Data Science 7 (2024) 29–53
DOI 10.3233/DS-240061
IOS Press

Electrocardiogram arrhythmia detection with novel signal processing and persistent homology-derived predictors

Hunter Dlugas [*]
Department of Mathematics, Wayne State University, MI, USA
E-mail: fy7392@wayne.edu; ORCID: https://orcid.org/0000-0002-6819-0045

Editor: Karin Verspoor (https://orcid.org/0000-0002-8661-1544)
Solicited reviews: Francesca D. Faraci (https://orcid.org/0000-0002-8720-1256); one anonymous reviewer

Received 29 December 2023
Accepted 22 April 2024

Abstract. Many approaches to computer-aided electrocardiogram (ECG) arrhythmia detection have been performed, several of which combine persistent homology and machine learning. We present a novel ECG signal processing pipeline and method of constructing predictor variables for use in statistical models. Specifically, we introduce an isoelectric baseline to yield non-trivial topological features corresponding to the P, Q, S, and T-waves (if they exist) and utilize the N-most persistent 1-dimensional homological features and their corresponding area-minimal cycle representatives to construct predictor variables derived from the persistent homology of the ECG signal for some choice of N. The binary classification of (1) Atrial Fibrillation vs. Non-Atrial Fibrillation, (2) Arrhythmia vs. Normal Sinus Rhythm, and (3) Arrhythmias with Morphological Changes vs. Sinus Rhythm with Bradycardia and Tachycardia Treated as Non-Arrhythmia was performed using Logistic Regression, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Naive Bayes, Random Forest, Gradient Boosted Decision Tree, K-Nearest Neighbors, and Support Vector Machine with a linear, radial, and polynomial kernel Models with stratified 5-fold cross validation. The Gradient Boosted Decision Tree Model attained the best results with a mean F1-score and mean Accuracy of (0.967, 0.946), (0.839, 0.946), and (0.943, 0.921) across the five folds for binary classifications of (1), (2), and (3), respectively.

Keywords: Arrhythmia classification, electrocardiogram, persistent homology, topological data analysis, signal analysis

1. Introduction

Cardiovascular diseases are among the leading causes of death per the World Health Organization and the Centers for Disease Control and Prevention [8,64]. Arrhythmias are heart rhythms other than normal sinus rhythm with a heart rate between 60 beats/minute and 100 beats/minute; that is, arrhythmias are heart rhythms that are either too fast, too slow, abnormal, and/or irregular. Most arrhythmias must be treated since they can either lead to 1) more chaotic electrical activity of cardiac muscle resulting in loss of cardiac output and/or 2) the formation of thromboemboli (e.g. as in atrial fibrillation) possibly

[*] Corresponding author. E-mail: fy7392@wayne.edu.

resulting in stroke [40]. The overall prevalence of arrhythmias among adults is estimated to be around 2% with atrial fibrillation being among the most common arrhythmias [13,30]. The global prevalence of atrial fibrillation has been estimated to be about 0.51% [35].

The contraction and relaxation of cardiac muscle cells is driven by ion movement across cell membranes and must be coordinated in order for the heart to pump blood effectively. This ion movement is governed by an electrochemical potential comprised of 1) ion concentration gradients and 2) electric potentials. The depolarization and subsequent repolarization of cardiac muscle cells causes changes in electric potential on the body surface which can be measured non-invasively using an electrocardiogram (ECG). ECG analysis is important for accurate diagnosis, treatment, and prevention of cardiovascular diseases.

Topological data analysis (TDA) refers to a collection of methods concerned with quantifying 'shapes' of data which are invariant under continuous deformations such as stretching and twisting. The main tool of TDA is persistent homology which quantifies the homology of structures within the data which *persist* over a range of scales. Persistent homology has been applied to many tasks across various fields such as electroencephalogram analysis [3], genomics [4,7,14,37,44,50,57,63], classifying skin lesions based on images [10], and tumor segmentation on histology slides [49]. Cycle representatives – which will be described in Section 1.1 – of topological features have shown utility in various fields outside of ECG analysis such as analyzing structures on the atomic scale [47] and in structural engineering [24].

Several approaches to computer-aided ECG rhythm classification have been performed, including neural networks [5,15,17,21,25,39,46,48,51,56,61,62,66–68], wavelet transformation and independent component analysis [31,65], using higher-order statistics of wavelet-packet decomposition coefficients as features [32], and support vector machines using projected and dynamic ECG features [9]. An overview of TDA applied to cardiovascular signals has recently been performed [23]. In the field of computer-aided ECG analysis, TDA has been used to construct metrics of heart rate variability [11,20]. Additionally, the Mapper algorithm has been applied to predict the presence and severity of heart disease [2]. Computer-aided ECG rhythm classification methods which utilize TDA include neural networks with topological-based features [16,53], fractal dimension in tandem with neural networks [55], mapping ECG signals to a higher dimensional space prior to computing topological features [26,27,34,36,41], and utilizing a sliding window and Fast Fourier Transform to process the ECG signal prior to computing topological features [43]. These approaches construct topological predictor variables utilizing information directly derived from the birth and death radii statistics along with extra information such as heart rate, fractal dimension statistics, and persistent entropy.

To the author's knowledge, constructing predictor variables for use in machine learning models to classify ECG rhythms based off information derived from cycle representatives has not yet been performed. Additionally, to our knowledge, there has been no computer-aided ECG analysis which utilizes only the N-most persistent topological features for use in rhythm classification, nor has there been an approach which introduces an isoelectric baseline into ECG signals to yield non-trivial topological features corresponding to P, Q, S, and T-waves (if they are present to begin with). Introducing an isoelectric baseline prior to computing persistent homology and utilizing the N-most persistent topological features and properties of their area-minimal cycle representatives for use in constructing predictor variables makes the approach taken here distinct from other combinations of TDA and machine learning described in the literature.

In Section 1.1, we give a brief overview of the aspects of persistent homology utilized in this study. Appendix A formalizes the intuition underlying persistent homology described in Section 1.1. The Methods portion is split into three parts: Section 2.1 describes the novel ECG processing pipeline, Section 2.2

describes the construction of predictor variables primarily based off the topological features of the processed ECG signal, and Section 2.3 describes the specific classification tasks along with the statistical models and evaluation metrics used. The Results/Discussion section presents the evaluation metrics and ROC curves for each statistical model used. The Conclusion section contains a brief comparison between the method proposed here and other methods which use TDA and machine learning for rhythm classification in addition to describing some future directions.

1.1. *Intuition behind persistent homology*

The background on persistent homology presented both here and in Appendix A is restricted to two-dimensional data and one-dimensional homology features. The methods discussed generalize to higher dimensions, but we restrict our focus to the relevant dimensions used in the ECG analysis presented here. A toy example dataset X and its persistent homology are used to build some intuition for persistent homology. The informal treatment of persistent homology described in this section is made rigorous in Appendix A.

Consider the set of points in the plane $\mathbb{R}^2$ shown in Fig. 1. Consider drawing a circle around each point, each with the same radius r. We will refer to the union of these circles as the Geometric Čech Complex of radius r, denoted $\check{C}_r(X)$, not to be confused with the Čech complex of radius r, which commonly refers to an abstract simplicial complex. Observe that for $r < 0.57$, none of the circles comprising $\check{C}_r(X)$ overlap around a "void" of non-overlapping space. Furthermore, observe that for $r \in [0.57, 0.81)$, the circles comprising the smaller loop of points nearby the point $(1, 1)$ overlap such that there is a "void" of non-overlapping space enclosed by their region of overlap. Hence for $r \in [0.57, 0.81)$, there exists a non-contractible loop within $\check{C}_r(X)$. "Non-contractible" here means that the loop drawn around the

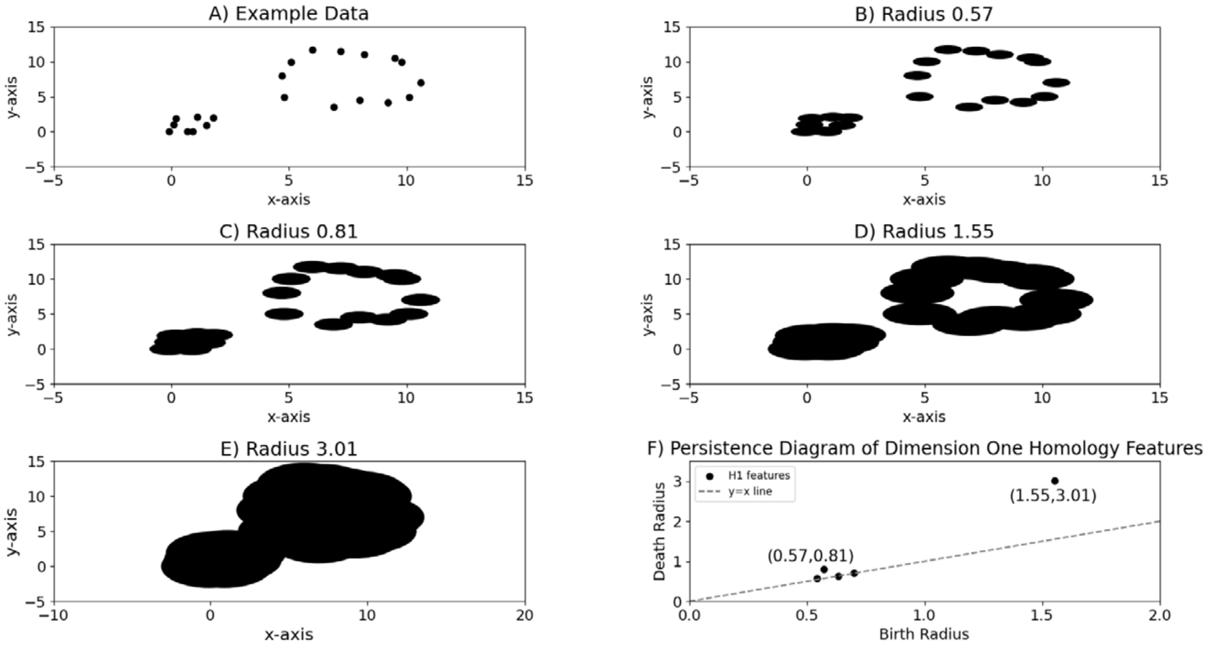

Fig. 1. **Example dataset with persistence diagram.** A: example dataset; B–E: radius 0.57, 0.81, 1.55, and 3.01 Geometric Čech Complex depicted in black, respectively; F: persistence diagram of equivalence classes of non-contractible loops.

void of non-overlapping space cannot be continuously deformed down to a single point without leaving $\breve{C}_r(X)$; that is, the loop gets "stuck" on the void encircled by $\breve{C}_r(X)$. This non-contractible loop can be continuously deformed to construct another non-contractible loop "stuck" around the same void. These two non-contractible loops are *homotopic* to one another. For example, the green and red loops in Fig. 1 are homotopic. The set of all possible non-contractible loops "stuck" around some void encircled by $\breve{C}_r(X)$ forms an equivalence class of non-contractible loops, i.e. a set of non-contractible loops where any two non-contractible loops in the set are homotopic. In practice, rather than homotopy – of which is relatively straightforward to garner intuition in the context of TDA – we use a weaker but more technically-involved equivalence relation on loops called homology to utilize efficient algorithms such as Ripser [6] and GUDHI [38] in computing topological features. For a rigorous treatment of homotopy and homology, see [22].

For a given two-dimensional dataset X such that there exists a non-contractible loop ℓ within $\breve{C}_r(X)$, we define the *birth radius* of the equivalence class of non-contractible loops containing ℓ as the smallest real number b such that some loop in $\breve{C}_r(X)$ which is equivalent to ℓ and which is contained in the subset $\breve{C}_b(X)$ of $\breve{C}_r(X)$ exists. Similarly, we define the *death radius* of the equivalence class of non-contractible loops containing ℓ as the smallest real number d such that $r \leqslant d$ and such that ℓ becomes contractible when regarded as a loop in $\breve{C}_d(X)$. That is, the birth radius of an equivalence class of non-contractible loops is the smallest radius at which the equivalence class of that non-contractible loop forms, and the death radius is the smallest radius at which it vanishes (i.e., becomes contractible). For $r \in [b, d]$, the equivalence class of non-contractible loops 'persists,' and this motivates the definition of the *persistence* of an equivalence class of non-contractible loops as the difference between the death radius and the birth radius. The two non-trivial equivalence classes of non-contractible loops in Fig. 1 have coordinates $(0.57, 0.81)$ and $(1.55, 3.01)$ in the persistence diagram and correspond to the subset of data clustered near $(1, 1)$ and the subset of data clustered near $(8, 8)$, respectively. Note that the larger loop-like structure of data in the upper-right corner of each subplot has a larger persistence than the smaller loop-like structure of data in the lower-left corner of each subplot (i.e. $3.01 - 1.55 = 1.46 > 0.81 - 0.57 = 0.24$).

The cycle representatives of a given equivalence class of non-contractible loops $\{\ell_\alpha\}_{\alpha \in I}$ (note that I is an uncountable indexing set) with birth radius b and death radius d are the subsets of the data which give rise to non-contractible loops in $\breve{C}_r(X)$ with birth radius b and death radius d. For example, the cycle representatives of the equivalence class of non-contractible loops with birth radius 0.5 and death radius $\frac{\sqrt{2}}{2} \approx 0.71$ in Fig. 2 are given by $\{\{a, b, c, d\}, \{a, b, c, d, e\}\}$. The Python package Homcloud can be used to identify cycle representatives which are optimal in some sense such as having the minimum number of points or spanning the minimum area among all cycle representatives [45]. Associating a single optimal cycle representative to each equivalence class of non-contractible loops is important 1) for reproducibility and 2) to select cycle representatives which more closely resemble the P, Q, S, and T-waves for the relevant equivalence classes of non-contractible loops.

2. Methods

The free and publicly available Shaoxing Hospital Zhejiang University School of Medicine electrocardiogram (ECG) database was used in this study [69]. This database consists of 10646 12-lead ECG signals, each spanning 10 seconds with a sampling frequency (i.e. the number of electric potential differences recorded per second) of 500 Hz, of which 10605 have non-empty Lead 2 signals. This study

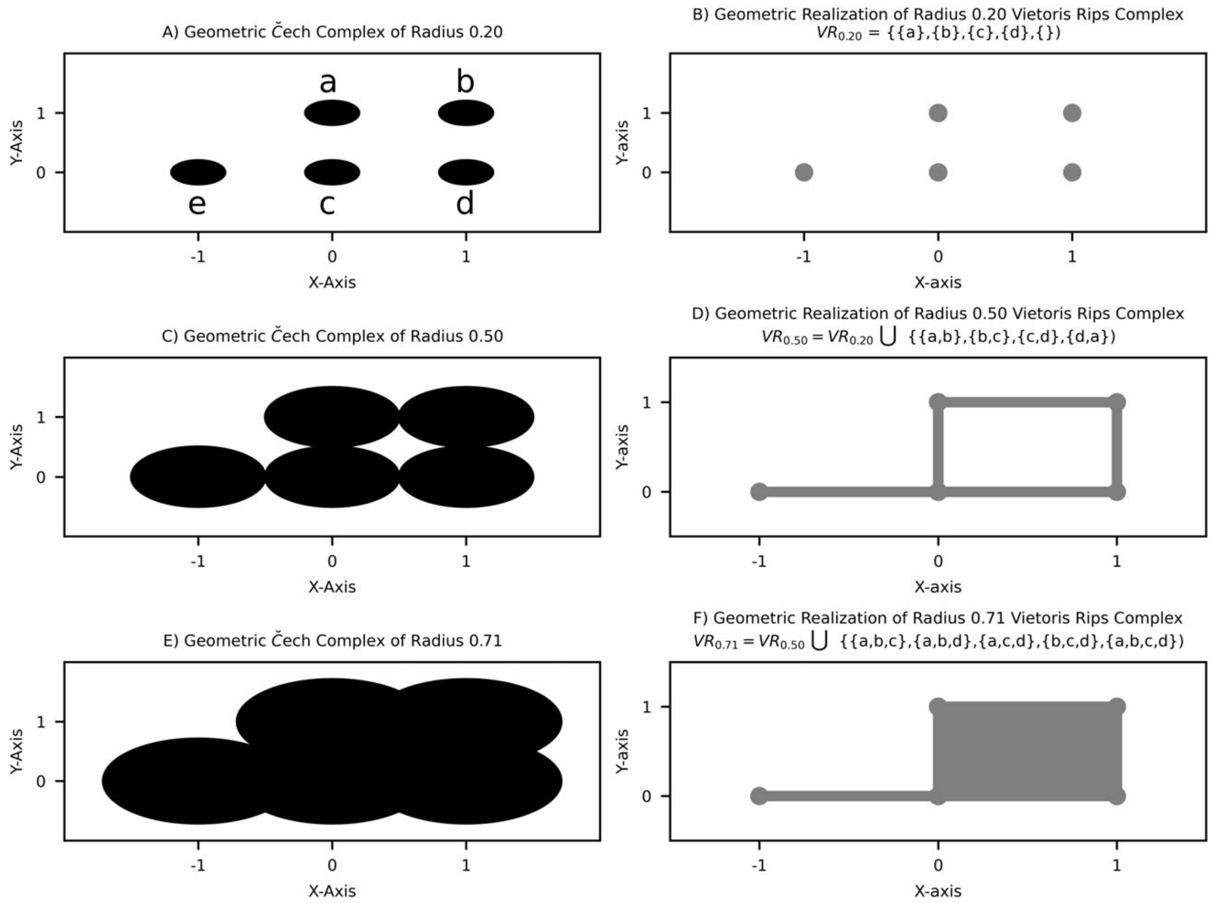
A) Geometric Čech Complex of Radius 0.20
a
b
e
c
d
Y-Axis
X-Axis
-1
0
1
B) Geometric Realization of Radius 0.20 Vietoris Rips Complex
$VR_{0.20}$ = {{a},{b},{c},{d},{})
Y-axis
X-axis
C) Geometric Čech Complex of Radius 0.50
D) Geometric Realization of Radius 0.50 Vietoris Rips Complex
$VR_{0.50} = VR_{0.20} \bigcup$ {{a,b},{b,c},{c,d},{d,a})
E) Geometric Čech Complex of Radius 0.71
F) Geometric Realization of Radius 0.71 Vietoris Rips Complex
$VR_{0.71} = VR_{0.50} \bigcup$ {{a,b,c},{a,b,d},{a,c,d},{b,c,d},{a,b,c,d})

Fig. 2. **Relationship between Geometric Čech Complex of radius r and geometric realization of radius r Vietoris Rips Complex.** A–C: Geometric Čech Complex of radius 0.2, 0.5, 0.71 depicted in black, respectively; D–F: geometric realization of radius 0.2, 0.5, 0.71 Vietoris Rips Complex, respectively.

strictly utilizes Lead 2, i.e. the ‘rhythm lead’, so the term ‘ECG signal’ is henceforth used to refer to Lead 2 ECG signals. Each ECG signal is labeled with one of 11 rhythms by professional experts. The distribution of these 11 rhythms across the 10605 ECG signals is shown in Table 1.

ECG signals are typically characterized as 1-dimensional lists of real numbers of length $F \cdot t_{\max}$ where F is the sampling frequency of the ECG machine (i.e. the number of electric potential differences recorded per second), $t_{\max}$ is the total amount of time (in seconds) over which the signal was gathered, and each real number in the list represents the signal amplitude at the given time index. In order to compute 1-dimensional topological features of an ECG signal, the ECG signal must be considered as a subset of $\mathbb{R}^2$. Therefore, rather than treat a given ECG signal S as a one-dimensional list with a sampling frequency F over a length of time $t_{\max}$, we use the equivalent formulation of S given by $S = \{(t, f(t)) | t \in D\} \subset \mathbb{R}^2$ where $D = \{\frac{i}{F} | i \in \{1, \ldots, F \cdot t_{\max}\}\}$ represents the set of time indices and $f : D \to \mathbb{R}$ defines the signal amplitude at each time index.

In the remainder of this section, we describe 1) ECG signal processing prior to extraction of topological features, 2) the construction of predictor variables derived from persistent homology, and 3) the

Table 1

Rhythm distribution

Rhythm	Count (total = 10605)	Percentage of all signals
Atrial Flutter	445	4.20%
Atrial Fibrillation	1780	16.78%
Atrial Tachycardia	121	1.14%
Atrioventricular Node Reentrant Tachycardia	16	0.15%
Atrioventricular Reentrant Tachycardia	8	0.08%
Sinoatrial Block	399	3.76%
Sinus Atrium to Atrial Wandering	7	0.07%
Sinus Bradycardia	3888	36.67%
Sinus Rhythm	1826	17.22%
Sinus Rachycardia	1568	14.79%
Supraventricular Tachycardia	547	5.16%

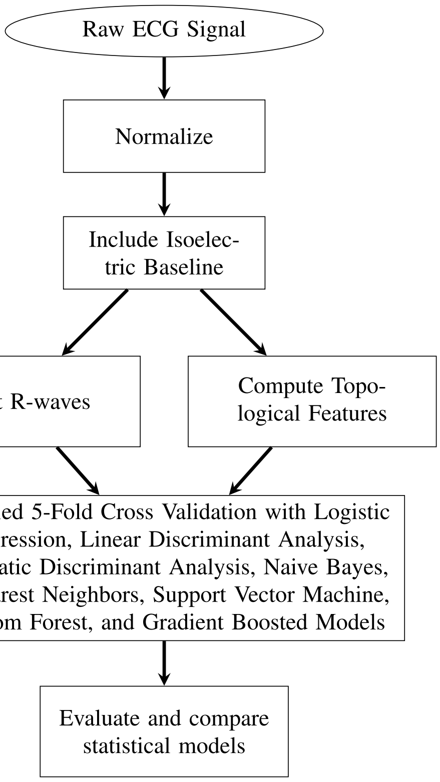
Raw ECG Signal
Normalize
Include Isoelectric Baseline
Detect R-waves
Compute Topological Features
Stratified 5-Fold Cross Validation with Logistic Regression, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Naive Bayes, K-Nearest Neighbors, Support Vector Machine, Random Forest, and Gradient Boosted Models
Evaluate and compare statistical models

Fig. 3. Flowchart of ECG signal processing and arrhythmia classification.

statistical modeling approaches and evaluation metrics used. A flowchart providing an overview of our approach to arrhythmia detection is shown in Fig. 3.

2.1. Electrocardiogram signal processing

Given a raw ECG signal $S = \{(t, f(t)) | t \in D\} \subset \mathbb{R}^2$ with time domain $D = \{\frac{h}{F} | h \in \{1, \ldots, F * t_{\max}\}\}$ and signal amplitude given by $f : D \to \mathbb{R}$, the signal is first normalized by applying the transformation $g : f(D) \to [0, 1]$ given by:

$$g\big(f(t)\big) = \frac{f(t) - \min\{f(D)\}}{\max\{f(D)\} - \min\{f(D)\}}. \tag{1}$$

The resulting signal $S_{\text{normalized}} = \{(t, g(f(t))) | t \in D\} \subset \mathbb{R}^2$ has maximum amplitude $\max\{g(f(D))\} = 1$ and minimum amplitude $\min\{g(f(D))\} = 0$. Since equivalence classes of non-contractible loops are not scale-invariant, this normalization is necessary for the magnitude of persistent homology-derived statistics to be comparable across ECG signals.

Next, an isoelectric baseline is included in $S_{\text{normalized}}$ in order to form ‘loop-like’ structures with non-trivial topological properties in the ECG signal corresponding to the P, Q, S, and T-waves (if they are present). The inclusion of this baseline emphasizes the shape of the P, Q, S, and T-waves (if they exist to begin with), as illustrated in Fig. 4. This is done by inserting the baseline value computed as the median of $g(f(D))$ at the beginning of the signal and between every pair of consecutive time indices, doubling the number of points of the signal while still spanning the same amount of time. More explicitly, after the inclusion of the isoelectric baseline to $S_{\text{normalized}}$, we obtain the signal:

$$\begin{aligned} &S_{\text{processed}} = \big\{\big(t, h\big(g\big(f(t)\big)\big)\big) | t \in E\big\}, \\ &E = \left\{\frac{i}{2F} \middle| i \in \{1, \ldots, 2 \cdot F \cdot t_{\max}\}\right\}, \\ &h : [0, 1] \to [0, 1] : g\left(f\left(\frac{i}{F}\right)\right) \mapsto \begin{cases} \text{median}\{g(f(D))\} & \text{if } i \text{ is odd} \\ g(f(\frac{i}{F})) & \text{if } i \text{ is even.} \end{cases} \end{aligned} \tag{2}$$

Note the appearance of highly-persistent equivalence classes of non-contractible loops around birth radius 0.005 once the isoelectric baseline is included in the normal sinus rhythm ECG signal in Fig. 5. Also observe in Fig. 5 that for a rhythm such as atrial fibrillation with the property of absent/attenuated P-waves, the isoelectric baseline does not produce additional highly-persistent equivalence classes of non-contractible loops to the same extent that it does for rhythms with normal wave-shape such as normal sinus rhythm.

The onset of each QRS-complex in the processed ECG signal $S_{\text{processed}}$ is identified using Zong, Moody, and Jiangs’ approach of “passing $S_{\text{processed}}$ through a low-pass filter, applying a transformation with a non-linear scaling factor to enhance the QRS-complexes and suppress unwanted noise, and applying adaptive thresholds to the signal to determine the onset of each QRS-complex” [70]. An illustration of the preprocessing transformations applied to a raw signal is shown in Fig. 5.

2.2. Construction of predictor variables

Each equivalence class of non-contractible loops with birth radius b and death radius d corresponds to a set of subsets of $S_{\text{processed}}$ given by $Y_{\text{processed}} = \{S^\star \subset S_{\text{processed}} | (\text{birth radius of } S^\star = b) \text{ and } (\text{death radius of } S^\star = d)\}$. That is, there may be multiple subsets of data which generate a given

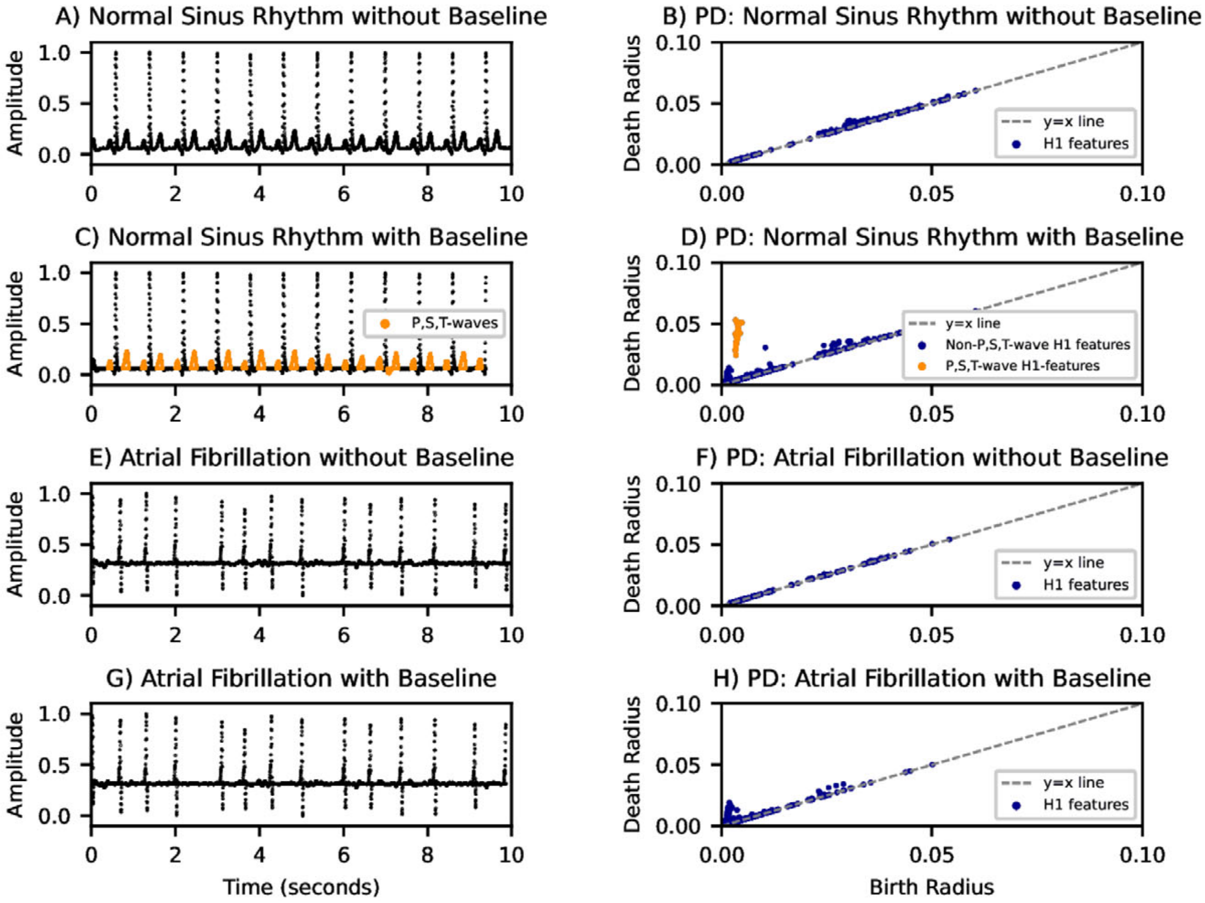

Fig. 4. Illustration depicting the effect of the isoelectric baseline on the persistence diagrams of ECG signals. PD: persistence diagram. A–B) normal sinus rhythm ECG signal without baseline and corresponding PD. C–D) same as A–B but with the isoelectric baseline included. Note the cluster of topological features that appeared and the P, S, and T-waves their area-minimal cycle representatives correspond to. E–F) atrial fibrillation without baseline included and corresponding PD. G–H) same as E–F but with the isoelectric baseline included. H1 features: equivalence classes of non-contractible loops.

equivalence class of non-contractible loops. Equivalently, for a single point in a persistence diagram, there may be multiple subsets of data such that the Geometric Čech complex births and vanishes the given equivalence class of non-contractible loops with the same birth and death radii. The Python package Homcloud is used to compute a single unique area-minimal cycle representative $S^{\star}$ from $Y_{\text{processed}}$ for each equivalence class of non-contractible loops in the signal $S_{\text{processed}}$ [45]. Given an equivalence class of non-contractible loops with centroid coordinates of the area-minimal cycle representative (T, A), the effective centroid coordinates (x, y) are computed as

- $x = t_R - T$ where t_R is the time-coordinate of the onset of the subsequent QRS-complex.
- $y = \frac{A-baseline}{1-baseline}$ where *baseline* represents the amplitude value of the isoelectric baseline median$\{g(f(D))\}$.

The computation of the effective centroid coordinates of an area-minimal cycle representative is depicted in Fig. 6. The equivalence classes of non-contractible loops with centroid time coordinate T larger than the largest of all onsets of the QRS-complexes are not considered to ensure that the effective centroid

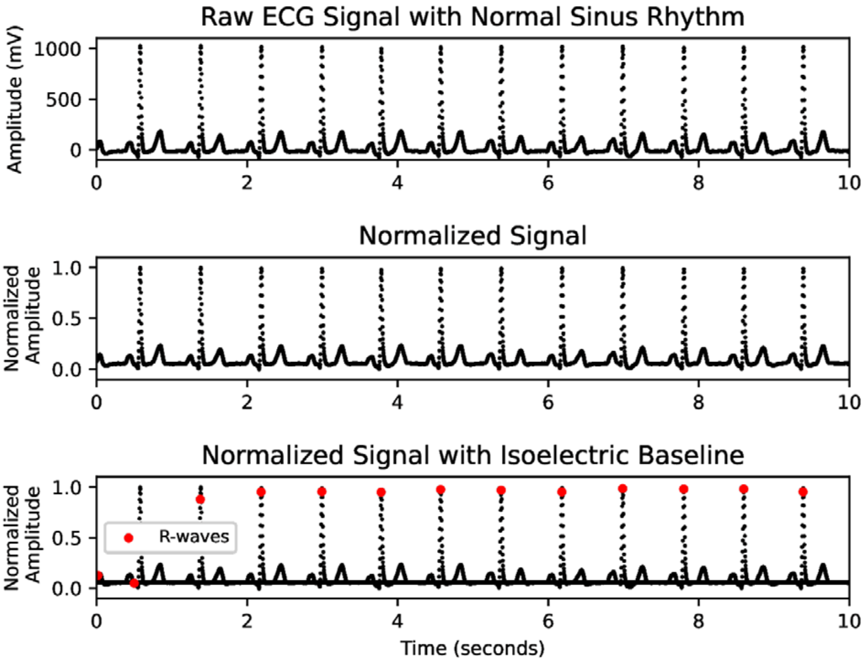

Fig. 5. Depiction of preprocessing transformations applied to a normal sinus rhythm ECG signal. A) raw ECG signal with normal sinus rhythm. B) normalized ECG signal with maximum amplitude 1 and minimum amplitude 0. C) normalized ECG signal with isoelectric baseline included and R-waves identified.

time coordinates can always be computed. This effectively trims $S_{\text{processed}}$ to end with a point representing the onset of a QRS-complex. Furthermore, all equivalence classes of non-contractible loops with area-minimal cycle representative centroid amplitude coordinate A larger than $\frac{1-baseline}{2}$ where $baseline = \text{median}\{g(f(D))\}$ are discarded to obtain a larger proportion of highly-persistent equivalence classes of non-contractible loops corresponding to clinically-relevant subsets of ECG signals such as P, Q, S, and T-waves. For example, the computation of the effective centroid time coordinate for area-minimal cycle representatives that represent P-waves is a proxy of the clinically-relevant PR-interval. The computation of the effective centroid amplitude coordinate normalizes the amplitude coordinates of centroids of area-minimal cycle representatives across signals with differing isoelectric baselines.

The persistent homology of the processed signal $S_{\text{processed}}$ is then computed, and the N most persistent equivalence classes of non-contractible loops are used to construct predictor variables for use in rhythm classification for $N \in \{5, 6, \dots, 29, 30\}$. Specifically, for each of the N-th most persistent equivalence classes of non-contractible loops, the persistence, birth radius, effective time-coordinate of the centroid of the area-minimal cycle representative relative to the subsequent QRS-complex, effective amplitude-coordinate of the centroid of the area-minimal cycle representative relative to the isoelectric baseline, and Shannon entropy of the vector $\frac{(a,b,c,d,e)}{\text{sum}((a,b,c,d,e))}$ where a = persistence, b = birth radius, c = death radius, d = centroid time-coordinate, and e = centroid amplitude-coordinate are used as predictor variables. Additional predictor variables include the mean and standard deviation of the persistences, birth radii, area-minimal cycle representative centroid time coordinates, and area-minimal cycle representative centroid amplitude coordinates of the N-most persistent equivalence classes of non-contractible loops along with the mean and standard deviation of the RR-intervals. Lastly, the total number of R-waves, the total number of equivalence classes of non-contractible loops, and the Shannon entropy of the normalized

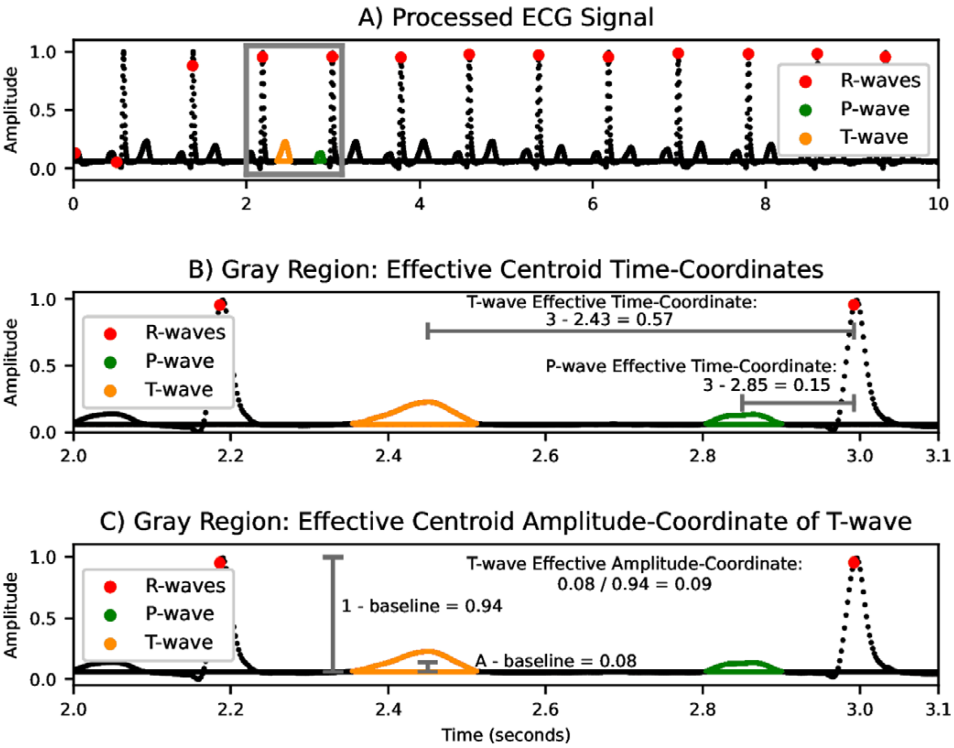

Fig. 6. Computation of the effective centroid coordinates of an area-minimal cycle representatives. A) processed ECG signal with normal sinus rhythm and R-waves, an area-minimal cycle representative corresponding to a P-wave, and an area-minimal cycle representative corresponding to a T-wave identified. B) zoomed-in region depicting the computations of the effective time-coordinates of the two area-minimal cycle representatives. C) zoomed-in region depicting the computation of the effective amplitude-coordinate of the area-minimal cycle representative corresponding to the T-wave.

distribution of all persistences are also used as predictor variables. Note that death radii statistics are not included as predictor variables since their inclusion would introduce undesired collinearity due to the persistence of a given equivalence class of non-contractible loops being the difference between the death radius and the birth radius.

2.3. *Statistical modeling and evaluation*

Three different binary classifications are carried out:

- Atrial Fibrillation vs. Non-Atrial Fibrillation
- Arrhythmia vs. Normal Sinus Rhythm
- Arrhythmias with Morphological Changes vs. Sinus Rhythm with Bradycardia and Tachycardia Treated as Non-Arrhythmia

For each of the three binary classifications, Logistic Regression, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Naive Bayes, Random Forest, Gradient Boosted Decision Tree, K-Nearest Neighbors, and Support Vector Machine with Linear, Radial, and Polynomial Kernel Models are constructed. For background on the theory and/or implementation of these statistical models, see [28]. Stratified 5-fold cross-validation is performed, and in each of the 5 folds, the true positives (TP),

		True Label	
		Positive	Negative
Predicted Label	Positive	TP	FP
	Negative	FN	TN

Fig. 7. Confusion matrix.

false positives (FP), false negatives (FN), and true negatives (TN) are recorded in a confusion matrix like that shown in Fig. 7 for each statistical model used. The mean and standard deviation of the F1-Scores, Accuracies, Sensitivities, Specificities, Positive Predictive Values (PPVs), and Negative Predictive Values (NPVs) across the five folds are recorded. Definitions of these evaluation metrics can be found in [52].

The optimal hyperparameters for the Random Forest, Gradient Boosted Decision Tree, K-Nearest Neighbors, and Support Vector Machines with Radial and Polynomial Kernel Models were chosen as the hyperparameters which yielded the largest mean F1-Score across all folds in 5-fold stratified cross validation. The grid search spaces of hyperparameters for the relevant models are:

- Random Forest:
  - ∗ number of trees $\in \{500, 1250, 2000, 3000\}$
  - ∗ number of variables randomly sampled $\in \{\text{int}(0.25 \cdot T), \text{int}(0.5 \cdot T), \text{int}(0.75 \cdot T), T\}$ where T is the total number of predictor variables.
- Gradient Boosted Decision Tree:
  - ∗ number of trees $\in \{500, 1250, 2000, 3000\}$
  - ∗ interaction depth $\in \{5, 10, 15, 20\}$
- K-Nearest Neighbors:
  - ∗ $K \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$.
- Support Vector Machine with Radial Kernel:
  - ∗ cost $= 1$.
  - ∗ $\gamma \in \{0.5, 1, 2, 3, 4, 5\}$.
- Support Vector Machine with Polynomial Kernel:
  - ∗ cost $= 1$.
  - ∗ degree $\in \{2, 3, 4, 5\}$.

For each of the three binary classifications, the relative influence of the predictor variables in the statistical model yielding the largest mean F1-score across the five folds is quantified using the methods described in Section 8.1 of "Greedy Function Approximation: A Gradient Boosting Machine" by Friedman [18].

3. Results/discussion

The mean and standard deviation across the five folds for the binary classifications of (i) Atrial Fibrillation vs. Non-Atrial Fibrillation, (ii) Arrhythmia vs. Normal Sinus Rhythm, and (iii) Arrhythmia

Table 2

Binary classification outcomes: atrial fibrillation vs. Non-atrial fibrillation

Model	F1-score	Accuracy	Sensitivity	Specificity	PPV	NPV	Optimal N
Logistic regression	0.938 ± 0.002	0.896 ± 0.004	0.947 ± 0.005	0.646 ± 0.018	0.930 ± 0.003	0.712 ± 0.018	24
Linear discriminant analysis	0.934 ± 0.002	0.890 ± 0.004	0.941 ± 0.004	0.637 ± 0.019	0.928 ± 0.003	0.686 ± 0.015	30
Quadratic discriminant analysis	0.917 ± 0.004	0.864 ± 0.008	0.908 ± 0.008	0.642 ± 0.063	0.927 ± 0.012	0.585 ± 0.019	25
Naive Bayes	0.890 ± 0.006	0.818 ± 0.009	0.880 ± 0.011	0.511 ± 0.034	0.899 ± 0.006	0.463 ± 0.024	5
Random forest	0.955 ± 0.004	0.925 ± 0.007	**0.964 ± 0.004**	0.734 ± 0.043	0.947 ± 0.008	0.803 ± 0.016	4
Gradient boosted model	**0.967 ± 0.003**	**0.946 ± 0.006**	0.959 ± 0.004	**0.880 ± 0.019**	**0.975 ± 0.004**	**0.813 ± 0.018**	20
K-Nearest Neighbors	0.942 ± 0.004	0.894 ± 0.007	0.925 ± 0.006	0.712 ± 0.021	0.952 ± 0.004	0.660 ± 0.035	23
Support Vector Machine: linear kernel	0.941 ± 0.003	0.898 ± 0.005	0.935 ± 0.004	0.706 ± 0.020	0.942 ± 0.006	0.705 ± 0.022	29
Support Vector Machine: radial kernel	0.927 ± 0.002	0.868 ± 0.003	0.869 ± 0.003	0.856 ± 0.025	0.991 ± 0.002	0.272 ± 0.016	5
Support Vector Machine: polynomial kernel	0.937 ± 0.004	0.890 ± 0.007	0.908 ± 0.005	0.749 ± 0.022	0.964 ± 0.002	0.539 ± 0.031	17

with Morphological Changes vs. Sinus Rhythm with Bradycardia and Tachycardia Treated as Non-Arrhythmia with the hyperparameters yielding the largest F1-Score are shown in Tables 2, 3, and 4, respectively. The results corresponding to the top-performing model with respect to each evaluation metric are displayed in bold. Observe that the Gradient Boosted Decision Tree Model outperforms all other models with respect to F1-Score and Accuracy across each of the three binary classification tasks, closely followed by the Random Forest Model. The maximum mean F1-Score attained by the Gradient Boosted Decision Tree Model across the five folds was 0.967, 0.839, and 0.943 for binary classification of Atrial Fibrillation vs. Non-Atrial Fibrillation, Arrhythmia vs. Normal Sinus Rhythm, and Arrhythmia with Morphological Changes vs. Sinus Rhythm with Bradycardia and Tachycardia Treated as Non-Arrhythmia, respectively. The corresponding mean Accuracy attained by the Gradient Boosted Decision Tree Model across the five folds was 0.946, 0.946, and 0.921 for binary classification of Atrial Fibrillation vs. Non-Atrial Fibrillation, Arrhythmia vs. Normal Sinus Rhythm, and Arrhythmia with Morphological Changes vs. Sinus Rhythm with Bradycardia and Tachycardia Treated as Non-Arrhythmia, respectively. The Gradient Boosted Decision Tree and Random Forest models outperformed all other models with respect to the area under the Receiver-Operator Characteristic Curves (AUC) for all three classification tasks as seen in Fig. 8, Fig. 9, and Fig. 10. This may be due to heterogeneity of the data; regardless, in computer-aided ECG analysis, interpretability of statistical models may be less important than the performance of said models, rendering more support in favor of ensemble and tree-based modeling approaches given their favorable performance.

Table 3

Binary classification outcomes: arrhythmia vs. Normal sinus rhythm

Model	F1-score	Accuracy	Sensitivity	Specificity	PPV	NPV	Optimal N
Logistic regression	0.634 ± 0.019	0.876 ± 0.004	0.622 ± 0.029	0.929 ± 0.002	0.647 ± 0.009	0.922 ± 0.005	10
Linear discriminant analysis	0.629 ± 0.023	0.867 ± 0.008	0.652 ± 0.028	0.912 ± 0.006	0.607 ± 0.022	0.927 ± 0.006	20
Quadratic discriminant analysis	0.481 ± 0.012	0.709 ± 0.010	0.783 ± 0.016	0.694 ± 0.011	0.347 ± 0.010	0.939 ± 0.004	24
Naive Bayes	0.460 ± 0.010	0.673 ± 0.016	0.809 ± 0.025	0.644 ± 0.022	0.322 ± 0.010	0.942 ± 0.006	24
Random forest	0.829 ± 0.010	0.942 ± 0.003	0.812 ± 0.017	0.969 ± 0.003	0.847 ± 0.014	0.961 ± 0.003	8
Gradient boosted model	**0.839 ± 0.011**	**0.946 ± 0.003**	0.815 ± 0.019	**0.974 ± 0.002**	**0.866 ± 0.009**	0.962 ± 0.004	12
K-Nearest Neighbors	0.722 ± 0.004	0.899 ± 0.007	0.747 ± 0.006	0.924 ± 0.021	0.664 ± 0.004	0.964 ± 0.035	5
Support Vector Machine: linear kernel	0.638 ± 0.003	0.889 ± 0.005	0.743 ± 0.004	0.910 ± 0.020	0.563 ± 0.006	0.958 ± 0.022	20
Support Vector Machine: radial kernel	0.720 ± 0.002	0.912 ± 0.003	**0.869 ± 0.003**	0.918 ± 0.025	0.612 ± 0.002	**0.982 ± 0.016**	5
Support Vector Machine: polynomial kernel	0.631 ± 0.004	0.891 ± 0.007	0.797 ± 0.005	0.902 ± 0.022	0.516 ± 0.002	0.975 ± 0.031	21

Recall that TDA quantifies the ‘shape’ of data. Thus, the motivation behind presenting the classifications of both (i) Arrhythmia vs. Normal Sinus Rhythm and (ii) Arrhythmias with Morphological Changes vs. Sinus Rhythm with Bradycardia and Tachycardia Treated as Non-Arrhythmia is to illustrate how the results are improved when TDA is used to classify two groups that primarily have different shapes, not frequencies. With this in mind, it may not be surprising that the presented TDA approach performs much better when classifying arrhythmias when the only two arrhythmias characterized solely by abnormal periodicity (assuming the individual has at most one rhythm as is the case in the data used in this study) – i.e. tachycardia and bradycardia – are not considered to be part of the arrhythmia group.

The relative influence [18] of each predictor variable in the top-performing model with respect to mean F1-score (i.e. Gradient Boosted Decision Tree model) across the five folds in the classifications of Atrial Fibrillation vs. Non-Atrial Fibrillation, Arrhythmia vs. Normal Sinus Rhythm, and Arrhythmia with Morphological Changes vs. Sinus Rhythm with Bradycardia and Tachycardia Treated as Non-Arrhythmia are shown in Fig. 11, Fig. 12, Fig. 13 in Appendix B. Atrial fibrillation is characterized by (1) absent/attenuated P-waves and (2) irregularly irregular frequency, so it is not surprising that the standard deviation of the RR-interval holds most influence for the classification of Atrial Fibrillation vs. Non-Atrial Fibrillation. Note that 9 of the 15 most influential predictor variables in the classification of Atrial Fibrillation vs. Non-Atrial Fibrillation stem from area-minimal cycle representatives and that $\frac{44}{115} = 38.3\%$ of all predictor variables stem from area-minimal cycle representatives. In the case of Arrhythmia vs. Normal Sinus Rhythm, 8 of the 15 most influential predictor variables stem from area-minimal cycle representatives and $\frac{28}{76} = 36.8\%$ of all predictor variables stem from area-minimal

Table 4

Binary classification outcomes: arrhythmia with morphological changes vs. Sinus rhythm with bradycardia and tachycardia treated as non-arrhythmia

Model	F1-score	Accuracy	Sensitivity	Specificity	PPV	NPV	Optimal N
Logistic regression	0.904 ± 0.002	0.865 ± 0.004	0.932 ± 0.003	0.717 ± 0.011	0.878 ± 0.004	0.828 ± 0.007	30
Linear discriminant analysis	0.905 ± 0.002	0.866 ± 0.003	0.927 ± 0.007	0.734 ± 0.012	0.884 ± 0.004	0.821 ± 0.012	30
Quadratic discriminant analysis	0.857 ± 0.004	0.797 ± 0.003	0.884 ± 0.014	0.607 ± 0.025	0.831 ± 0.007	0.706 ± 0.018	25
Naive Bayes	0.859 ± 0.004	0.794 ± 0.006	0.912 ± 0.008	0.536 ± 0.013	0.812 ± 0.005	0.735 ± 0.018	27
Random forest	0.933 ± 0.005	0.906 ± 0.007	0.952 ± 0.005	0.805 ± 0.012	0.915 ± 0.005	0.885 ± 0.012	10
Gradient boosted model	**0.943 ± 0.004**	**0.921 ± 0.006**	**0.955 ± 0.004**	0.847 ± 0.013	0.932 ± 0.005	**0.896 ± 0.009**	10
K-Nearest Neighbors	0.905 ± 0.004	0.861 ± 0.007	0.883 ± 0.006	0.807 ± 0.021	0.923 ± 0.004	0.741 ± 0.035	19
Support Vector Machine: linear kernel	0.905 ± 0.003	0.866 ± 0.005	0.886 ± 0.004	0.815 ± 0.020	0.923 ± 0.006	0.745 ± 0.022	29
Support Vector Machine: radial kernel	0.883 ± 0.002	0.828 ± 0.003	0.813 ± 0.003	**0.896 ± 0.025**	**0.968 ± 0.002**	0.507 ± 0.016	5
Support Vector Machine: polynomial kernel	0.897 ± 0.004	0.845 ± 0.007	0.848 ± 0.005	0.833 ± 0.022	0.949 ± 0.002	0.637 ± 0.031	16

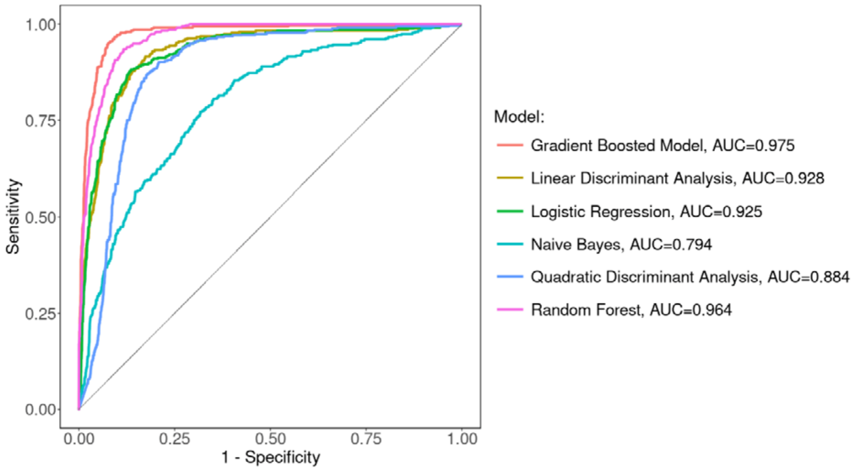

Fig. 8. Receiver operator characteristic curve for classification of atrial fibrillation vs. Non-atrial fibrillation.

cycle representatives. Lastly, for the classification of Arrhythmia with Morphological Changes vs. Sinus Rhythm with Bradycardia and Tachycardia Treated as Non-Arrhythmia, 11 of the 15 most influential predictor variables are derived from area-minimal cycle representatives while $\frac{24}{66} = 36.4\%$ of all pre-

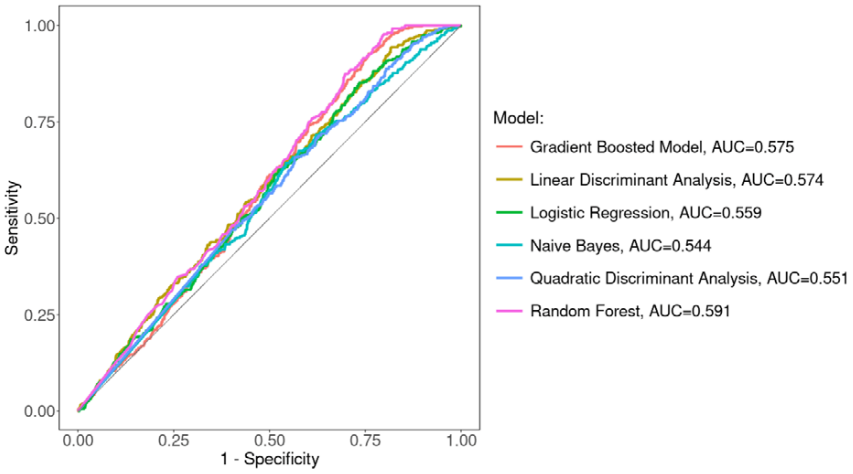

Fig. 9. Receiver operator characteristic curve for classification of arrhythmia vs. Sinus rhythm.

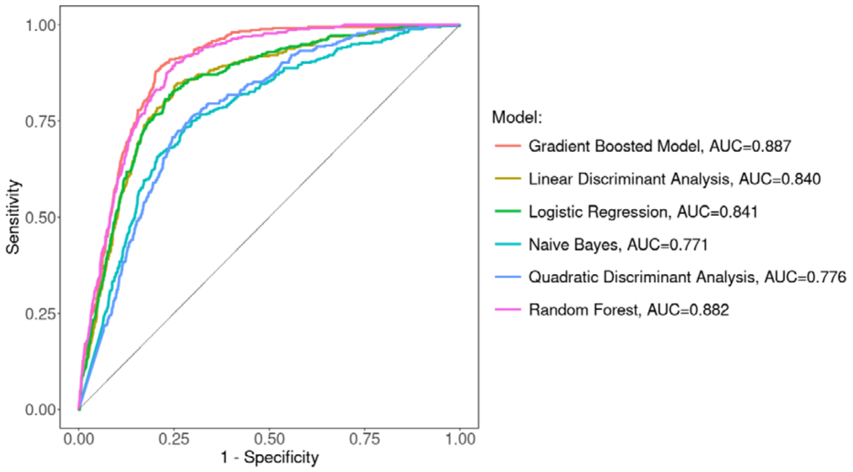

Fig. 10. Receiver operator characteristic curve for classification of arrhythmia with morphological changes vs. Sinus rhythm with bradycardia and tachycardia treated as non-arrhythmia.

dictor variables are derived from area-minimal cycle representatives. These results demonstrate that the predictor variables constructed from properties of the area-minimal cycle representatives are highly influential in the top-performing model regardless of the arrhythmia classification task.

The methods used in other studies that approach computer-aided ECG rhythm classification through a combination of TDA and machine learning are summarized in Table 5. Due to the wide range of classification tasks performed and evaluation metrics used in these studies, the classification tasks and evaluation metrics are not shown in Table 5 to avoid (i) presenting misleading comparisons and (ii) subjectivity in choosing the results from other studies to present. These other studies use a variety of databases [19,42] and sometimes a sample size on the scale of tens or hundreds, in addition to having longer – and consequently more informative – signals compared to the database used in this study [69]. Another factor to consider when comparing analyses of TDA and machine learning in ECG rhythm classification is the fact that different ECG databases often have signals labeled with different rhythms

Table 5

Comparison of studies applying TDA and machine learning to arrhythmia classification

Title	Database(s)	Preprocessing	Features	Model(s)
Topological Data Analysis for Arrhythmia Detection Through Modular Neural Networks [16]	PhysioNet MIT-BIH Normal Sinus Rhythm, Arrhythmia, Supraventricular Arrhythmia, Malignant Ventricular Arrhythmia, and Long Term Database	Resample as different frequency, remove baseline, finite impulse response filter, Kalman filter, rescale, translate	Coefficients from Discrete Fourier Transform of sliding windows; linear relationships between P, Q, R, S, & T-waves; extrema, mean, standard deviation, kurtosis, skewness, entropy, crossing-overs, PCA reduction of persistence statistics	Autoencoder
Nonlinear dynamic approaches to identify atrial fibrillation progression based on topological methods [55]	PhysioBank long-term atrial fibrillation database; PhysioNet MIT-BIH normal sinus rhythm database	Normalize, time-delay embedding	Number and persistence of 1-dimensional topological features and fractal dimension	Feed-forward back propagation neural network
Classification of Single-Lead Electrocardiograms: TDA Informed Machine Learning [26]	Alivecor	Butterworth filter and time-delay embedding	Sum, mean, standard deviation, skewness, kurtosis of birth, death, and/or persistence of 0, 1, & 2-dimensional topological features	Random forest classifier
Persistence Landscape-based Topological Data Analysis for Personalized Arrhythmia Classification [36]	PhysioNet MIT-BIH Long-Term database	Resample at different frequency, Butterworth filter, detect R waves and segment signal, time-delay embedding, downsample	Persistence landscape-derived statistics	Random forest classifier
Early Ventricular Fibrillation Prediction Based on Topological Data Analysis of ECG Signal [34]	PhysioNet CUDB, SDDB, PTBDB	Resample at different frequency, moving average filter, normalization, time-delay embedding	Sum, mean, and variance of persistences of 0, 1, & 2-dimensional topological features; box-counting features; heart rate variability features	Logistic regression, decision trees, SVM, KNN classifier
Ventricular Fibrillation and Tachycardia Detection Using Features Derived from Topological Data Analysis [41]	AHA 2000 series and PhysioNet MIT-BIH Malignant Arrhythmia Database	Infinite impulse response filter, time-delay embedding	Derived from representations of time domain signal, embedded signal, persistence diagram, persistence landscape representation, weighted silhouettes representation	KNN classifier
A Topology Informed Random Forest Classifier for ECG Classification [27]	PhysioNet/Computing in Cardiology Challenge 2020	Time-delay embedding	Persistence entropy and statistics derived from persistence diagram and persistence landscape	Two-level random forest classifier
A Novel Heart Disease Classification Algorithm based on Fourier Transform and Persistent Homology [43]	PhysioNet MIT-BIH Arrhythmia Database	Butterworth filter, sliding window fast Fourier Transform to embed signal in higher dimension	Persistence entropy and persistence statistics	SVM

that may not be found in other ECG databases. The approach presented here attains similar results as these previous studies with respect to classification outcomes while utilizing a novel ECG signal processing pipeline and topological predictor variable construction, particularly with respect to using information derived from area-minimal cycle representatives.

4. Conclusion

The method presented here differs from other methods utilizing TDA and machine learning in three main ways:

- by using information about optimal cycle representatives of equivalence classes of non-contractible loops when constructing topological predictor variables.
- by focusing only on the N-most persistent equivalence classes of non-contractible loops when constructing topological predictor variables.
- by introducing an isoelectric baseline to create non-trivial equivalence classes of non-contractible loops corresponding to the P, Q, S, and T-waves (if they are present to begin with).

This novel approach to ECG signal processing and construction of topological predictors yields classification results on par with other methods proposed in the literature and demonstrates the utility of optimal cycle representatives in TDA. Future directions include multiclass rhythm classification, other methods of defining the isoelectric baseline to account for baseline wander in longer ECG signals, including statistics derived from optimal cycle representatives in other approaches such as sliding window and Fast Fourier Transform embeddings, and including an isoelectric baseline prior to embedding ECG signals in higher dimensions. Several studies have used TDA-derived statistics as input to neural networks [16,53,55]; however, to the author's knowledge, there has been no study performed which utilizes persistence images [1] as the TDA-derived input for neural networks in arrhythmia detection, yielding another direction for future work.

There have been people working on computer-aided ECG analysis since the invention of the ECG machine. Over the past 20 years, there have been many machine learning approaches taken, yielding encouraging results. Some of these methods have involved TDA. Regardless of the type of method taken in computer-aided ECG analysis and the goodness of the evaluation metrics, we must take care to not rush to replace ECG interpretation by skilled health care professionals, however tempting the potential time and cost savings may be. In addition to the obvious danger of automated arrhythmia classification algorithms missing a harmful arrhythmia that a skilled healthcare professional would not have missed, bells and whistles from automated arrhythmia detection algorithms can lead to unnecessary medical staff fatigue and an increase in stress and adverse outcomes in hospitalized patients [12,29,33,54,58,59].

The data used in this study are free and publicly available at https://figshare.com/collections/ChapmanECG/4560497/2 [69]. The code used in this study is free and publicly available and can be found on GitHub: https://github.com/hdlugas/ekg_tda_arrhythmia_detection.

Acknowledgement

The computations of topological features and all optimal parameter searches were performed on Wayne State University's High-Performance Computing Grid.

Appendix A. Formalization of persistent homology intuition

We now set out to formalize the notion of “equivalence classes of non-contractible loops that persist for a given range of radius values.” Given a set of data X represented as a finite set of points in $\mathbb{R}^2$, a simplicial complex is constructed as a topological space that approximates the structure of the data.

Definition A.1. *A* simplicial complex *is a collection K of subsets of a finite set V such that:*

- *$\{v\} \in K$ for all $v \in V$, and*
- *if $\tau \subset \sigma$ for $\sigma \in K$, then $\tau \in K$.*

An element of V is referred to as a vertex, *and an element of K with cardinality $n + 1$ is referred to as an n-*simplex.

There are several ways to construct a simplicial complex given a finite set of points in $\mathbb{R}^2$, and to be consistent with the geometry of the toy examples discussed in Section 1.1, we consider the Radius r Vietoris–Rips complex, a simplicial complex constructed by considering a circle of radius r around each point in our dataset and then including $S \subset X$ as a simplex if the intersection of the balls of radius r for each point in S is non-empty. An example of the Radius r Vietoris–Rips complex and its corresponding geometric realization for several values of r is shown in Fig. 2.

Definition A.2. *Given a dataset X represented as a finite subset of $\mathbb{R}^2$, and given a positive real number r, the* radius r Vietoris–Rips complex of *X, denoted $VR_r(X)$, is the simplicial complex given by the collection of all subsets U of X with the property that if $x_1, x_2 \in U$, then $|x_1 - x_2| < r$.*

Note that if $S \subset U$ for $U \in VR_r(X)$, then $|x_1 - x_2| < r$ for all $x_1, x_2 \in U$ implies $|x_1 - x_2| < r$ for all $x_1, x_2 \in S$. Thus the radius r Vietoris Rips complex of a finite subset of $\mathbb{R}^2$ defines a simplicial complex.

We are now in a position to be more concrete about the notion of an “equivalence class of non-contractible loops” within the geometric Čech complex, as discussed in Section 1.1. By an “equivalence class of non-contractible loops,” we are referring to an element of the 1-dimensional homology group of some radius r Vietoris–Rips complex, which we now set out to define.

Let X be a finite subset of $\mathbb{R}^2$, let r be a positive real number, and let C_n be the vector space over $\mathbb{F}_2$ with basis consisting of the elements of $VR_r(X)$ of cardinality $n + 1$ for $n = 0, 1, 2$. Furthermore, suppose there is an ordering on $VR_r(X)$. Consider $0 \xleftarrow{\delta_{-1}} C_0 \overset{\delta_0}{\leftarrow} C_1 \overset{\delta_1}{\leftarrow} C_2 \overset{\delta_2}{\leftarrow} 0$ where $\delta_n([x_0, \ldots, x_n]) = \sum_{i=0}^{i}(-1)^n[x_0, \ldots, \hat{x}_i, \ldots, x_n]$ and $\hat{x}_i$ indicates that x_i is omitted from the ordered simplex. The elements of C_1 are referred to as 1-chains, the elements of $\ker(\delta_0)$ are referred to as 1-cycles, and elements of $\mathrm{im}(\delta_1)$ are referred to as 1-boundaries. Since $\delta_0(\delta_1(v)) = 0$ for all $v \in C_2$, every 1-boundary is an 1-cycle. However, it is not necessarily true that every 1-cycle is an 1-boundary. Intuitively, if we think of X as a point cloud in the plane $\mathbb{R}^2$, the 1-dimensional homology group of $VR_r(X)$ is defined such that its dimension over $\mathbb{F}_2$ counts the number of “holes” in that point cloud.

Definition A.3. *Given $r > 0$ and $VR_r(X)$ where X is a finite subset of $\mathbb{R}^2$, we follow the construction of $\mathbb{F}_2$-vector spaces C_0, C_1, C_2 and linear transformations δ_{-1}, δ_0, δ_1, δ_2 as outlined above and define the first homology group of $VR_r(X)$ as the quotient vector space $H_1(VR_r(X)) = \ker(\delta_0)/\operatorname{im}(\delta_1)$. The $\mathbb{F}_2$-vector space dimension $\beta_1 = dim(H_1(VR_r(X))) = dim(\ker(\delta_0)) - dim(\operatorname{im}(\delta_1))$ of $H_1(VR_r(X))$ is called the first Betti number.*

By increasing r, we create a sequence of Vietoris–Rips Complexes where $VR_r(X) \subset VR_{r'}(X)$ for $r < r'$. We then construct

$$VR_{r_0}(X) \xrightarrow{i_0} VR_{r_1}(X) \xrightarrow{i_1} \cdots \xrightarrow{i_{m-1}} VR_{r_m}(X)$$

where $VR_{r_i}(X)$ is a proper subset of $VR_{r_j}(X)$ for $i < j$ and $i_0, i_1, \ldots, i_{m-1}$ are inclusion homomorphisms. This induces a sequence of $\mathbb{F}_2$-linear functions $i_0^*, i_1^*, \ldots, i_{m-1}^*$ such that

$$H_1\big(VR_{r_0=0}(X)\big) \xrightarrow{i_0^*} H_1\big(VR_{r_1}(X)\big) \xrightarrow{i_1^*} \cdots \xrightarrow{i_{m-1}^*} H_1\big(VR_{r_m}(X)\big)$$

and $i_n^*([c]_V) = [i_n(c)]_W$ for $V = VR_{r_n}(X)$, $W = VR_{r_{n+1}}(X)$, and all $n = 0, 1, \ldots, m-1$. We now give a name to the "smallest" and "largest" $r > 0$ such that a given 1-cycle belongs to $H_1(VR_r(X))$.

Definition A.4. *Let $[c] \in H_1(VR_r(X))$ for some $r > 0$. The birth filtration of $[c]$ is defined as the greatest lower bound of the set of all $\epsilon > 0$ such that $[c]$ is in the range of the $\mathbb{F}_2$-linear function $H_1(VR_\epsilon(X)) \to H_1(VR_r(X))$. Similarly, the death filtration of $[c]$ is defined as the least upper bound of the set of all $\epsilon > 0$ such that $[c]$ maps to zero under the $\mathbb{F}_2$-linear function $H_1(VR_r(X)) \to H_1(VR_\epsilon(X))$. The persistence of $[c]$ is defined as the difference between the death filtration and the birth filtration.*

Up to a scaling factor in the variable r, the Geometric Čech complex of radius r is homotopy equivalent to the Radius r Vietoris–Rips complex due to the Nerve Lemma (see Corollary 4G.3 in Hatcher) [22]. Consequently, the definitions of the birth and death radius of an equivalence class of non-contractible loops presented in Section 1.1 are equivalent to the definitions of the birth and death filtration of a class $[c] \in H_1(VR_r(X))$ given in Definition A.4. For a more thorough treatment of persistent homology, see [60].

Appendix B. Relative influence of predictor variables in top-performing models

Figures depicting the relative influence of the top-performing models with respect to F1-score in each of the three classification tasks are displayed. Note that for each of the three classification tasks, the top-performing model with respect to F1-score was the Gradient Boosted Decision Tree Model.

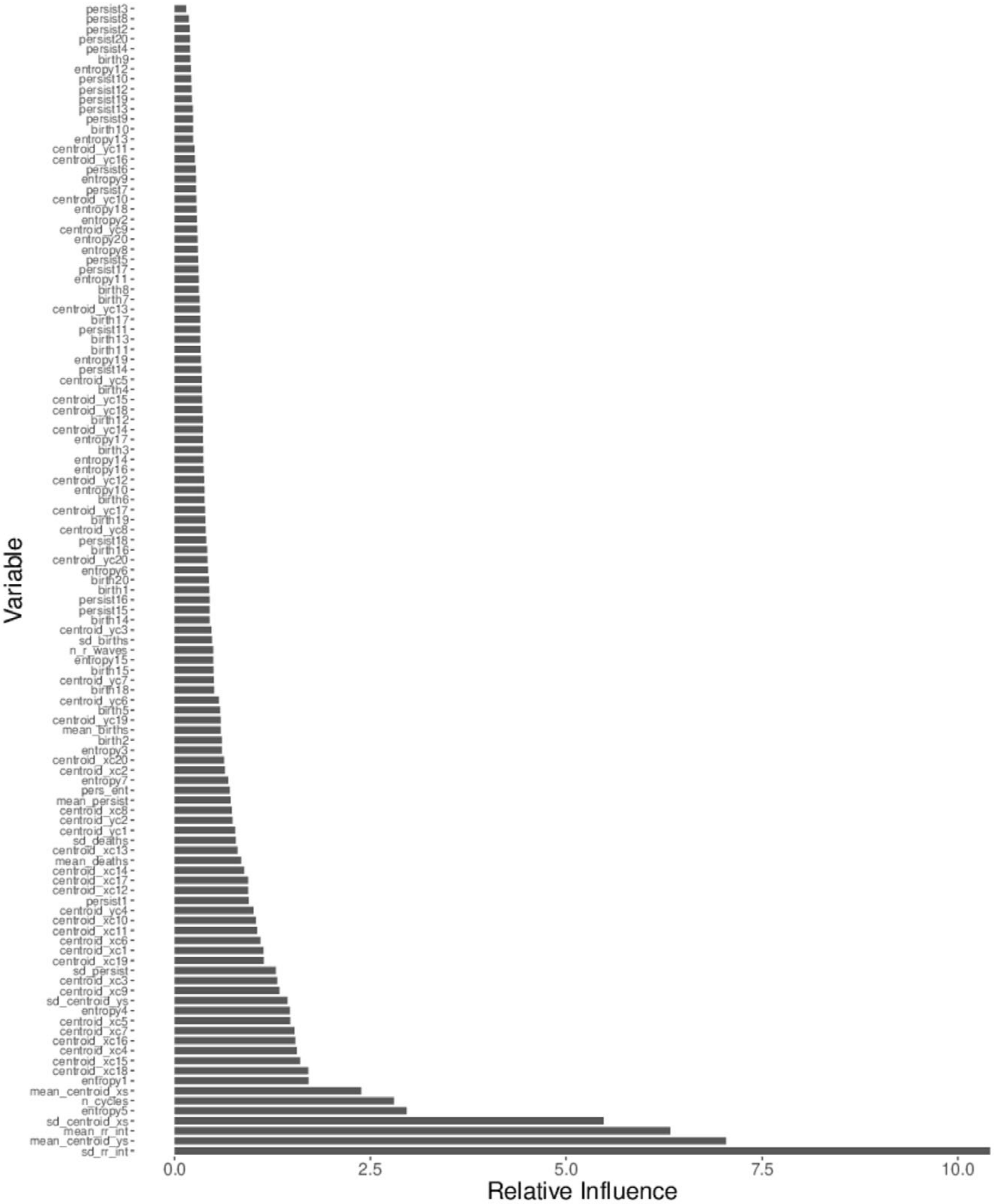

Fig. 11. Relative influence of predictor variables in top-performing gradient boosted decision tree model in classification of atrial fibrillation vs. Non-atrial fibrillation.

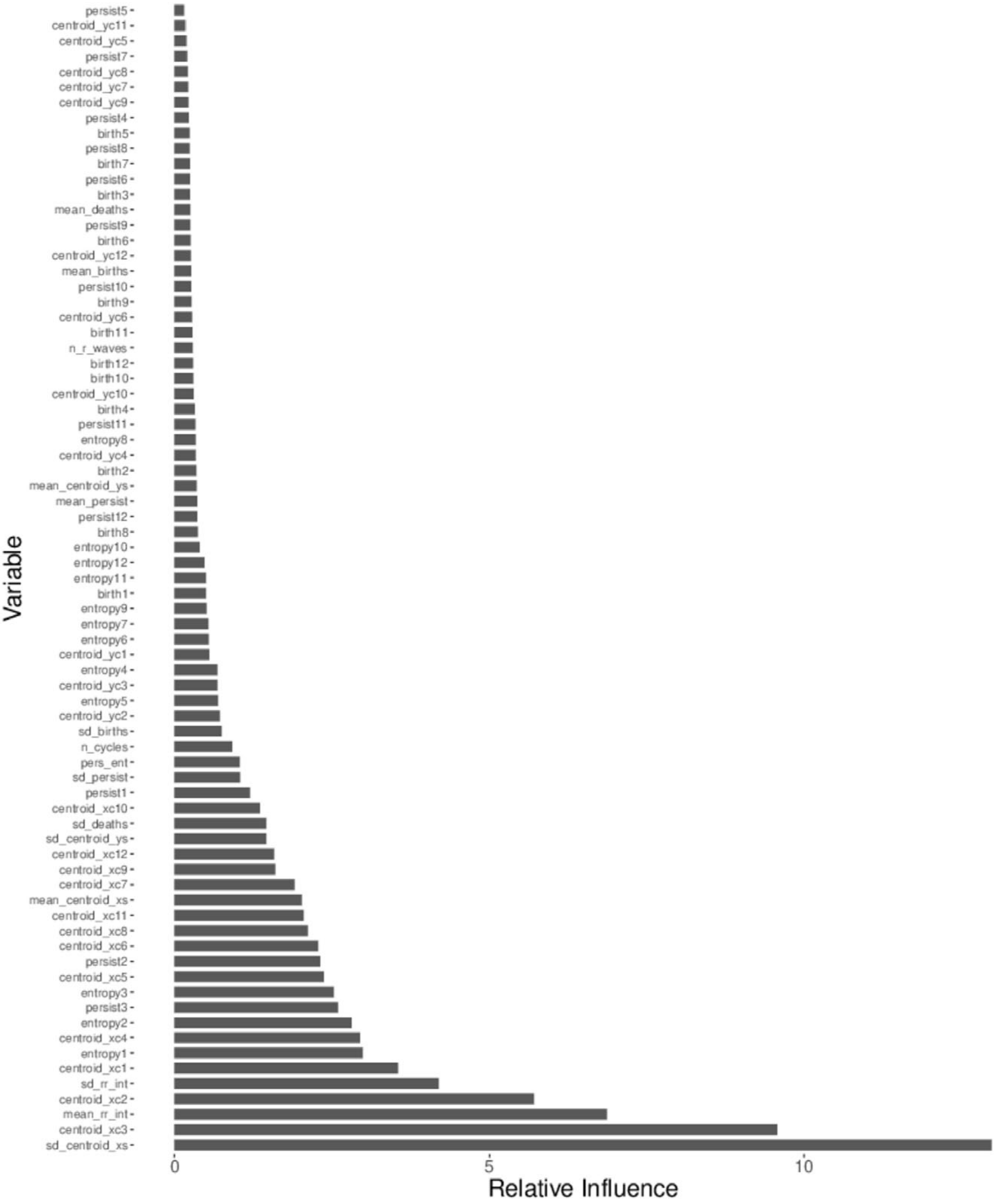

Fig. 12. Relative influence of predictor variables in top-performing gradient boosted decision tree model in classification of arrhythmia vs. Normal sinus rhythm.

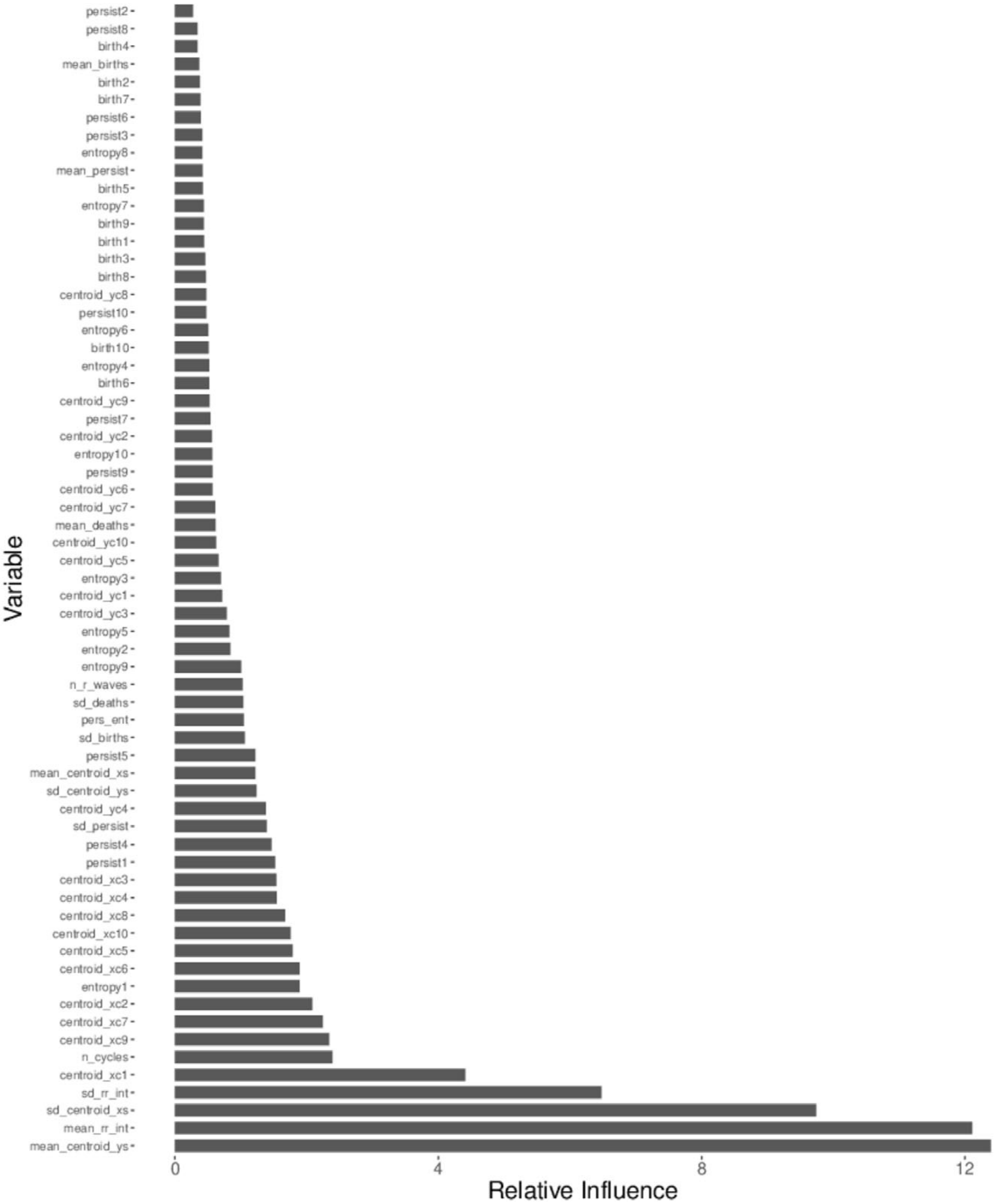

Fig. 13. Relative influence of predictor variables in top-performing gradient boosted decision tree model in classification of arrhythmia with morphological changes vs. Sinus rhythm with bradycardia and tachycardia treated as non-arrhythmia.

References

[1] H. Adams, T. Emerson, M. Kirby, R. Neville, C. Peterson, P. Shipman, S. Chepushtanova, E. Hanson, F. Motta and L. Ziegelmeier, Persistence images: A stable vector representation of persistent homology, *J. Mach. Learn. Res.* **18** (2017), 218–252, https://arxiv.org/abs/1507.06217.

[2] A. Aljanobi and J. Lee, *Topological Data Analysis for Classification of Heart Disease Data, 2021 IEEE International Conference on Big Data and Smart Computing (BigComp)*, Jeju Island, Korea (South), 2021, pp. 210–213. doi:10.1109/BigComp51126.2021.00047.

[3] F. Altındiş, B. Yılmaz, S. Borisenok and K. İçöz, Parameter investigation of topological data analysis for EEG signals, *Biomedical Signal Processing And Control.* **63** (2021), 102196. doi:10.1016/j.bspc.2020.102196.

[4] J. Arsuaga, N. Baas, D. Dewoskin, H. Mizuno, A. Pankov and C. Park, Topological analysis of gene expression arrays identifies high risk molecular subtypes in breast cancer, *Appl. Algebra Eng., Commun. Comput.* **23** (2012), 3–15. doi:10.1007/s00200-012-0166-8.

[5] A. Asgharzadeh-Bonab, M. Amirani and A. Mehri, Spectral entropy and deep convolutional neural network for ECG beat classification, *Biocybernetics And Biomedical Engineering.* **40** (2020), 691–700. doi:10.1016/j.bbe.2020.02.004.

[6] U. Bauer, Ripser: Efficient computation of Vietoris–Rips persistence barcodes, *Journal Of Applied And Computational Topology* (2021), https://arxiv.org/abs/1908.02518.

[7] P.G. Camara, D.I. Rosenbloom, K.J. Emmett, A.J. Levine and R. Rabadan, Topological data analysis generates high-resolution, genome-wide maps of human recombination. *Cell Syst.* **3**(1) (2016), 83–94. doi:10.1016/j.cels.2016.05.008.

[8] Centers for Disease Control and Prevention, Leading causes of death, https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm.

[9] S. Chen, W. Hua, Z. Li, J. Li and X. Gao, Heartbeat classification using projected and dynamic features of ECG signal, *Biomedical Signal Processing And Control.* **31** (2017), 165–173. doi:10.1016/j.bspc.2016.07.010.

[10] Y. Chung, C. Hu, A. Lawson and C. Smyth, Topological approaches to skin disease image analysis, in: *2018 IEEE International Conference on Big Data (Big Data)*, 2018, pp. 100–105. doi:10.1109/BigData.2018.8622175.

[11] Y. Chung, C. Hu, Y. Lo and H. Wu, A persistent homology approach to heart rate variability analysis with an application to sleep-wake classification, *Frontiers In Physiology.* **12** (2021), 202, https://arxiv.org/abs/1908.06856.

[12] S.A. Dee, J. Tucciarone, G. Plotkin and C. Mallilo, Determining the impact of an alarm management program on alarm fatigue among ICU and telemetry RNs: An evidence based research project, *SAGE Open Nurs.* **13**(8) (2022), 23779608221098713. doi:10.1177/23779608221098713.

[13] D.S. Desai and H.S. Arrhythmias, StatPearls [internet]. Treasure Island (FL): StatPearls publishing, 2024. Available from: https://www.ncbi.nlm.nih.gov/books/NBK558923/.

[14] D. DeWoskin, J. Climent, I. Cruz-White, M. Vazquez, C. Park and J. Arsuaga, Applications of computational homology to the analysis of treatment response in breast cancer patients, *Topology And Its Applications.* **157** (2010), 157–164. doi:10.1016/j.topol.2009.04.036.

[15] S. Dhyani, A. Kumar and S. Choudhury, Arrhythmia disease classification utilizing ResRNN, *Biomedical Signal Processing And Control.* **79** (2023), 104160. doi:10.1016/j.bspc.2022.104160.

[16] M. Dindin, Y. Umeda and F. Chazal, Topological data analysis for arrhythmia detection through modular neural networks. *Advances In Artificial Intelligence.* (2020).

[17] F. Elhaj, N. Salim, A. Harris, T. Swee and T. Ahmed, Arrhythmia recognition and classification using combined linear and nonlinear features of ECG signals, *Computer Methods And Programs In Biomedicine.* **127** (2016), 52–63. doi:10.1016/j.cmpb.2015.12.024.

[18] J.H. Friedman, Greedy function approximation: A gradient boosting machine, *Ann. Statist.* **29**(5) (2001), 1189–1232. doi:10.1214/aos/1013203451.

[19] A. Goldberger, L. Amaral, L. Glass, J. Hausdorff, P.C. Ivanov, R. Mark and H.E. Stanley, PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals, *Circulation [Online]* **101**(23) (2000), e215–e220. doi:10.1161/01.cir.101.23.e215.

[20] G. Graff, Persistent homology as a new method of the assessment of heart rate variability, *PLOS ONE* **16** (2021), 1–24. doi:10.1371/journal.pone.0253851.

[21] L. Guo, G. Sim and B. Matuszewski, Inter-patient ECG classification with convolutional and recurrent neural networks, *Biocybernetics And Biomedical Engineering.* **39** (2019), 868–879, https://www.sciencedirect.com/science/article/pii/S0208521618304200. doi:10.1016/j.bbe.2019.06.001.

[22] A. Hatcher, *Algebraic Topology*, Cambridge University Press, ISBN: 9780521795401 2002.

[23] E. Hernández-Lemus, P. Miramontes and M. Martínez-García, Topological data analysis in cardiovascular signals: An overview, *Entropy* **26**(1) (2024), 67. doi:10.3390/e26010067.

[24] Y. Hiraoka, T. Nakamura, A. Hirata, E. Escolar, K. Matsue and Y. Nishiura, Hierarchical structures of amorphous solids characterized by persistent homology, *Proceedings Of The National Academy Of Sciences* **113** (2016), 7035–7040. doi:10.1073/pnas.1520877113.

[25] Y. Huang, H. Li and X. Yu, A novel time representation input based on deep learning for ECG classification, *Biomedical Signal Processing And Control.* **83** (2023), 104628. doi:10.1016/j.bspc.2023.104628.

[26] P. Ignacio, C. Dunstan, E. Escobar, L. Trujillo and D. Uminsky, Classification of single-lead electrocardiograms: TDA informed machine learning, in: *2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2019, pp. 1241–1246. doi:10.1109/ICMLA.2019.00204.

[27] P.S. Ignacio, J.A. Bulauan and J.R. Manzanares, in: *A Topology Informed Random Forest Classifier for ECG Classification, 2020 Computing in Cardiology*, Rimini, Italy, 2020, pp. 1–4. doi:10.22489/CinC.2020.297.

[28] G. James, D. Witten, T. Hastie and R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*, Springer, New York, NY, ISBN: 1461471370, 2021.

[29] K.R. Johnson, J.I. Hagadorn and D.W. Sink, Alarm safety and alarm fatigue, *Clin Perinatol.* **44**(3) (2017), 713–728. Epub 2017 Jul 14. doi:10.1016/j.clp.2017.05.005.

[30] S. Khurshid, S.H. Choi, L.C. Weng, E.Y. Wang, L. Trinquart, E.J. Benjamin, P.T. Ellinor and S.A. Lubitz, Frequency of cardiac rhythm abnormalities in a half million adults. *Circ Arrhythm Electrophysiol.* **11**(7) (2018), e006273. doi:10.1161/CIRCEP.118.006273.

[31] M. Kumar, R. Pachori and U. Rajendra Acharya, Automated diagnosis of atrial fibrillation ECG signals using entropy features extracted from flexible analytic wavelet transform, *Biocybernetics And Biomedical Engineering.* **38** (2018), 564–573. doi:10.1016/j.bbe.2018.04.004.

[32] Y.K. Kutlu, Feature extraction for ECG heartbeats using higher order statistics of WPD coefficients. *Comput Methods Programs Biomed.* (2012).

[33] K. Lewandowska, M. Weisbrot, A. Cieloszyk, W. Mędrzycka-Dąbrowska, S. Krupa and D. Ozga, Impact of alarm fatigue on the work of nurses in an intensive care environment-a systematic review. *Int J Environ Res Public Health.* **17**(22) (2020), 8409. doi:10.3390/ijerph17228409.

[34] T. Ling, Z. Zhu, Y. Zhang and F. Jiang, Early ventricular fibrillation prediction based on topological data analysis of ECG, *Signal. Applied Sciences* **12**(20) (2022), 10370. doi:10.3390/app122010370.

[35] G. Lippi, F. Sanchis-Gomar and G. Cervellin, Global epidemiology of atrial fibrillation: An increasing epidemic and public health challenge. *Int J Stroke* **16**(2) (2021), 217–221. Epub 2020 Jan 19. Erratum in: Int J Stroke. 2020 Jan 28. doi:10.1177/1747493019897870.

[36] Y. Liu, L. Wang and Y. Yan, Persistence landscape-based topological data analysis for personalized arrhythmia classification, in: *2023 IEEE 19th International Conference on Body Sensor Networks (BSN)*, IEEE, 2023, pp. 1–6. doi:10.1109/BSN58485.2023.10331360.

[37] S. Lockwood and B. Krishnamoorthy, Topological features in cancer gene expression data, https://arxiv.org/abs/1410.3198, 2015.

[38] C. Maria, J. Boissonnat, M. Glisse and M. Yvinec, *The Gudhi Library: Simplicial Complexes and Persistent Homology. Mathematical Software – ICMS 2014*, 2014. doi:10.1007/978-3-662-44199-2_28.

[39] W. Midani, W. Ouarda and M. Ayed, DeepArr: An investigative tool for arrhythmia detection using a contextual deep neural network from electrocardiograms (ECG) signals, *Biomedical Signal Processing And Control.* **85** (2023), 104954. doi:10.1016/j.bspc.2023.104954.

[40] I. Migdady, A. Russman and A.B. Buletko, Atrial fibrillation and ischemic stroke: A clinical review. *Semin Neurol.* **41**(4) (2021), 348–364. Epub 2021 Apr 13. doi:10.1055/s-0041-1726332.

[41] A. Mjahad, J.V. Frances-Villora, M. Bataller-Mompean and A. Rosado-Muñoz, Ventricular fibrillation and tachycardia detection using features derived from topological data analysis, *Appl. Sci.* **12** (2022), 7248. doi:10.3390/app12147248.

[42] G.B. Moody and R.G. Mark, The impact of the MIT-BIH arrhythmia database. *IEEE Eng in Med and Biol* **20**(3), 45–50. doi:10.1109/51.932724.

[43] Y. Ni, F. Sun, Y. Luo, Z. Xiang and H. Sun, A novel heart disease classification algorithm based on fourier transform and persistent homology, 2021.

[44] 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, *Proc Natl Acad Sci USA* **108**(17) (2011), 7265–7270, Epub 2011 Apr 11.. doi:10.1073/pnas.1102826108.

[45] I. Obayashi, Volume-optimal cycle: Tightest representative cycle of a generator in persistent homology, *SIAM Journal On Applied Algebra And Geometry.* **2** (2018), 508–534. doi:10.1137/17M1159439.

[46] S. Oh, E. Ng, R. Tan and U. Acharya, Automated diagnosis of arrhythmia using combination of CNN and LSTM techniques with variable length heart beats, *Computers In Biology And Medicine.* **102** (2018), 278–287. doi:10.1016/j.compbiomed.2018.06.002.

[47] D.W. Ormrod Morley, Persistent homology in two-dimensional atomic networks. *J Chem Phys.* (2021). doi:10.1063/5.0040393.

[48] B. Pyakillya, N. Kazachenko and N. Mikhailovsky, Deep learning for ECG classification, *Journal Of Physics: Conference Series.* **913** (2017), 012004. doi:10.1088/1742-6596/913/1/012004.

[49] T. Qaiser, K. Sirinukunwattana, K. Nakane, Y. Tsang, D. Epstein and N. Rajpoot, Persistent homology for fast tumor segmentation in whole slide histology images, *Procedia Computer Science.* **90** (2016), 119–124, 20th Conference on Medical Image Understanding and Analysis (MIUA 2016). doi:10.1016/j.procs.2016.07.033.

[50] R. Rabadán, Y. Mohamedi, U. Rubin, T. Chu, A.N. Alghalith, O. Elliott, L. Arnés, S. Cal, Á.J. Obaya, A.J. Levine and P.G. Cámara, Identification of relevant genetic alterations in cancer using topological data analysis. *Nat Commun.* **11**(1) (2020), 3808. doi:10.1038/s41467-020-17659-7.

[51] M. Rahhal, Y. Bazi, H. AlHichri, N. Alajlan, F. Melgani and R. Yager, Deep learning approach for active classification of electrocardiogram signals, *Information Sciences* **345** (2016), 340–354. doi:10.1016/j.ins.2016.01.082.

[52] O. Rainio, J. Teuho and R. Klén, Evaluation metrics and statistical tests for machine learning. *Sci Rep.* **14**(1) (2024), 6086. doi:10.1038/s41598-024-56706-x.

[53] Y. Ren, F. Liu, S. Xia, S. Shi, L. Chen and W.Z. Dynamic, ECG signal quality evaluation based on persistent homology and GoogLeNet method. *Front Neurosci.* **17** (2023), 1153386. doi:10.3389/fnins.2023.1153386.

[54] K.J. Ruskin and D. Hueske-Kraus, Alarm fatigue: Impacts on patient safety. *Curr Opin Anaesthesiol.* **28**(6) (2015), 685–690. doi:10.1097/ACO.0000000000000260.

[55] B. Safarbali and S. Hashemi Golpayegani, Nonlinear dynamic approaches to identify atrial fibrillation progression based on topological methods, *Biomedical Signal Processing And Control.* **53** (2019), 101563. doi:10.1016/j.bspc.2019.101563.

[56] G. Sannino and G. De Pietro, A deep learning approach for ECG-based heartbeat classification for arrhythmia detection, *Future Generation Computer Systems.* **86** (2018), 446–455. doi:10.1016/j.future.2018.03.057.

[57] L. Seemann, J. Shulman and G.H. Gunaratne, A robust topology-based algorithm for gene expression profiling. *ISRN Bioinform.* **2012** (2012), 381023. doi:10.5402/2012/381023.

[58] S. Sendelbach and M. Funk, Alarm fatigue: A patient safety concern. *AACN Adv Crit Care.* **24**(4) (2013), 378–386. quiz 387-8. doi:10.1097/NCI.0b013e3182a903f9.

[59] J. Storm and H.C. Chen, The relationships among alarm fatigue, compassion fatigue, burnout and compassion satisfaction in critical care and step-down nurses. *J Clin Nurs.* **30**(3–4) (2021), 443–453. Epub 2020 Nov 28. doi:10.1111/jocn.15555.

[60] Y. Tamal Dey, *Computational Topology for Data Analysis*, Cambridge University Press, ISBN: 978-1009098168, https://www.cs.purdue.edu/homes/tamaldey/book/CTDAbook/CTDAbook.pdf, 2021.

[61] G. Wang, C. Zhang, Y. Liu, H. Yang, D. Fu, H. Wang and P. Zhang, A global and updatable ECG beat classification system based on recurrent neural networks and active learning, *Information Sciences* **501** (2019), 523–542. doi:10.1016/j.ins.2018.06.062.

[62] J. Wang, Automated detection of atrial fibrillation and atrial flutter in ECG signals based on convolutional and improved Elman neural network, *Knowledge-Based Systems.* **193** (2020), 105446. doi:10.1016/j.knosys.2019.105446.

[63] T. Wang, T. Johnson, J. Zhang and K. Huang, Topological methods for visualization and analysis of high dimensional single-cell RNA sequencing data, *Pac Symp Biocomput.* **24** (2019), 350–361, PMCID: PMC6417818. PMID:30963074.

[64] World Health Organization, The top 10 causes of death, https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death.

[65] C.C. Ye, Heartbeat classification using morphological and dynamic features of ECG signals. *IEEE Trans Biomed Eng.* (2012). doi:10.1109/TBME.2012.2213253.

[66] Ö. Yildirim, A novel wavelet sequence based on deep bidirectional LSTM network model for ECG signal classification, *Computers In Biology And Medicine.* **96** (2018), 189–202. doi:10.1016/j.compbiomed.2018.03.016.

[67] O. Yildirim, U. Baloglu, R. Tan, E. Ciaccio and U. Acharya, A new approach for arrhythmia classification using deep coded features and LSTM networks, *Computer Methods And Programs In Biomedicine.* **176** (2019), 121–133. doi:10.1016/j.cmpb.2019.05.004.

[68] Ö. Yıldırım, P. Pławiak, R. Tan and U. Acharya, Arrhythmia detection using deep convolutional neural network with long duration ECG signals, *Computers In Biology And Medicine.* **102** (2018), 411–420. doi:10.1016/j.compbiomed.2018.09.009.

[69] J. Zheng, J. Zhang, S. Danioko et al., A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients, *Sci Data* **7** (2020), 48. doi:10.1038/s41597-020-0386-x.

[70] W. Zong, G. Moody and D. Jiang, A robust open-source algorithm to detect onset and duration of QRS complexes, *Computers In Cardiology* **2003** (2003), 737–740. doi:10.1109/CIC.2003.1291261.