Detecting PVC Beats by Beat-by-beat Analysis of ECG Signals Using Machine Learning Classifiers for Real-time Predictive Cardiac Health Monitoring

I Hua Tsai, *Student Member, IEEE*, and Bashir I. Morshed, *Senior Member, IEEE*

Department of Computer Science, Texas Tech University, Lubbock, TX 79409, USA

***Abstract*— Premature Ventricular Contraction (PVC) episodes are redundant heartbeats that disrupt the normal rhythm of the heart. The use of wearable sensors for remote heart monitoring and the implementation of trusted artificial intelligence (AI) algorithms are improvements in the field of smart health (sHealth) using cyber-physical systems (CPS) for telemedicine systems. We detect PVC beats by analyzing electrocardiogram (ECG/EKG) data and perform automatic classification to achieve high accuracy in real-time. In this study, we used a number of PVC heartbeat recordings from the MIT BIH supraventricular arrhythmia database We divided the recordings into a training dataset, which contains 39 ECG data, and a test dataset, which contains the remaining 39 ECG data. Both datasets contain approximately 80,000 samples of normal heartbeats and 7,000 samples of ventricular ectopic We extract combination of signal-specific features and signal-independent features for feature selection and ranking. We apply four algorithms, receiver operator characteristic (ROC) and the area under the ROC curve (AUC) (ROCAUC), constant, quasi constant and duplicate feature removal (Univariate) (CQCDFR), analysis of variance (ANOVA), and root mean square deviation (RMSE) to select and rank the feature. For each algorithm, it has its own selection of signal-independent features, which we combine separately with signal-specific features and test their accuracy. Then, we train the top 10 ranked combined features of each algorithm separately and check the highest performance. We explored the random forest (RF) classifier and support vector machine (SVM) classifier. Compared with other algorithms, the performance of feature selection using ANOVA algorithm before feature ranking is the lowest. The ANOVA algorithm achieved the highest accuracy after picking out the top 10 features. We further separately evaluate the accuracy, sensitivity, precision, specificity and F1 score of the top-ranked features according to the best accuracy obtained by different feature selection algorithms. The classification ANOVA algorithm from RF selects the top 7 features with 97% accuracy, 97.5 sensitivity, 98.1% specificity, 98.1 Precision%, and 95.0% F1 Score. This method can accurately monitor cardiac disease in real-time and analyze ECG beats so that patients can get accurate feedback.**

Keywords- ECG classification, Cardiac episodes, unsupervised monitoring, machine learning.

I. Introduction

Cyber-physical systems (CPS) are used to control and monitor embedded computers of cyber-physical processes, often using feedback loops [1]. A small wearable monitoring system can transmit the detected premature ventricular contractions (PVCs) in the ECG signal to the central control system for effective inspection. When an abnormality occurs, the control system can issue an alarm to the actuation system. CPS can detect hidden heart disease early and provide an effective program. Different wearables are available to capture ECG signal from different lead positions [2]. Today's ECG devices are small in size to allow for continuous, remote monitoring of health conditions [3].

The heart's extra heartbeat starts in one of the two lower pumping chambers (the ventricles) called a premature ventricular contraction (PVC). These extra heartbeats can disrupt the normal heart rhythm. They can occur in a variety of people, including those without structural heart disease and those with all types and degrees of heart disease. Patients with premature ventricular contractions usually have no specific symptoms, but may present with palpitations, chest tightness, dizziness, fatigue and other symptoms. In severe cases, angina pectoris, hypotension, and heart failure may also occur. An electrocardiogram (ECG) can detect premature ventricular beats and discover type and origin [4-14]. Identifying these disorders from raw electrocardiogram (ECG) signals is tedious and expensive [4]. Smart method for automatic heartbeat classification in ECG recordings will be of great help to clinicians in diagnosing heart disease [5]. Automated analysis techniques rely on machine learning (ML) models trained to identify PVCs.

Stationary wavelets transform (SWT) is used for peak-to-average power ratio and logarithmic energy entropy to extract features and detect with support vector machine (SVM). [6]. ECG features based on morphological features, temporal features, and peak-top features use machine learning classification methods such as decision trees, random forests, and gradient boosted trees (GDB). [7]. Advances in electrocardiogram (ECG) data and dynamic neural networks

Table I: The criteria of AAMI labeling class of MIT-BIH Supraventricular Arrhythmia Database with the full database, and training and testing dataset.

AAMI Heartbeat class	**N**	**S**	**V**	**F**	**Q**	**Total**
Description	Normal beat	Supraventricular ectopic beat (SVEB)	Ventricular ectopic beat (VEB)	Fusion beat	Unknown beat	
Label	N, L, R	S, e, j, A, a, J	V, E	F	Q, /, f	
Full data *(Total 78 records)*	162,338	8,990	9,441	23	3,223	184,555
Training *(First 39 records)*	83,292	1,252	7,397	5	1,090	93,036
Testing *(Rest 39 records)*	79,046	7,738	7,044	18	2,133	95,979

(DNNs), particularly algorithms such as long short-term memory (LSTM) and convolutional neural networks (CNN), are based on the features used. [8]. Develop methods for different features, including RR beat interval, time domain, frequency domain, and distribution features, and evaluate the efficiency of the method with linear classifiers, SVMs, and quadratic neural networks. [9]. Use an ML classifier to achieve the performance of feature selection techniques and perform feature selection on AF morphological features and heart rate variability [10].

Identifying features that convert a 1D ECG into a 2D form through short-time Fourier transform and wavelet transform use SVM to detect PVCs. [11]. Trained using deep admixture feature selection extracted from the last fully connected layer of ten CNNs. These classifiers involve Support Vector Machines (SVM), Random Forests (RF), K-Nearest Neighbors (KNN), Linear Discriminant Classifiers (LDA), Quadratic Discriminant Analysis (QDA), and Decision Trees (DT) [12]. Detect impending heart disease using Naive Bayes, Artificial Neural Networks, Support Vector Machines, Random Forests, Simple Logistic Regression, and accompanying physiological features. [13]. Electrocardiograms (ECGs) combined with multiple support vector machines (SVMs) rely on time intervals and their feature automatic classification methods for ECG characterization. [14].

This work is a further extension and study of previous tasks. In the previous task, we used hand-selected signal-specific features for classification and took 80%, 20% to form training and test sets. This results in an imbalance of our training and test data for normal and abnormal heartbeats and an inability to determine the effectiveness of specific feature selection. We form more explicit training and testing data in order to classify normal and abnormal heartbeats. Detect PVC heartbeats and increase the number of features as well as feature selection and ranking to test the accuracy of the top-ranked features to find the best performance. We measure the statistical performance metrics of the top performers using random forest classifiers and support vector machines. We form feature vectors for classification using signal-specific features and signal-independent features. The focus of this work is to analyze each heartbeat and apply it to wearable devices through an efficient method. We select effective features and minimize the number of features used to reduce the power consumption of the monitoring device. Our special real-time beat-to-beat classification method enables efficient analysis of ECG data from wearable heart monitors.

II. Methodology

This study used MIT BIH database in which criteria of the AAMI to classify heartbeat types [15]. There are five heartbeat categories that MIT-BIH uses for subsequent processing. Each category includes multiple types of heartbeats, as shown in Table I. Normal and bundle branch block heartbeat present in class N. Supraventricular ectopic heartbeat (SVEB) present in class S. Ventricular ectopic beats (VEB) present in class present in class V. Fusion present in class F. Unknown beat present in class Q. We accessed the PhysioNet database from the open source of public data. We pick up MIT BIH supraventricular arrhythmia database (svdb) to achieve our goal [16]. The beats we classify use PVC beats and normal beats. Figure 1. shows the flow diagram of process to classify PVC beat.

A. ECG data

In our work, we used a number of PVC heartbeat recordings from the MIT BIH supraventricular arrhythmia database [16]. The database contains two lead ECG signals. Total of 78 records of duration around 30 minutes. Data has 128 Hz sampling rate and filter with bandpass. This study used the AAMI criteria to classify MIT BIH heartbeat types. The heartbeat types listed on Table I. The normal beat (N) has the largest category with over 162,338 samples, and the fusion beat (F) has the smallest category with only around 23 samples. The number of individual heartbeat types varies widely. We selected only normal beats (N) and ventricular ectopic beats (V) for analysis. In addition to this, we removed supraventricular ectopic beats (SVEB), fused beats (F), and unknown beats (Q) from the analysis [17]. The record we divided into a training dataset, and a testing dataset. The training dataset contains first of 39 ECG data records, and the test dataset contains rest of 39 ECG data records. Both datasets contain approximately 80,000 samples of normal heartbeat and 7,000 samples of ventricular ectopic beat. The performance of the classifier is evaluated using the training dataset. Final performance evaluation of the heartbeat classification system on the test dataset. Table I shows the breakdown of each dataset by heartbeat type.

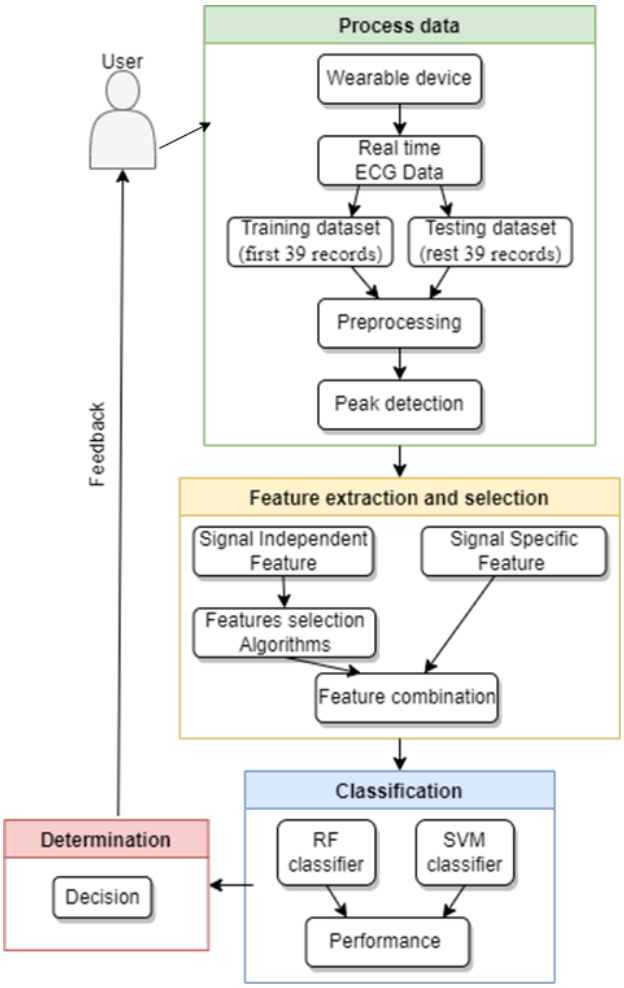


Fig. 1: Flowchart of ECG signal demonstrating machine learning algorithm for preprocessing, analysis, and classification.

B. Signal preprocessing

The human electrocardiogram (ECG) signal is a weak physiological signal with nonlinear, non-stationary and strong randomness. The main noise of electrocardiogram (ECG) signal includes baseband drift, EMG interference, power frequency noise, other noise interference, etc. This noise and distortion results in a low signal-to-noise ratio (SNR). The waveform of the disturbed ECG signal will be deformed, which will affect the doctor's recognition of the ECG. The frequency of the baseline noise is relatively low, and the ECG signal itself contains very rich low-frequency signals, so a low-pass filter is used to remove the baseline drift. The traditional methods for removing baseline drift include median filtering, wavelet transform, algorithm average filtering, and EMD decomposition.

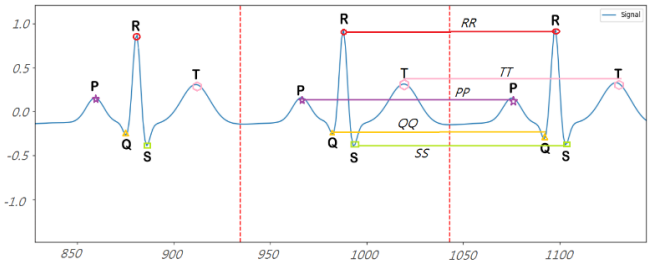


Fig. 2: P, Q, R, S, T peaks in detail to form a heartbeat.

C. Peak detection

Identifying and detecting R peaks occupies a very important and necessary position in ECG heartbeat recognition. Figure 2 shows P, Q, R, S, T peaks and a heartbeat. In this work we detect the exact position of the R peak using the pan_tompkin algorithm [18]. The correct position of the P, Q, S, T, T′ peaks is affected by the detection accuracy of the R peak. We detected R peaks using the Pan Tompkin algorithm in the open-source code "The Python Toolbox for Neurophysiological Signal Processing" provided by neurokit 2 [19]. When the position of the R peak is detected, we can calculate the RR interval and the mean RR interval. RR interval and average RR interval can be used to find the P, Q, S, T and T' peaks. We find three R peaks to form the heartbeat. As the starting point of the beat, we use the middle value between the first R peak and the second R peak to present. As the end point of the beat, we use the middle value between the second R peak and the third R peak as the presentation. See Table II for specifications.

Table II: Standard of P, Q, S, T, T' peak and one heartbeat detection.

Peak	Standard
P	Max of 3/8*RR* before *RQ* interval
Q	Min 1/8 *RR* before *RR* interval
S	Min of *R* to 1/4*RR* before *R* peak
T	1/4*RR* to max of 3/8*RR*
T'	1/4*RR* to min of 3/8*RR*
One heartbeat	Min of 1/2*RR* to max of 1/2*RR*

D. Feature extraction and selection

The feature extraction process takes two ECG beats at a time, and slides one beat when the process finishes extracting (i.e. that beat runs the next beat). We extract combination of signal-specific features and signal-independent features for feature selection and ranking. We obtain a total of 7 amplitude features, 6 frequency features and 15 statistical features in the signal specific features. The total of 175 signal independent features be extracted using Time series feature extraction library (TSFEL) which is a python package [20]. We apply four different algorithms to do the feature selection and ranking. Algorithms we use receiver operator characteristic (ROC) and the area under the ROC curve (AUC) (ROCAUC), constant, quasi constant and duplicate feature removal (Univariate) (CQCDFR), analysis of variance (ANOVA), and root mean square deviation (RMSE). For each algorithm, it has its own choice of signal-independent features, which we combine with signal-specific features, separately. We use the combined features from each algorithm to make the top 10 features for classification. Table III and IV shows the summaries of signal specific features and signal independent features and the top 10 ranked features based on different feature selection algorithms in detail. Due to its advantages in handling imbalanced and cost-sensitive data, ROC curve and AUC (area under the ROC curve) are used to determine classification accuracy in supervised learning. Constant, quasi constant and duplicate feature removal (Univariate) (CQCDFR) methods rank individual features according to certain criteria and then select the top N features. Different types of ranking criteria are used for univariate filtering methods. Analysis of variance (ANOVA) is used to analyze differences between group means in a sample. A collection of statistical models and their associated estimation procedures. The root mean square error (RMSE) is the standard deviation of the residuals (prediction error). RMSE measures how distributed these residuals are and how concentrated the data are around the line of best fit.

E. ML classification

We label ventricular ectopic beats (V) as 1, and label normal heartbeats (N) as 0. We train the machine learning model of random forest (RF) and support vector machine (SVM) with use a 10-fold cross-validation technique to do the classification. The training and testing datasets had similar proportions of numbers of ventricular ectopic and normal heartbeat types. Training datasets include 90,689 sample intervals and testing datasets include 86,090 sample intervals of normal heartbeats and ventricular ectopic beats, separately. Table I shows that each dataset is divided by heartbeat type. By using RF and SVM classifier, we first classify combination of signal specific features and signal independent features based on different feature selection algorithms to check the performance, separately. After that, we use four different algorithms to do the feature ranking for identify useful features which can get more precise result. In each of algorithm, we run the machine learning model for training and testing to classify top 10 ranked features. We train the first-ranked feature for first run. After training completed, we select the second-ranked feature and add with the first-ranked feature together for second test. After training completed, we select a third-ranked feature and add with the first-ranked feature the second-ranked feature together for third test, and so on. Until tenth test, we train all top 10 ranked features together.

F. Performance

We use various statistical metrics to exhibit model performance and observe performance evaluations. We evaluated the sensitivity, specificity, accuracy, precision, and

Table III: Signal specific features and signal independent features used.

Feature type	Features summary
Signal Specific Feature	***Total of 28 features.***
7 amplitude features, 6 frequency features, 15 statistical features	Q peak, S peak, T peak, R peak, QRS position, QT length, RR length. (amplitude features) heart rate, frequency of Q, R, S and T, instance heart rate. (frequency features) mean peak value of R, Q, S, T and QT, mean value of LF and HF, ratio of LF/HF, maximum frequency, very low frequency, stationary wavelet transforms, PNN50, index of sympathetic and parasympathetic modulation of the autonomic nervous system, root mean square of successive differences. (statistical features)
Signal Independent Feature	**Total 175 features be extracted**
Time series feature extraction library (Tsfel)	histogram, signal distance, ECDF percentile count, empirical cumulative distribution function (ECDF), root mean square, slope, wavelet energy, wavelet entropy, spectral centroid, autocorrelation, max power spectrum, mean absolute deviation, spectral skewness…, and so on.

Table IV: Extract the top 10 ranked signal-independent features according to the algorithm used.

Signal independent feature (base on algorithms)	Features summary
Receiver operator characteristic (ROC) and the area under the ROC curve (AUC) (ROCAUC)	***Total of 115 features be selected***
Top 10 Signal-Specific and Signal-Independent Features (in order)	empirical cumulative distribution function (ECDF), root mean square, slope, wavelet energy, autocorrelation, absolute energy, max power spectrum, mean absolute deviation, spectral skewness, median
Root mean square error (RMSE)	***Total of 115 features be selected***
Top 10 Signal-Specific and Signal-Independent Features (in order)	fast Fourier transform mean coefficient, power bandwidth, histogram, signal distance, empirical cumulative distribution function (ECDF), root mean square, autocorrelation, neighborhood peaks, median diff, total energy, percentile count
Constant, quasi constant and duplicate feature removal (Univariate) (CQCDFR)	***Total of 62 features be selected***
Top 10 Signal-Specific and Signal-Independent Features (in order)	ECDF percentile count, root mean square, max power spectrum, mean absolute deviation, negative turning points, neighborhood peaks, peak to peak distance, wavelet standard deviation, wavelet variance, fast Fourier transform mean coefficient
Analysis of variance (ANOVA)	***Total of 115 features be selected***
Top 10 Signal-Specific and Signal-Independent Features (in order)	root mean square, max power spectrum, mean absolute deviation, power bandwidth, spectral distance, median absolute diff, median diff, spectral entropy, absolute energy, linear prediction cepstral coefficients (LPCC)

F1 score for combination of signal specific features and signal independent features base on different selection algorithms to check the classification performance. We also evaluate the top-ranked features separately based on the best accuracy obtained by different feature selection algorithms. Sensitivity indicates that the actual positives be correctly identified. Specificity means correct identification of a measure of the true proportion of negatives. Precision represents a measure of recognition accuracy. Accuracy means determining how close a measured value is to the true value. The F1 score represents a test accuracy.

III. Results

The sliding window runs every two beat intervals and moves one beat at a time until it ends. We extract R peaks using a sliding window. Use the Pan-Tompkin algorithm to find R-points. When the R point is detected, we can find the P, Q, S, T points. Figure 3 shows the top of the R peak with a dashed purple line. The two R peaks detected for two ECG beats and located at sample interval 1100 and 1207, respectively. After we detected the R peak, we can find P, Q, S, and T points. Table II is the P, Q, S, T, T' peak detection standards. Number 0 represents the T peak position and mark as the purple dashed line, number 1 represents the P peak position and mark as the green dashed line, number 2 represents the Q peak position and mark as the yellow dashed line, and number 3 represents the S peak position and mark as the red dashed line. The sampling interval is from 1050 to 1250. Figure 4 shows the top of P, Q, S, T points run every two ECG beats.

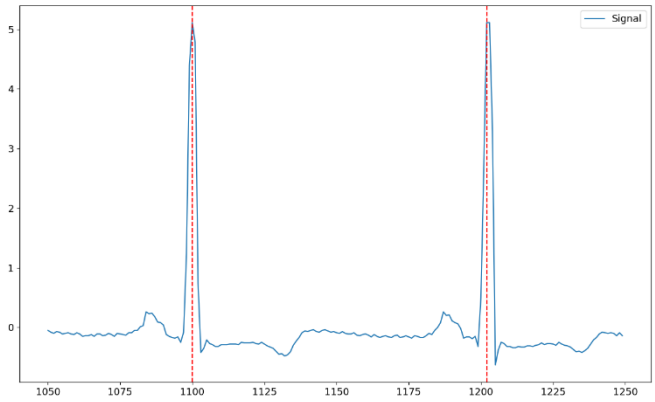


Fig. 3: R peak detected for 2 ECG beats.

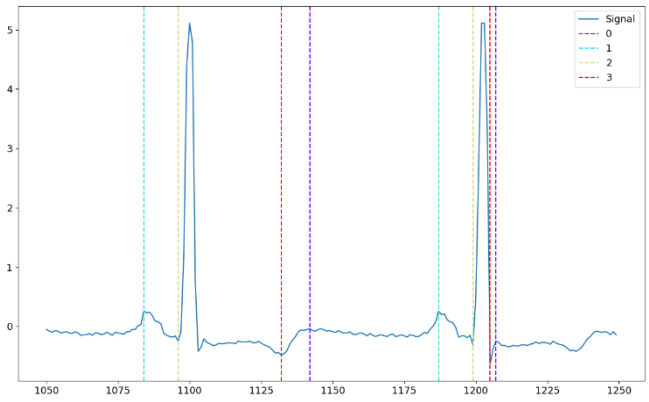


Fig. 4: P, Q, S, T peaks detected for 2 ECG beats

We divided the data of all patients into training and testing datasets. We test a combination of signal-specific features and signal-independent features using random forest (RF) and support vector machine (SVM) classifiers.

ROCAUC algorithm extract the total of 175 signal independent features. There are 115 features be extracted and combined with the signal specific feature to do the classification. By using RF, we get 81.8% sensitivity, 79.2% specificity, 78% accuracy, 80.0% precision, and 82.3% F1 score. By using SVM, we get 78.2% sensitivity, 74.1%, 72% accuracy, specificity, 74.2% precision, and 77.5% F1 score. RMSE algorithm extract the total of 175 signal independent features. There are 115 features be extracted and combined with the signal specific feature to do the classification. By using RF, we get 77.3% sensitivity, 75.2% specificity, 72% accuracy, 75.3% precision, and 76.8% F1 score. By using SVM, we get 80.3% sensitivity, 79.1% specificity, 76% accuracy, 79.3% precision, and 80.1% F1 score.

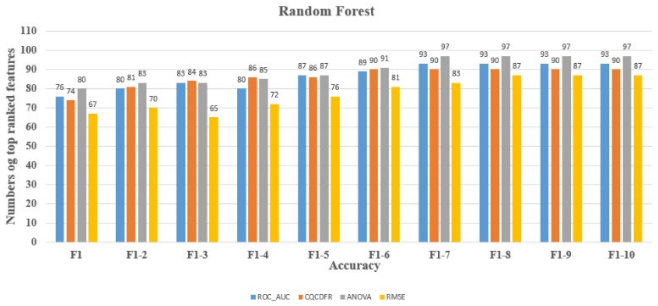
CQCDFR algorithm extract the total of 175 signal independent features. There are 62 features be extracted and combined with the signal specific feature to do the classification. By using RF, we get 77.1% sensitivity, 75.4% specificity, 77% accuracy, 76.2% precision, and 78.3% F1 score. By using SVM, we get 83.7% sensitivity, 82.1% specificity, 78% accuracy, 82.2% precision, and 84.7% F1 score. ANOVA algorithm extract the total of 175 signal independent features. There are 115 features be extracted and combined with the signal specific feature to do the classification. By using RF, we get 74.2% sensitivity, 75.2% specificity, 75% accuracy, 74.0% precision, and 77.3% F1 score. By using SVM, we get 84.1% sensitivity, 80.5% specificity, 77% accuracy, 80.5% precision, and 85.2% F1 score. Table V shows the comparison of random forest (RF) and support vector machine (SVM) performance based on different feature extraction algorithms for the signal specific feature and the signal independent feature combination.

Table V: Comparison of random forest (RF) and support vector machine (SVM) performance based on different feature selection algorithms by using signal specific features and signal independent features combination.

Performent	***ML***	**Sensitivity (%)**	**Specificity (%)**	**Accuracy (%)**	**Precision (%)**	**F1 Score (%)**
ROCAUC						
Total of 115 features	*RF* *SVM*	81.8 78.2	79.2 74.1	78 72	80.0 74.2	82.3 77.5
RMSE						
Total of 115 features	*RF* *SVM*	77.3 80.3	75.2 79.1	72 76	75.3 79.3	76.8 80.1
CQCDFR						
Total of 62 features	*RF* *SVM*	77.1 83.7	75.4 82.1	77 78	76.2 82.2	78.3 84.7
ANOVA						
Total of 115 features	*RF* *SVM*	74.2 84.1	75.2 80.5	75 77	74.0 80.5	77.3 85.2

ROCAUC: Receiver Operator Characteristic (ROC) and the area under the ROC curve (AUC). CQCDFR: Constant, quasi constant and duplicate feature removal (Univariate). ANOVA: Analysis of variance. RMSE: Root mean square deviation.

We combined the signal specific feature and the signal independent feature to pick up top 10 ranked features and find the performance of each selection algorithm by using RF and SVM. ROCAUC, RMSE, CQCDFR and AVONA algorithms are the method to select and rank useful features which can get more precise accuracy. In each training of RF classifier, the ROCAUC algorithm achieves improved accuracy in the first 7 training runs. The accuracy remains the same after 7 features training. The highest accuracy 93%, located on top of 7 features ranked. The CQCDFR algorithm achieves improved accuracy in the first 6 training runs. The accuracy remains the same after 6 features training. The highest accuracy 90%, located on top of 6 features ranked. The ANOVA algorithm achieves improved accuracy in the first 7 training runs. The accuracy remains the same after 7 features training. The highest accuracy 97%, located on top of 7 features ranked. The RMSE algorithm achieves improved accuracy in the first 8 training runs. The accuracy remains the same after 8 features training. The highest accuracy 87%, located on top of 8 features ranked. Figure 5 presents the accuracy of random forest classifier for top 10 ranked based on of different selection and ranking algorithms by using signal specific feature and the signal independent feature. The blue bar presents the accuracy of RF based on ROCAUC algorithm. The orange bar presents the accuracy of RF based on CQCDFR algorithm. The grey bar presents the accuracy of RF based on ANOVA algorithm. The yellow bar presents the accuracy of RF based on RMSE algorithm.



F1: First feature, F1-2: First and second features together, F1-3: First, second and third features together,…F1-10: First, second,…tenth features together. SVM: Support Vector Machine. ROCAUC: Receiver Operator Characteristic (ROC) and the area under the ROC curve (AUC). CQCDFR: Constant, quasi constant and duplicate feature removal (Univariate). ANOVA: Analysis of variance. RMSE: Root mean square deviation.

Fig. 5: Comparison the accuracy of random forest classifier for top 10 ranked based on of different selection and ranking algorithms by using signal specific feature and the signal independent feature.

Table VI summaries the performance of RF for top ranked features based on different feature selection algorithms. The performance evaluation has sensitivity, specificity, accuracy, precision, and F1. In each training of RF classifier, the ROCAUC algorithm achievethe highest accuracy located on top 7 features train. We select top 7 features to do the performance evalution. We achieved 95.1% sensitivity, 91.4% specificity, 93% accuracy, 91.3% precision, and 93.8% F1 score. The CQCDFR algorithm achievethe highest accuracy located on top 6 features train. We select top 6 features to do the performance evalution. We achieved 91.5% sensitivity, 90.4% specificity, 90% accuracy, 90.3% precision, and 90.4% F1 score. The ANOVA algorithm achievethe highest accuracy located on top 7 features train. We select top 7 features to do the performance evalution. We achieved 97.5% sensitivity, 98.1% specificity, 97% accuracy, 98.1% precision, and 95.0% F1 score. The RMSE algorithm achievethe highest accuracy located on top 8 features train. We select top 8 features to do the performance evalution. We achieved

88.2% sensitivity, 89.3% specificity, 87% accuracy, 89.3% precision, and 88.7% F1 score.

Table VI: Summary of RF (random forest) performance for top ranked features based on different feature selection algorithms.

RF Performance	Sensitivity (%)	Specificity (%)	Accuracy (%)	Precision (%)	F1 Score (%)
ROCAUC					
top 6 features	94.4	90.2	89	90.1	92.6
top 7 features	**95.1**	**91.4**	**93**	**91.3**	**93.8**
top 8 features	95.1	91.4	93	91.3	93.8
CQCDFR					
top 5 features	88.6	87.8	86	88.2	87.2
top 6 features	**91.5**	**90.4**	**90**	**90.3**	**90.4**
top 7 features	91.5	90.4	90	90.3	90.4
ANOVA					
top 6 features	90.7	94.4	91	92.7	92.1
top 7 features	**97.5**	**98.1**	**97**	**98.1**	**95.0**
top 8 features	97.5	98.1	97	98.1	95.0
RMSE					
top 7 features	84.1	84.5	83	84.5	85.9
top 8 features	**88.2**	**89.3**	**87**	**89.3**	**88.7**
top 9 features	88.2	89.3	87	89.3	88.7

In each training of SVM classifier, the ROCAUC algorithm achieves improved accuracy in the first 7 training runs. The accuracy remains the same after 7 features training. The highest accuracy 89%, located on top of 7 features ranked. The CQCDFR algorithm achieves improved accuracy in the first 6 training runs. The accuracy remains the same after 6 features training. The highest accuracy 92%, located on top of 6 features ranked. The ANOVA algorithm achieves improved accuracy in the first 5 training runs. The accuracy remains the same after 5 features training. The highest accuracy 93%, located on top of 5 features ranked. The RMSE algorithm achieves improved accuracy in the first 7 training runs. The accuracy remains the same after 7 features training. The highest accuracy 87%, located on top of 7 features ranked. Figure 6 presents the accuracy of support vector machine classifier for top 10 ranked based on of different selection and ranking algorithms by using signal specific feature and the signal independent feature. The blue bar presents the accuracy of SVM based on ROCAUC algorithm. The orange bar presents the accuracy of SVM based on CQCDFR algorithm. The grey bar presents the accuracy of SVM based on ANOVA algorithm. The yellow bar presents the accuracy of SVM based on RMSE algorithm.

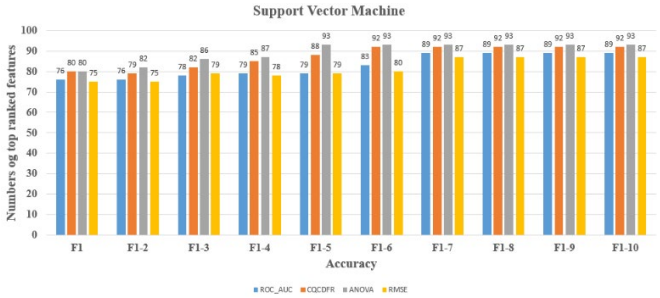


Fig. 6: Comparison the accuracy of support vector machine classifier for top 10 ranked based on of different selection and ranking algorithms by using signal specific feature and the signal independent feature.

Table VII summaries the performance of SVM for top ranked features based on different feature selection algorithms. In each training of SVM classifier, the ROCAUC algorithm achievethe highest accuracy located on top 7 features train. We select top 7 features to do the performance evalution. We achieved 93.8% sensitivity, 91.4% specificity, 89% accuracy, 91.3% precision, and 90.1% F1 score. The CQCDFR algorithm achievethe highest accuracy located on top 6 features train. We select top 6 features to do the performance evalution. We achieved 92.7% sensitivity, 94.4% specificity, 92% accuracy, 94.3% precision, and 93.4% F1 score. The ANOVA algorithm achievethe highest accuracy located on top 5 features train. We select top 5 features to do the performance evalution. We achieved 94.3% sensitivity, 93.5% specificity, 93% accuracy, 93.6% precision, and 91.5% F1 score. The RMSE algorithm achievethe highest accuracy located on top 7 features train. We select top 7 features to do the performance evalution. We achieved 88.7% sensitivity, 89.6% specificity, 87% accuracy, 89.6% precision, and 88.0% F1 score.

Table VII: Summary of SVM (support vector machine) performance for top ranked features based on different feature selection algorithms.

SVM Performance	Sensitivity (%)	Specificity (%)	Accuracy (%)	Precision (%)	F1 Score (%)
ROC_AUC					
top 6 features	88.0	84.5	83	84.6	86.2
top 7 features	**93.8**	**91.4**	**89**	**91.3**	**90.1**
top 8 features	93.8	91.4	89	91.3	90.1
CQCDFR					
top 5 features	90.1	88.9	88	89.0	90.9
top 6 features	**92.7**	**94.4**	**92**	**94.3**	**93.4**
top 7 features	92.7	94.4	92	94.3	93.4
ANOVA Test					
top 4 features	92.0	89.0	87	89.2	91.6
top 5 features	**94.3**	**93.5**	**93**	**93.6**	**91.5**
top 6 features	94.3	93.5	93	93.6	91.5
RMSE					
top 6 features	84.1	82.9	80	82.8	83.6
top 7 features	**88.7**	**89.6**	**87**	**89.6**	**88.0**
top 8 features	88.7	89.6	87	89.6	88.0

IV. Conclusions

This work aims to use machine learning algorithms to detect cardiac conditions in patients from single-lead ECG data in a database of supraventricular arrhythmias and applying wearables device. A premature ventricular contraction (PVC) is an unwanted heartbeat that begins in one of the heart's two lower pumping chambers (the ventricles). These extra heartbeats disrupt the normal heart rhythm and sometimes feel like a throbbing or throbbing in the chest. Early monitoring and treatment can be of great help to patients. We use MIT BIH supraventricular arrhythmia database (svdb) which is the public source to do the analysis. We extract combination of signal-specific features and signal-independent features from the ECG signal to form feature vectors. Signal specific features contain amplitude features, frequency features, and statistical features. Signal-independent features contain statistical features from the TSFELs. We apply four different algorithms to do the feature selection and ranking. Algorithms we use receiver operator characteristic (ROC) and the area under the ROC curve (AUC) (ROCAUC), constant, quasi constant and duplicate feature removal (Univariate) (CQCDFR), analysis of variance (ANOVA), and root mean square deviation (RMSE). We train the machine learning model of random forest (RF) and support vector machine (SVM) to find out the best feature selection method. Based on different feature selection algorithms, we analyze the ECG database and compare the accuracy of the combinations of signal-specific and signal-independent features with the top-ranked feature combinations to find the best performance. When the feature selection algorithm is different, the feature selection will also show different results. By training RF and SVM classifier, feature selection using ANOVA algorithm without ranking shows the lowest performance compare with other algorithms. After doing the feature ranking, ANOVA algorithm gets the highest accuracy compare with other algorithms. We further analyze the most accurate feature selection methods base on different algorithms to test performance such as the accuracy, sensitivity, precision, specificity and F1 score. ANOVA algorithm shoes the best performance on RF classifier and SVM classifier. Therefore, The ANOVA algorithm can help us select effective features and eliminate useless ones for testing of real-time wearable heart monitoring devices. Effective feature selection not only reduces the number of features used, but also reduces the power consumption of the monitoring device. As the amount of training data increases, the accuracy and performance of the model can be improved by relying on more features. The ability to improve the classification accuracy and features are more suitable for real-world needs.

Acknowledgement

This material is based upon work supported by the National Science Foundation under Grant No. 1932281.

References

[1] Edward A Lee,"Cyber physical systems: design challenges,"technical report no. UCB/EECS-2008- 8 January 23, 2008.

[2] M. Rahman and B. I. Morshed, "Extraction of Respiration Rate from Wrist ECG Signals," 2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), 2021, pp. 0565-0570, doi: 10.1109/UEMCON53757.2021.9666489.

[3] M. M. R. Momota and B. I. Morshed, "Inkjet Printed Flexible Electronic Dry ECG Electrodes on Polyimide Substrates Using Silver Ink," 2020 IEEE International Conference on Electro Information Technology (EIT), 2020, pp. 464-468, doi: 10.1109/EIT48999.2020.9208322.

[4] S. Khatun and B. I. Morshed, "Detection of myocardial infarction and arrhythmia from single-lead ECG data using Bagging Trees classifier," 2017 IEEE International Conference on Electro Information Technology (EIT), 2017, pp. 520-524, doi: 10.1109/EIT.2017.8053417.

[5] He J, Sun L, Rong J, Wang H, Zhang Y, "A pyramid-like model for heartbeat classification from ECG recordings, " 2018 PLoS ONE 13(11): e0206593. https://doi.org/10.1371/journal.pone.0206593.

[6] Asgari S, Mehrnia A, Moussavi M. Automatic detection of atrial fibrillation using stationary wavelet transform and support vector machine. Comput Biol Med. 2015 May;60:132-42. doi: 10.1016/j.compbiomed.2015.03.005. Epub 2015 Mar 14. PMID: 25817534.

[7] Alarsan, F.I., Younes, M. Analysis and classification of heart diseases using heartbeat features and machine learning algorithms. J Big Data 6, 81 (2019). https://doi.org/10.1186/s40537-019-0244-x.

[8] S. Liaqat, K. Dashtipour, A. Zahid, K. Assaleh, K. Arshad, and N. Ramzan, "Detection of Atrial Fibrillation Using a Machine Learning Approach," Information, vol. 11, no. 12, p. 549, Nov. 2020, doi: 10.3390/info11120549.

[9] Sadr, N.; Jayawardhana, M.; Pham, T.T.; Tang, R.; Balaei, A.T.; de Chazal, P, "A low-complexity algorithm fordetection of atrial fibrillation using an ECG, " Physiol. Meas. 2018, 39, 064003 .

[10] Lim, H.W.; Hau, Y.W.; Lim, C.W.; Othman, M.A, "Artificial intelligence classification methods of atrialfibrillation with implementation technology, " Comput. Assist. Surg. 2016, 21, 154–161.

[11] Xia, Y.; Wulan, N.; Wang, K.; Zhang, H, "Detecting atrial fibrillation by deep convolutional neural networks.Comput, " Biol. Med. 2018, 93, 84 92.

[12] Attallah O, "An Intelligent ECG-Based Tool for Diagnosing COVID-19 via Ensemble Deep Learning Techniques, " 2022 Biosensors (Basel), 12(5):299. doi: 10.3390/bios12050299. PMID: 35624600; PMCID: PMC9138764.

[13] ashif, Shadman & Raihan, Rakib & Islam, Md Rasedul & Imam, Mohammad, "Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System," 2018 World Journal of Engineering and Technology. 06. 854-873. 10.4236/wjet.2018.64057.

[14] Mondéjar-Guerra, Víctor M. & Novo, Jorge & Rouco, José & Gonzalez, Manuel & Ortega, Marcos, "Heartbeat classification fusing temporal and morphological information of ECGs via ensemble of classifiers," 2018 Biomedical Signal Processing and Control. 47. 41-48. 10.1016/j.bspc.2018.08.007.

[15] ANSI/AAMI. Testing and Reporting Performance Results of Cardiac Rhythm and ST Segment Measurement Algorithms, 1998. Association for the Advancement of Medical lnstrumentation.

[16] Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation 101(23):e215-e220 [Circulation Electronic Pages; http://circ.ahajournals.org/content/101/23/e215.full]; 2000 (June 13).

[17] de Chazal P, O'Dwyer M, Reilly RB. Automatic classification of heartbeats using ECG morphology and heartbeat interval features. IEEE Trans Biomed Eng. 2004 Jul;51(7):1196-206. doi: 10.1109/TBME.2004.827359. PMID: 15248536.

[18] J. Pan and W.J. Tompkins, "A REAL-TIME QRS DETECTION ALGORITHM," Ieee Transactions on Biomedical Engineering, vol. 32, no. 3, 1985, pp. 230-236.

[19] Makowski, D., Pham, T., Lau, Z.J. et al. NeuroKit2: A Python toolbox for neurophysiological signal processing. Behav Res 53, 1689–1696 (2021). https://doi.org/10.3758/s13428-020-01516-y.

[20] Barandas M., Folgado D., Fernandes L., Santos S., Abreu M., Bota P., Liu H., Schultz T., Gamboa H. Tsfel: Time series feature extraction library SoftwareX, 11 (2020), Article 100456, 10.1016/j.softx.2020.100456.