^+$ in which n is the signal sampling index. For a time delay embedding operation, let $\mathcal{S} \in \mathbb{Z}^+$ as the parameter of delay step, the dimension of the topological space to be embedded into is $d \in \mathbb{Z}^+$, then the time delay embedding at the time $t \in \mathbb{Z}^+$ can be illustrated as:

$$DE(f,t;s,d) = \{f(t), f(t+s), \cdots, f(t+(d-1)s)\} \tag{1}$$

The reconstruction of the phase space could convert the signals into higher dimensional phase-space, which approximate the phase-state of the real dynamics. The central problem in doing the phase space reconstruction is the determination of the parameters of time delay parameter τ and the embedded dimension d. Appropriate d and τ could approximate the dynamics better which could help for further analysis. Estimating good value for D and τ is quite challenging, lots of methods had been developed to choose the parameters [55,57]. In the traditional practice, τ was determined first and then d. However, in this work we set the dimension of the embedding space as $d = 2$. As stated in former study of [55] and [28], a larger d does not necessarily increase the classification performance. With time delay embedding, each segmented ECG beat waveform has been converted into a 2-D point cloud. The distribution on the 2-D varies when different arrhythmia type occurs, we give the point cloud illustration for different pattern of signal in Section 4.2.

3.3. Topological Data Analysis

Geometrical structures of the point clouds like components or existence of holes and voids in the space are the central elements for TDA. In TDA, the homology invariants could provide distinct different topological structure. Mathematically, homology is associates one vector space $H_i(X)$ to the space X for each natural number $i \in \{0, 1, 2, \ldots\}$, and represent the corresponding geometrical structures. For example, $H_0(X)$ counts for the linking path components, $H_1(X)$ counts for the number of holes, and $H_2(X)$ counts for voids, etc. These algebraic structures are robust to the underlying space transformation. As Figure 2 illustrated, suppose the data points clouds are lying on some topological spaces, the specific topological space could not be achieved, but with the approximation of the topological space with mesh, we can still get the properties of the topological space.

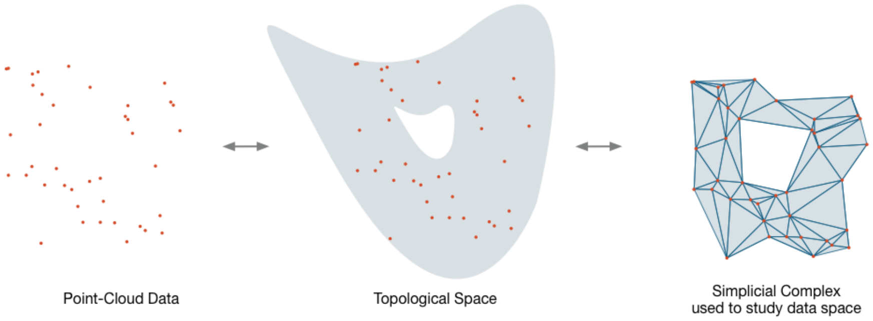

Figure 2. Topological space and simplicial complex.

Algorithmically, the simplicial complexes theory is the solution for the approximation task. We do not know the topological space X the data points lied on; instead, we use the data points to construct a discrete set S to build a simplicial complex to approximate the space. The $C\check{e}ch$ complex [58] and $Alpha$ complex [59] had been proved to be a useful technique to construct such simplicial complex with theoretical support of Nerve Theorem. However, the $C\check{e}ch$ complex needs intensive computation. Several reduction complexes had been proposed, such as $Vietoris - Rips$ complex [60,61], $Graph - induced$ complex [62], and $Sparsified\ C\check{e}ch$ complex [63]. The $Vietoris - Rips$ complex had already proved to be computational efficiently and used in data analysis; therefore we use the $Vietoris - Rips$ complex to approximate the data point cloud topological space, and later for topological analysis.

3.3.1. Simplicial Complex Graph Building

The embedded data point cloud is assumed as points with different distance notations, sometimes similarity or dissimilarity between separate points. With mathematical components simplicial complex or filtration (i.e., a nested family of simplicial complexes) in a topological space, they can reflect the structure of the data at different scales. Here we give an intuitive illustration for the simplicial complex building; more content can be referred from the respective reference.

A topology can be considered as a mathematical description of geometry objects, a topology on a set is defined as a collection of subsets, including the empty set and the whole set. With the data cloud generated with the data samples from the specific applications, simplicial complexes are used to study the shapes in the TDA methods. The simplicial complex can be considered as a generalized higher dimensional graph. A single point can be considered as 0-simplex, a line for a 1-simplex, a triangle for a 2-simplex and a tetrahedron for 3-simplex. The faces of a simplex are its boundaries, such as for a 1-simplex the faces are points, while for a 2-simplex the faces are line segments, and for a 3-simplex the faces are its three triangles (Figure 3). Generally, a subset consisted of $(k+1)$ data points is called a k-simplex. A simplicial complex is the collection of simplexes, together with their faces, an exhaustive description can be found in [64].

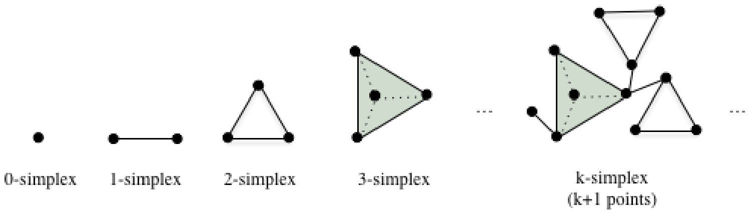

Figure 3. Simplex and simplicial complex.

Compared to the graph-based methods, the simplicial complex based methods could deal with triangles, tetrahedrons, and even more abstract objects. One can consider the simplicial complex as a generalization of graphs with higher-order relationships among the nodes [65]. Graphs are focused on modeling pairwise interaction while simplicial complex on higher-order interaction. After we get the data point cloud, the next task is to build the simplicial complex of the data, and the original data forms the vertex set. With the simplicial complex tools, next, we can get the filtration of a data point cloud, which is a series of simplicial complexes from the original data point cloud, with different settings.

3.3.2. Persistent homology

Based on the simplicial complex tools, complicate geometrical structures (i.e., homology) can be described, like loops and voids. The persistent homology illustrated the persistent features from the point cloud with a ϵ-ball radius for each vertex. The simplicial complex with the radius $\epsilon = 0$ is the original data point cloud. If we consider increasing the ϵ radius gradually, a simplicial complex series are generated. With the increasing of ϵ, data point pairs are connected when their Euclidean distance is less than ϵ. Consequently, a new edge born which forms a new simplicial complex structure.

For example, in Figure 4, the original point cloud is vertices without any edges with $\epsilon = 0.08$, as $\epsilon = 0.08$ increases any points with distance less than 0.08 are connected, thus some edges appeared. Keeping on increasing ϵ, finally, any two points are linked thus generated a fully connected graph ($r = 0.334$). At each stage, the homology of the simplicial complex is changing, some components born while some died. Consider the quadrilateral hole in Figure 4, it borns at an event time before $\epsilon = 0.41$ and dies at an event time after $\epsilon = 0.48$. Then we can say the hole is alive between $\epsilon = 0.41$ and $\epsilon = 0.48$. Similarly, there are much more higher-dimensional holes when we consider a more complicated point cloud in practical applications. A similar process happens for all the homology.

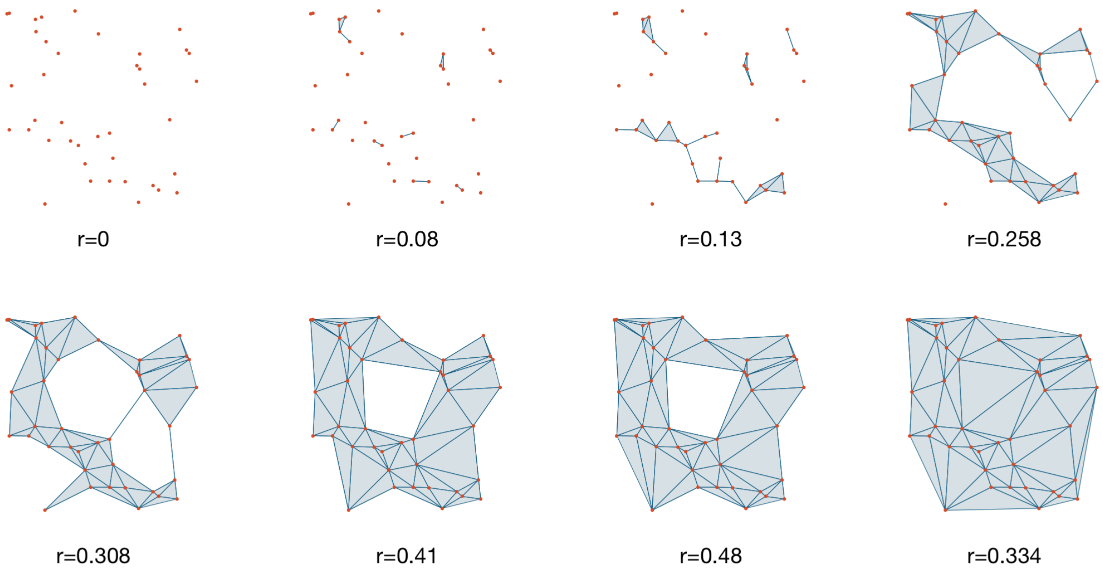

Figure 4. Persistent homology.

For each distance ϵ we build the space of S_ϵ, which is composed of points, edges, triangles and high dimensional objects. Consider the ϵ increasing process: an edge appears when the Euclidean distance between two points is less than ϵ; a triangle appears when all its edges are in the space of S_ϵ; an tetrahedron appears when all its faces are in S_ϵ; and so on. From the process we have a series of ϵ-based spaces, and for two spaces' distance parameters ϵ_1 and ϵ_2 with $\epsilon_1 \leq \epsilon_2$, we could infer that the space S_{ϵ_1} is contained in S_{ϵ_2} [66]. Then the process yields a nested sequence of spaces, the sequence is termed as filtration, consider the description in [66,67] as follows. The increasing sequence of ϵ (i.e. distance) value produces a $filtration$: given a set X, the K-simplex $\{\sigma_1, \sigma_2, \ldots, \sigma_{k+1}\}$, then we have the definition of filtration.

Definition 1. *A filtration of a (finite) simplicial complex K is a sequence of sub-complexes such that*

1. $\emptyset = K^0 \subset K^1 \subset K^2 \cdots \subset K^m = K$
2. $K^{i+1} = K^i \cup \sigma^{i+1}$ *where* σ^{i+1} *is a simplex of K*

3.4. Topological Signature: from Barcode to Persistence Landscape

The space of S_ϵ can be studied with the tool of homology from the algebraic topology; the essential features in each space are the components of topological objects and their lifetime. The lifetime can be described as barcode intervals illustrated in Figure 5. The starting point of the interval can be considered as the time when the corresponding object first appears (born time in works of literature), and the endpoint is when this object finally disappear (death time). The strict mathematical theory of persistent homology and barcode can be referred in [46,47]. In this study, based on the Barcodes, we can extract features from the data point clouds of related time series samples. The horizontal axis in the figure indicates the distance parameter of ϵ. As ϵ increases from zero to infinity, components merge gradually, and finally, only one object exists as indicated by the longest arrow bar.

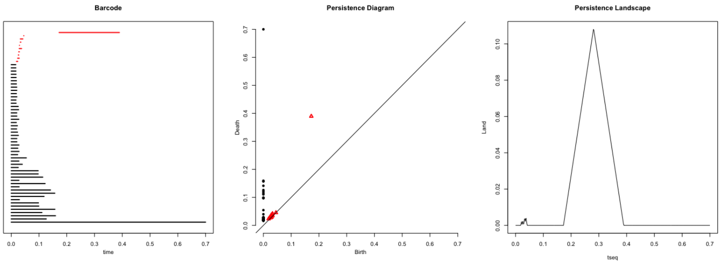

Figure 5. Barcode Illustrations.

From the perspective of machine learning applications, barcode generated with the topological method can be considered as an alternative feature sets. Some applications directly consider the barcode as an estimator in statistics or features [29], i.e. directly use the barcode intervals as representation. Further, using distance measures like the Bottleneck and Wasserstein distance for comparison the topological similarity between persistence diagrams for applications in protein binding analysis [68]. Other essential works are distance-based signal classification, [20,69]. In this work, we use the persistence landscape constructed from the Wasserstein distance description for our analysis.

Mathematically, the persistence diagram P_k encoded from the k-dimensional homology α information in all scales. The homology α was "born" at b_α and "die" at d_α which make a pair (b_α, d_α). This pair can be considered as point $z_\alpha \in \mathcal{R}$. Then the barcode graph can be transformed into a persistence diagram graph with birth indices on horizontal axis and death indices on the vertical axis as in Figure 5. The Wasserstein distance is often used as a standard metric to analysis the persistence diagram space as:

$$W_p(P_k^1, P_k^2) = inf_\phi[\sum_{q \in P_k^1} ||x - \phi(x)||_\infty^p]^{\frac{1}{p}} \tag{2}$$

The equation is termed as the p-th Wasserstein distance, when $p = \infty$ the metric is known as Bottleneck distance. Based on the Wasserstein metric, a representation termed Persistence Landscape had ben proposed for statistical analysis by [31].

For each birth-death point $(b_\alpha, d_\alpha) \in P_k$, a piecewise linear function:

$$f(b_\alpha, d_\alpha) = \begin{cases} x - b_\alpha, & if \quad x \in \left(b_\alpha, \dfrac{b_\alpha + d_\alpha}{2}\right] \\ -x + d_\alpha, & if \quad x \in \left(\dfrac{b_\alpha + d_\alpha}{2}, d_\alpha\right) \\ 0, & if \quad x \notin (b_\alpha, d_\alpha) \end{cases} \tag{3}$$

with which a sequence of functions λ can be given by:

$$\lambda(x) = k - max\{f_{(b_\alpha, d_\alpha)}(x) | (b_\alpha, d_\alpha) \in P_k\} \tag{4}$$

where the k-max denotes the k-th largest value of a function. With the persistence landscape in Banach space, statistical methods can be involved. More theory description of persistence landscape can be referred from [31].

3.5. Random Forest Classifier

With the extracted topological representations, we use the random forests classifier to perform the multi-class classification task. Random forests (RF), some times, random decision forests, are a kind of ensemble learning method for classification. It is widely used in different types of classification and regression problems. Then persistence landscape features are used as the input for the RF classifier. Random forests classifier had been widely used in the classification task. The RF classifier design is out of scope in this study, and we recommend [70] for reference of RF techniques.

3.6. Performance Evaluation

The proposed method could be evaluated as a usual classification task parameters, we calculate the four parameters in confusion matrix: the correct classification number (TP as true positive samples); the false classification number (FN as false negatives); correct normal classification number (TN true negatives); and false classification number (FP as false positives). We use the normalized confusion matrix for the performance evaluation. In the normalized confusion matrix, a row represents an instance of the class with its actual label, and a column represents the predicted class. Then the values of the diagonal elements are the correctly classified proportions. In the meantime, the off-diagonal elements are misclassified information. Then higher values in the diagonal elements mean better results.

4. Experiments

4.1. Materials and pre-processing

The systematic approach presented in Section 3.1 is applied to the long-term ECG data in the Physionet MIT-BIH Long-Term database (PhysioBank) [71]. The long-term ECG database contains six two-channel ECG signals sampled at 128 Hz per channel with 12-bit resolution, and one three-channel ECG sampled at 128 Hz per channel with 10-bit resolution. The seven long-term ECG record IDs are 14046, 14134, 14149, 14157, 14172, 14184, and 15814.

We resample the ECG signals to 340 Hz, which is a typical sampling rate in ECG algorithm studies. The baseline wander (caused by perspiration, respiration and body movements), power line interference and muscle noise removed with a Butterworth filter. Then the filtered ECG signals were segmented into individual heartbeat waveforms based on the detected R peaks. With the extracted R peaks, each 340-point signal is determined in an ad-hoc rule: set the R position as the 141 points, and extract the anterior 140 points and the following 199 points from the original time series. Then the extracted samples are illustrated as in Table 1.

	Arrhythmia Type				
User ID	N	S	V	F	Total Samples
14046	105,408	2	9,767	95	115,272
14134	38,769	29	9,836	992	49,626
14149	144,548	0	264	0	144,812
14157	83,422	244	4,369	63	88,098
14172	58,318	1,152	6,529	1	66,000
14184	78,104	39	23,383	11	101,537
15814	91,628	34	9,942	1,744	103,348

Table 1. The 7 records' detail of MIT-BIH Long Term Dataset

We can see that for each records that the data samples are severe imbalanced. So we further re-design an experiment with the following settings:

1. Data samples with a limited number are ignored, like the *F* class in ID 14172, and *V* in ID 15814;
2. As some samples are ignored, the classification task in ID 14184, and ID 14046 become a two-class classification task.

Then we only focus on the 3-class classification task in ID 15814 ID 14134, and ID 14172. For each task, we random select the correspond signal samples to make the signal sample distribution balanced, and also for the reduction of computation consumption. As Table 2 illustrated, the distributions for the signal samples are 1:2:4 for each task. While for a small sample validation, we designed an extra experiments with 100 samples from each class from ID 14172.

The software packages and tools used in this work are listed in the Appendix A.

		Arrhythmia Type			
Experiment Number	User ID	N	S	V	F
Exp #1	14134	3,968	0	1,984	992
Exp #2	15814	6,976	0	3,488	1,744
Exp #3	14172	4,608	1,152	2,304	0
Exp #4	14172	100	100	100	0

Table 2. 4 designed experiments from the MIT-BIH Long-term Database

4.2. *Time delay embedding*

For the abstract phase space construction, we use a time delay embedding method to convert the time series into data point cloud. The time delay embedding parameters are the embedding dimension and time lag, here we use an empirical setting as time lag $\tau = 4$ and embedded dimension $d = 2$.

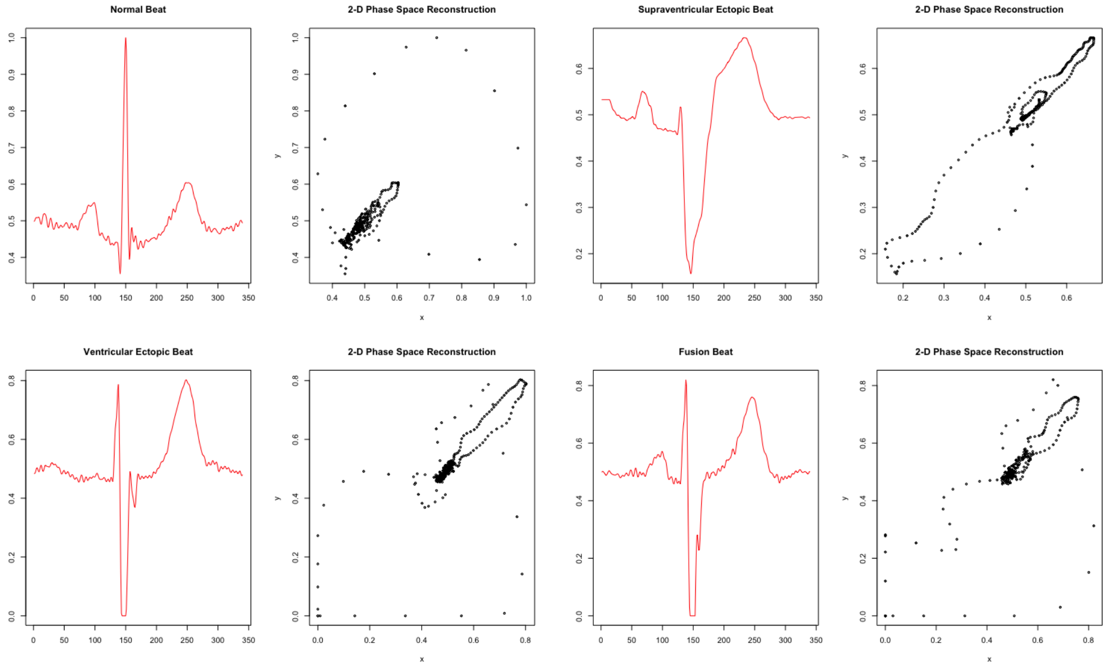

Figure 6. Time Delay Embedding for the arrhythmia types.

As Figure 6 illustrated, for different heartbeat types, the 2-dimensional time delay embedding has quite different data point cloud.

4.3. *Point cloud subsampling*

With the embedded point clouds, we try to downsample the point numbers for computational efficient. Here we consider to downsample the point cloud into 50 points. For downsampling strategy, we choose the landmark selection algorithm [72], which could keep the geometry similarity, i.e. the algorithm can still track the original shape of the point cloud. As Figure 7 illustrated, the subsampled data point clouds share the sample shape from the original embedded point cloud.

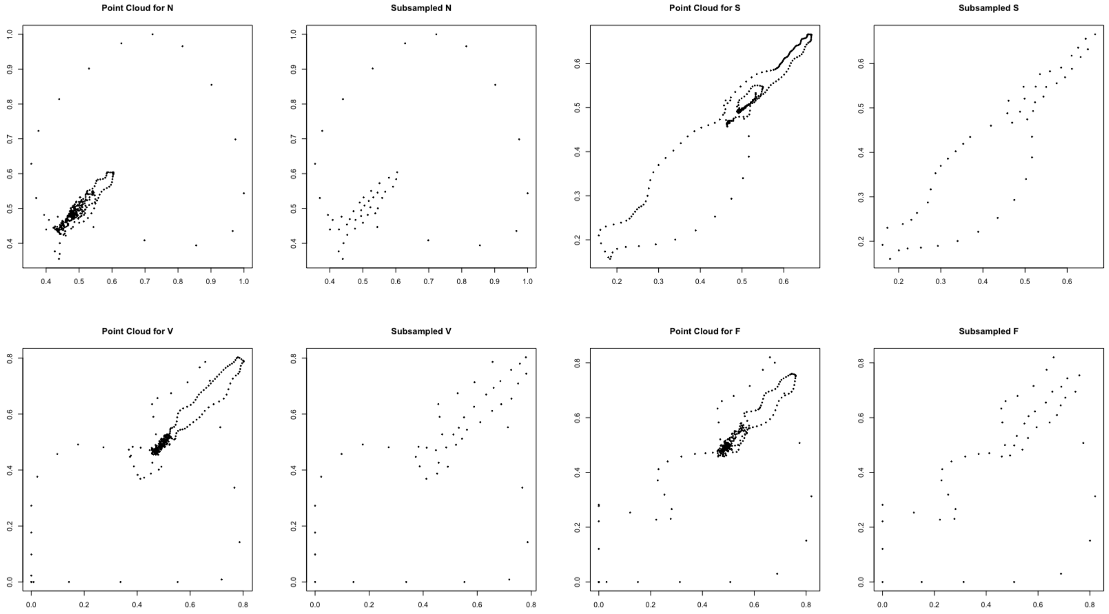

Figure 7. The subsampling of data point clouds.

4.4. Topological representations for point clouds

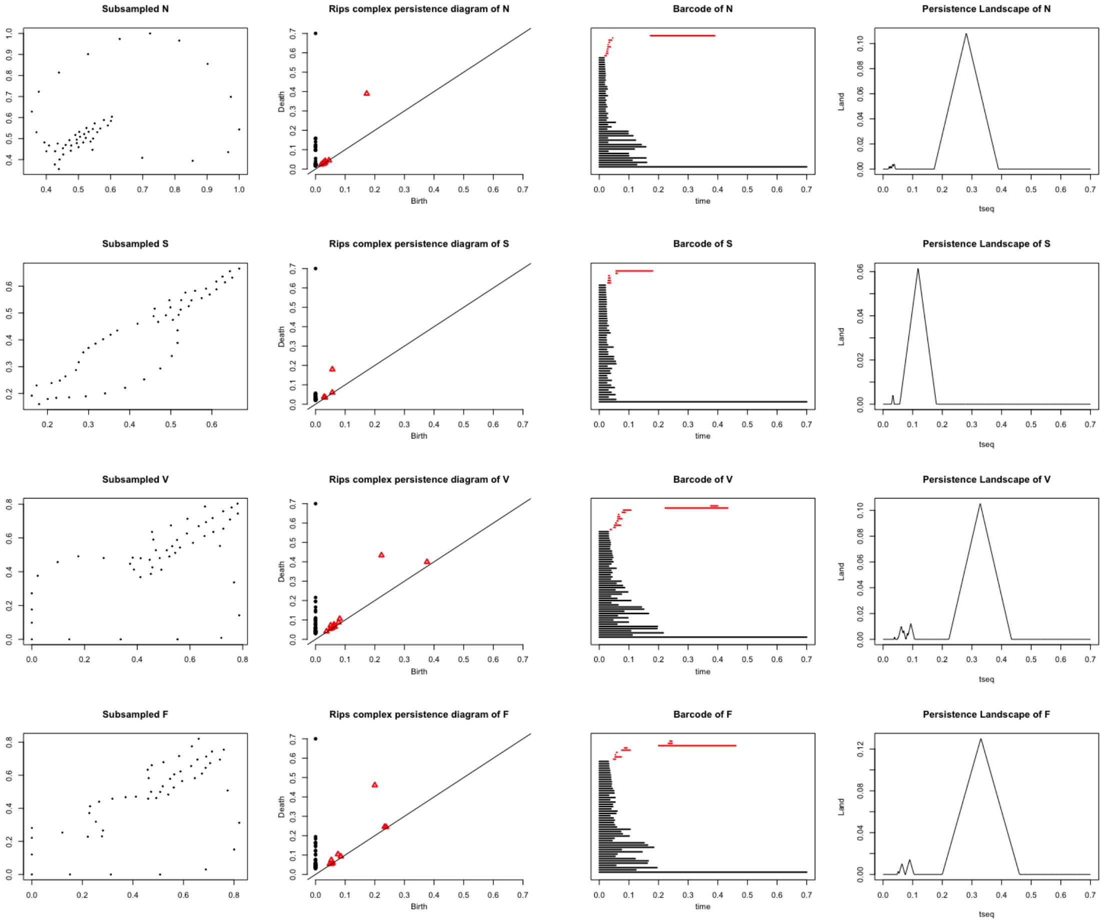

Figure 8. The topological signatures for each class

As Figure 8 illustrated, for each arrhythmia type, using the subsampled point cloud we generate the corresponding persistence diagram using the Rips complex, we can see that for each class the persistence diagram, Barcode, and landscape distributed differently. For the first column, the data point clouds are subsampled, which is only illustrations for 50 points. We can see that the sketches are different. As the radius of each point increase, the homology of the point cloud start to born and later die gradually. This birth-death process can be illustrated in the third column as Barcodes. Each bar in the Barcode graph means a homology of the point cloud. The dark bars are the 0-homology and red for 1-homology. In the second column, the persistence diagrams are alternative representations of Barcode. In the fourth column, the persistence landscapes are the features we extracted for the classification. For each ECG class, we can see that the distribution for each class are different, with which the random forest classifiers are performed.

4.5. Experiment Results

With the experiments settings from Section 4.1, we evaluated the features with the proposed framework with the random forest classifier. We can see that in Figure 8, the persistence landscapes for each class are distributed differently. With the landscapes for each data sample, we set the experiments with a testing partition size into 20%, 40%, 60%, and 80%. The rest of the data samples are treat as testing set. The performance are evaluated with classification confusion matrixes, which are illustrated in Figure 9.

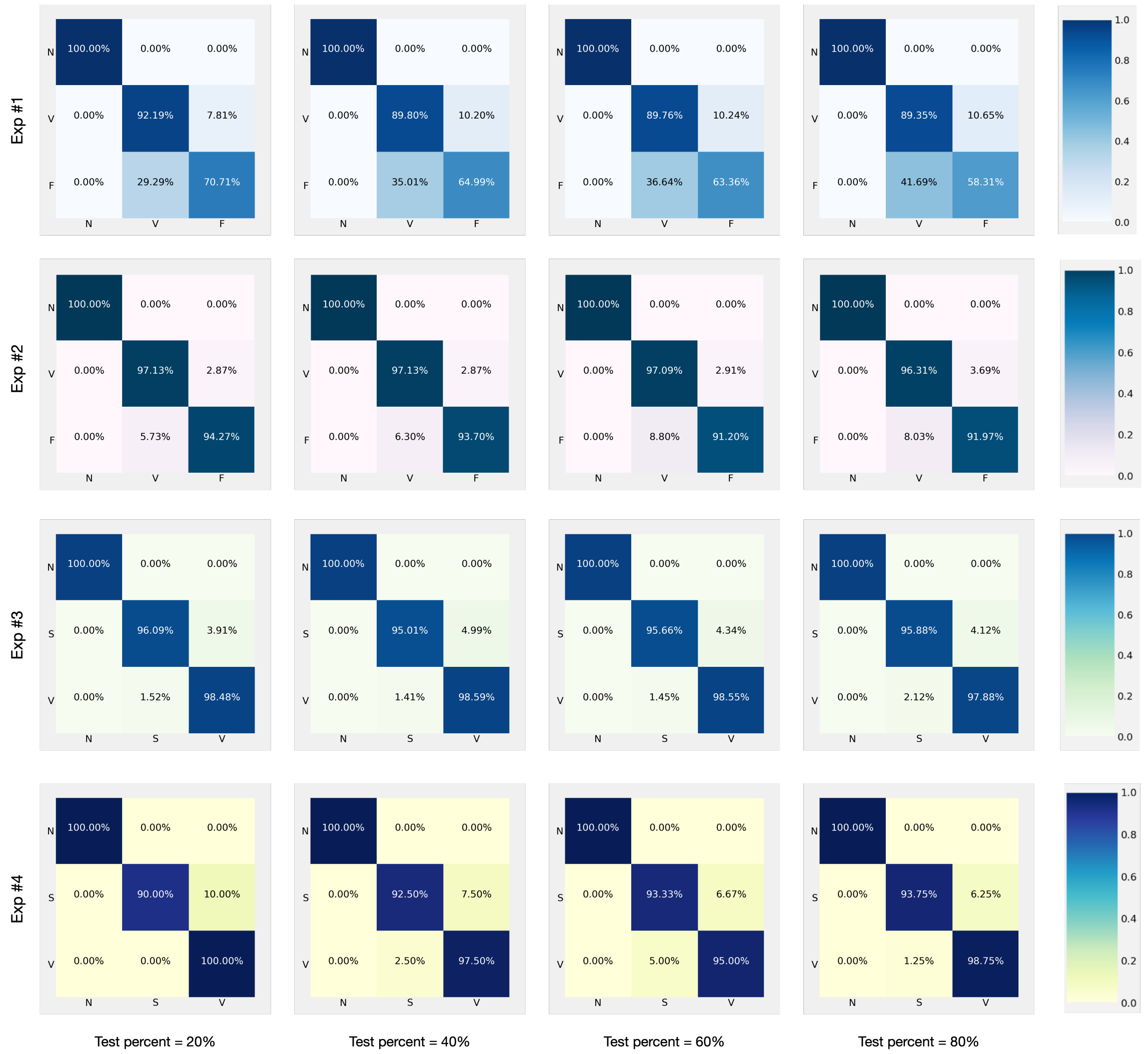

Figure 9. The topological signatures for each class

The experiment results of Exp #1 are shown in the first row of Figure 9. In Exp #1, a 3-class classification task for N, V, F was evaluated with the proposed method. We can see that the V class was well classified, even with only 20% percent of samples used in the training process, the test result can achieve 89.35%. The results for the F class recognition, the performance dropped rapidly as the training sample reduced from 70.71% to 58.31%. From Table 2 we can see that in Exp #1, there are only 992 F samples for evaluation. When the training size is set into 20%, with only about 200 training samples, the recognition rate still can achieve 58.31%.

In the second row of Figure 9, for Exp #2 the N, V, F oriented 3-class classification task was evaluated. We can see with 20% percents for the training set, the recognition rates of V and F class achieve 96.31% and 91.97% respectively. In the third row for Exp #3, the N, S, V oriented 3-class classification task was evaluated, and the recognition rates of V and F are 95.88% and 97.88% respectively when training size set into 20%. The fourth row illustrated the results of Exp #4, and we use a small data size; the results show excellent performance as well.

5. Related Work and Discussion

In this paper, a new personalized ECG beat classification and arrhythmia analysis scheme have been introduced using persistence landscape from algebra topology field. To our knowledge, this is the first work ever the TDA representation persistence landscape applied to the personalized arrhythmia analysis. The scheme has been proved as an efficient representation for ECG analysis, especially when using in small training size, the technique shows great potential. Firstly Employing this scheme include a time-delay embedding technique to convert signals into points that illustrate the underlying signal system. We offer a 2-dimensional embedding from the signals, and a time lag parameter was set to 4, the reconstructed phase space for each class via time delay embeddings has been illustrated in Figure 6. Secondly, to reduce computational consumption, a sub-sampling technique has been used to reduce the sample number, the technique function can be referred to in Figure 7. Thirdly, different topological signatures have been extracted from the sub-sampled data point clouds, which include Barcode, persistence diagram, and persistence landscape. Finally, the persistence landscapes of the signals are considered as signal representations, then used in a classification task.

Compare to the previous work in ECG classification or arrhythmia analysis. We have the following different settings:

1. In lots of ECG analysis tasks, they validate the method using a mixture of different patients. However, in this work, we focus more on personalized settings. The reason for this setting is because, for different individuals, the cardiac system could differ in the physiological system. The settings could bring a sample deviation because of the dynamics difference, which departures the purpose of this study.
2. The topological method was used in this study to extract representations differ from the statistical features' work. This study is more related to the nonlinear signal analysis, which uses TDA instead of other nonlinear parameters.
3. There is a limited study on the small training set experiments; the proposed experiments show that the proposed method could achieve good performance for arrhythmia analysis task. This specialty could be meaningful for the wearable or clinical applications when dealing with a high prediction performance under limited data source condition.

However, in Exp #1, we find that the fusion type heartbeat could be considered as the ventricular heartbeat type when using a rather small training size. The possible explanation for this phenomena is that the fusion heartbeat is a mixture of a ventricular and normal beat; the physiological process could be similar to the ventricular physiological process. A fusion beat occurs when electrical impulses from different sources act upon the same region of the heart at the same time[73]. So the classifier needs more training samples to learning this type.

Since the TDA method could be time-consuming, we use different approximation techniques like using only 2-dimension as the phase space, choose Rips complex to build the simplicial complex, and adopt the landmark selection method to reduce the point cloud amount. These kinds of approximation may cause an information loss for the classification task. However, the proposed methodology could still be useful in some case as an alternative option to statistic-based features, or recently widely used deep learning methods.

6. Conclusions

The TDA methods provide alternative insights compare to the statistical features. We show that a TDA-based signal feature, namely persistence landscape, can provide useful representation for personalized arrhythmia analysis. The proposed method shows excellent performance when dealing with personalized arrhythmia analysis in our experiments. We also find that the proposed method is robust to the size of the training set. When dealing with only 20% of the full training dataset, it achieves a 100% accuracy for normal heartbeats, 97.13% for ventricular beats, 94.27% for supra-ventricular beats, and 94.27% for fusion beats. Thus the method can be trained for a single individual, allowing for personalized analysis systems. With the present study, we show that TDA could be a useful tool for biomedical signal analysis, with potentially promising application in the personalized data processing.

Author Contributions: Conceptualization, Yan Yan. and Y.Y.; methodology, Yan Yan; software, Yan Yan; validation, Yan Yan; formal analysis, Yan Yan; investigation, Yan Yan; resources, Yan Yan; data curation, Yan Yan; writing–original draft preparation, Yan Yan; writing–review and editing, Yan Yan and Kamen Ivanov; visualization, Yan Yan; supervision, Jian Cen and Lei Wang; project administration, Qiu-Hua Liu and Lei Wang; funding acquisition, Yan Yan and Lei Wang.

Funding: This research was funded by the National Natural Science Foundation of China grant number 71531004, Shenzhen Innovation Project KQJSCX20170731163142116, and the Enhancement Project for Shenzhen Biomedical Electronics Technology Public Service Platform.

Acknowledgments: The author would like to thank the effort for the anonymous reviewers.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

TDA	Topological Data Analysis
ECG	Electrocardiagraphy
PL	Persistence Landscape
N	Normal Heartbeat
S	Supra-ventricular Heartbeat
V	Ventricular Heartbeat
F	Fusion Beat
RF	Random Forests

Appendix A Software and tools used

1. ECG data sources: https://www.physionet.org/content/ltafdb/
2. Data pre-processing: can be accessed from https://github.com/tygrin/sensorsPLECG.git
3. Time delay embedding: the R package "nonlinearTseries"
4. Point cloud downsampling: javaplex tutorial from [74]
5. Topological data analysis: the R package "TDA"
6. Classification: the Python package “sklearn”

References

1. Go, A.S.; Hylek, E.M.; Phillips, K.A.; Chang, Y.; Henault, L.E.; Selby, J.V.; Singer, D.E. Prevalence of diagnosed atrial fibrillation in adults: national implications for rhythm management and stroke prevention: the AnTicoagulation and Risk Factors in Atrial Fibrillation (ATRIA) Study. *Jama* **2001**, *285*, 2370–2375.
2. Baig, M.M.; Gholamhosseini, H.; Connolly, M.J. A comprehensive survey of wearable and wireless ECG monitoring systems for older adults. *Medical & biological engineering & computing* **2013**, *51*, 485–495.

3. Glaros, C.; Fotiadis, D.; Likas, A.; Stafylopatis, A. A wearable intelligent system for monitoring health condition and rehabilitation of running athletes. 4th International IEEE EMBS Special Topic Conference on Information Technology Applications in Biomedicine, 2003. IEEE, 2003, pp. 276–279.
4. Lee, J.; Reyes, B.A.; McManus, D.D.; Maitas, O.; Chon, K.H. Atrial fibrillation detection using an iPhone 4S. *IEEE Transactions on Biomedical Engineering* **2012**, *60*, 203–206.
5. Chan, P.H.; Wong, C.K.; Poh, Y.C.; Pun, L.; Leung, W.W.C.; Wong, Y.F.; Wong, M.M.Y.; Poh, M.Z.; Chu, D.W.S.; Siu, C.W. Diagnostic performance of a smartphone-based photoplethysmographic application for atrial fibrillation screening in a primary care setting. *Journal of the American Heart Association* **2016**, *5*, e003428.
6. Richter, M.; Schreiber, T. Phase space embedding of electrocardiograms. *Physical Review E* **1998**, *58*, 6392.
7. Kennel, M.B.; Brown, R.; Abarbanel, H.D. Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Physical review A* **1992**, *45*, 3403.
8. Frank, J.; Mannor, S.; Precup, D. Activity and gait recognition with time-delay embeddings. Twenty-Fourth AAAI Conference on Artificial Intelligence, 2010.
9. Voss, H.U. Real-time anticipation of chaotic states of an electronic circuit. *International Journal of Bifurcation and Chaos* **2002**, *12*, 1619–1625.
10. Han, M.; Xi, J.; Xu, S.; Yin, F.L. Prediction of chaotic time series based on the recurrent predictor neural network. *IEEE transactions on signal processing* **2004**, *52*, 3409–3416.
11. Xu, B.; Jacquir, S.; Laurent, G.; Bilbault, J.M.; Binczak, S. Phase space reconstruction of an experimental model of cardiac field potential in normal and arrhythmic conditions. 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2013, pp. 3274–3277.
12. Perc, M. The dynamics of human gait. *European journal of physics* **2005**, *26*, 525.
13. Ali, S.; Basharat, A.; Shah, M. Chaotic invariants for human action recognition. 2007 IEEE 11th International Conference on Computer Vision. IEEE, 2007, pp. 1–8.
14. Valenza, G.; Allegrini, P.; Lanatà, A.; Scilingo, E.P. Dominant Lyapunov exponent and approximate entropy in heart rate variability during emotional visual elicitation. *Frontiers in neuroengineering* **2012**, *5*, 3.
15. Bolea, J.; Laguna, P.; Remartínez, J.M.; Rovira, E.; Navarro, A.; Bailón, R. Methodological framework for estimating the correlation dimension in HRV signals. *Computational and mathematical methods in medicine* **2014**, *2014*.
16. Chen, C.K.; Lin, C.L.; Lin, S.L.; Chiu, Y.M.; Chiang, C.T. A chaotic theoretical approach to ECG-based identity recognition [application notes]. *IEEE Computational Intelligence Magazine* **2014**, *9*, 53–63.
17. Desai, U.; Martis, R.J.; Acharya, U.R.; Nayak, C.G.; Seshikala, G.; SHETTY K, R. Diagnosis of multiclass tachycardia beats using recurrence quantification analysis and ensemble classifiers. *Journal of Mechanics in Medicine and Biology* **2016**, *16*, 1640005.
18. Di Marco, L.Y.; Raine, D.; Bourke, J.P.; Langley, P. Recurring patterns of atrial fibrillation in surface ECG predict restoration of sinus rhythm by catheter ablation. *Computers in biology and medicine* **2014**, *54*, 172–179.
19. Gong, Y.; Lu, Y.; Zhang, L.; Zhang, H.; Li, Y. Predict defibrillation outcome using stepping increment of poincare plot for out-of-hospital ventricular fibrillation cardiac arrest. *BioMed research international* **2015**, *2015*.
20. Seversky, L.M.; Davis, S.; Berger, M. On time-series topological data analysis: New data and opportunities. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016, pp. 59–67.
21. Duponchel, L. Exploring hyperspectral imaging data sets with topological data analysis. *Analytica chimica acta* **2018**, *1000*, 123–131.
22. Li, L.; Cheng, W.Y.; Glicksberg, B.S.; Gottesman, O.; Tamler, R.; Chen, R.; Bottinger, E.P.; Dudley, J.T. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Science translational medicine* **2015**, *7*, 311ra174–311ra174.
23. Lum, P.Y.; Singh, G.; Lehman, A.; Ishkanov, T.; Vejdemo-Johansson, M.; Alagappan, M.; Carlsson, J.; Carlsson, G. Extracting insights from the shape of complex data using topology. *Scientific reports* **2013**, *3*, 1236.
24. Muszynski, G.; Kurlin, V.; Kashinath, K.; Wehner, M.; Prabhat, M. Topological Data Analysis and Machine Learning for Classifying Atmospheric River Patterns in Large Climate Datasets. EGU General Assembly Conference Abstracts, 2018, Vol. 20, p. 10825.

25. Lamar-León, J.; Garcia-Reyes, E.B.; Gonzalez-Diaz, R. Human gait identification using persistent homology. Iberoamerican Congress on Pattern Recognition. Springer, 2012, pp. 244–251.
26. Dirafzoon, A.; Lokare, N.; Lobaton, E. Action classification from motion capture data using topological data analysis. 2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP). IEEE, 2016, pp. 1260–1264.
27. Emrani, S.; Gentimis, T.; Krim, H. Persistent homology of delay embeddings and its application to wheeze detection. *IEEE Signal Processing Letters* **2014**, *21*, 459–463.
28. Zhang, Z.; Song, Y.; Cui, H.; Wu, J.; Schwartz, F.; Qi, H. Topological analysis and Gaussian decision tree: Effective representation and classification of biosignals of small sample size. *IEEE Transactions on Biomedical Engineering* **2016**, *64*, 2288–2299.
29. Safarbali, B.; Golpayegani, S.M.R.H. Nonlinear dynamic approaches to identify atrial fibrillation progression based on topological methods. *Biomedical Signal Processing and Control* **2019**, *53*, 101563.
30. Dindin, M.; Umeda, Y.; Chazal, F. Topological Data Analysis for Arrhythmia Detection through Modular Neural Networks. *arXiv preprint arXiv:1906.05795* **2019**.
31. Bubenik, P. Statistical topological data analysis using persistence landscapes. *The Journal of Machine Learning Research* **2015**, *16*, 77–102.
32. Bubenik, P.; Dłotko, P. A persistence landscapes toolbox for topological statistics. *Journal of Symbolic Computation* **2017**, *78*, 91–114.
33. Bubenik, P. The persistence landscape and some of its properties. *arXiv preprint arXiv:1810.04963* **2018**.
34. Christov, I.; Gómez-Herrero, G.; Krasteva, V.; Jekova, I.; Gotchev, A.; Egiazarian, K. Comparative study of morphological and time-frequency ECG descriptors for heartbeat classification. *Medical engineering & physics* **2006**, *28*, 876–887.
35. Ye, C.; Kumar, B.V.; Coimbra, M.T. Heartbeat classification using morphological and dynamic features of ECG signals. *IEEE Transactions on Biomedical Engineering* **2012**, *59*, 2930–2941.
36. Banerjee, S.; Mitra, M. Application of cross wavelet transform for ECG pattern analysis and classification. *IEEE transactions on instrumentation and measurement* **2013**, *63*, 326–333.
37. Stamkopoulos, T.; Diamantaras, K.; Maglaveras, N.; Strintzis, M. ECG analysis using nonlinear PCA neural networks for ischemia detection. *IEEE Transactions on Signal Processing* **1998**, *46*, 3058–3067.
38. He, R.; Wang, K.; Zhao, N.; Liu, Y.; Yuan, Y.; Li, Q.; Zhang, H. Automatic detection of atrial fibrillation based on continuous wavelet transform and 2d convolutional neural networks. *Frontiers in physiology* **2018**, *9*, 1206.
39. Yan, Y.; Qin, X.; Wu, Y.; Zhang, N.; Fan, J.; Wang, L. A restricted Boltzmann machine based two-lead electrocardiography classification. BSN, 2015, pp. 1–9.
40. Owis, M.I.; Abou-Zied, A.H.; Youssef, A.B.; Kadah, Y.M. Study of features based on nonlinear dynamical modeling in ECG arrhythmia detection and classification. *IEEE transactions on Biomedical Engineering* **2002**, *49*, 733–736.
41. Nayak, S.K.; Bit, A.; Dey, A.; Mohapatra, B.; Pal, K. A review on the nonlinear dynamical system analysis of electrocardiogram signal. *Journal of healthcare engineering* **2018**, *2018*.
42. Sezgin, N. Nonlinear analysis of electrocardiography signals for atrial fibrillation. *The Scientific World Journal* **2013**, *2013*.
43. Henry, B.; Lovell, N.; Camacho, F. Nonlinear dynamics time series analysis. *Nonlinear biomedical signal processing: Dynamic analysis and modeling* **2001**, *2*, 1–39.
44. Takens, F. Detecting strange attractors in turbulence. In *Dynamical systems and turbulence, Warwick 1980*; Springer, 1981; pp. 366–381.
45. Jaeger, J.; Monk, N. Bioattractors: dynamical systems theory and the evolution of regulatory processes. *The Journal of physiology* **2014**, *592*, 2267–2281.
46. Carlsson, G. Topology and data. *Bulletin of the American Mathematical Society* **2009**, *46*, 255–308.
47. Ghrist, R. Barcodes: the persistent topology of data. *Bulletin of the American Mathematical Society* **2008**, *45*, 61–75.
48. Cohen-Steiner, D.; Edelsbrunner, H.; Harer, J. Stability of persistence diagrams. *Discrete & Computational Geometry* **2007**, *37*, 103–120.
49. Fasy, B.T.; Lecci, F.; Rinaldo, A.; Wasserman, L.; Balakrishnan, S.; Singh, A.; others. Confidence sets for persistence diagrams. *The Annals of Statistics* **2014**, *42*, 2301–2339.

50. Carriere, M.; Cuturi, M.; Oudot, S. Sliced wasserstein kernel for persistence diagrams. Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017, pp. 664–673.
51. Zieliński, B.; Lipiński, M.; Juda, M.; Zeppelzauer, M.; Dłotko, P. Persistence Bag-of-Words for Topological Data Analysis. *arXiv preprint arXiv:1812.09245* **2018**.
52. Perea, J.A.; Harer, J. Sliding windows and persistence: An application of topological methods to signal analysis. *Foundations of Computational Mathematics* **2015**, *15*, 799–838.
53. Gidea, M.; Katz, Y. Topological data analysis of financial time series: Landscapes of crashes. *Physica A: Statistical Mechanics and its Applications* **2018**, *491*, 820–834.
54. Umeda, Y. Time series classification via topological data analysis. *Information and Media Technologies* **2017**, *12*, 228–239.
55. Bradley, E.; Kantz, H. Nonlinear time-series analysis revisited. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **2015**, *25*, 097610.
56. Kliková, B.; Raidl, A. Reconstruction of phase space of dynamical systems using method of time delay. Proceedings of WDS, 2011, Vol. 11, pp. 83–87.
57. Kantz, H.; Schreiber, T. *Nonlinear time series analysis*; Vol. 7, Cambridge university press, 2004.
58. Kerber, M.; Sharathkumar, R. Approximate Čech complex in low and high dimensions. International Symposium on Algorithms and Computation. Springer, 2013, pp. 666–676.
59. Edelsbrunner, H.; Kirkpatrick, D.; Seidel, R. On the shape of a set of points in the plane. *IEEE Transactions on information theory* **1983**, *29*, 551–559.
60. Edelsbrunner, H.; Harer, J. *Computational topology: an introduction*; American Mathematical Soc., 2010.
61. Vietoris, L. Über den höheren Zusammenhang kompakter Räume und eine Klasse von zusammenhangstreuen Abbildungen. *Mathematische Annalen* **1927**, *97*, 454–472.
62. Dey, T.K.; Fan, F.; Wang, Y. Graph induced complex on point data. Proceedings of the twenty-ninth annual symposium on Computational geometry. ACM, 2013, pp. 107–116.
63. Osting, B.; Palande, S.; Wang, B. Spectral Sparsification of Simplicial Complexes for Clustering and Label Propagation. *arXiv preprint arXiv:1708.08436* **2017**.
64. Munch, E. A user's guide to topological data analysis. *Journal of Learning Analytics* **2017**, *4*, 47–61.
65. Marchese, A.; Maroulas, V.; Mike, J. K- means clustering on the space of persistence diagrams. Wavelets and Sparsity XVII. International Society for Optics and Photonics, 2017, Vol. 10394, p. 103940W.
66. Otter, N.; Porter, M.A.; Tillmann, U.; Grindrod, P.; Harrington, H.A. A roadmap for the computation of persistent homology. *EPJ Data Science* **2017**, *6*, 17.
67. Zomorodian, A.J. *Topology for computing*; Vol. 16, Cambridge university press, 2005.
68. Kovacev-Nikolic, V.; Bubenik, P.; Nikolić, D.; Heo, G. Using persistent homology and dynamical distances to analyze protein binding. *Statistical applications in genetics and molecular biology* **2016**, *15*, 19–38.
69. Marchese, A.; Maroulas, V. Signal classification with a point process distance on the space of persistence diagrams. *Advances in Data Analysis and Classification* **2018**, *12*, 657–682.
70. Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5–32. doi:10.1023/A:1010933404324.
71. Goldberger, A.L.; Amaral, L.A.; Glass, L.; Hausdorff, J.M.; Ivanov, P.C.; Mark, R.G.; Mietus, J.E.; Moody, G.B.; Peng, C.K.; Stanley, H.E. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* **2000**, *101*, e215–e220.
72. Carlsson, G.; Ishkhanov, T.; De Silva, V.; Zomorodian, A. On the local behavior of spaces of natural images. *International journal of computer vision* **2008**, *76*, 1–12.
73. Conover, M.B. *Understanding electrocardiography*; Elsevier Health Sciences, 2002.
74. Tausz, A.; Vejdemo-Johansson, M.; Adams, H. JavaPlex: A research software package for persistent (co)homology. Proceedings of ICMS 2014; Hong, H.; Yap, C., Eds., 2014, Lecture Notes in Computer Science 8592, pp. 129–136. Software available at http://appliedtopology.github.io/javaplex/.

Sample Availability: The code for data preprocessing and segmentation are available from .