Beat-by-Beat Classification of ECG Signals Using Machine Learning Algorithms to Detect PVC Beats 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*—A premature ventricular contraction (PVC) disrupt the normal heart rhythm and indicate underlying cardiac disease. We aim to detect these PVC beats from electrocardiogram (ECG/EKG) data by automatically classifying these ECG beats with high accuracy in real-time. In this study, we used MIT BIH Long-Term Electrocardiogram Database (ltdb) dataset from the PhysioNet database. We extract signal-specific features and signal-independent features and combine them for feature ranking. We use principal component analysis (PCA), elastic net regularization (ENR), univariate filter of constant, quasi constant and duplicate feature removal (CQCDFR) and analysis of variance test (ANOVA) for feature selection. We take the top 10 features for four methods and classify them separately. The machine learning model explored is the random forest classifier. In our analysis, elastic net regularization performed best in terms of accuracy in cardiac patients. We further use the feature with the best accuracy in four algorithms to test sensitivity, specificity, accuracy, precision, f1-score to evaluate statistics. The overall accuracy of elastic net regularization for classifying the highest first 8 feature data is 97.8%. The sensitivity was 94.7% and the specificity was 99.6%. The accuracy rate is 99.6%, and the F1 score is 97.1%. The method can accurately detect ECG beats and analyze categories for real-time cardiac monitoring for feedback to the use patient. Efficient feature selection minimizes the number of features used and reduces the power consumption of the monitoring device.**

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

I. INTRODUCTION

PVCs are one of the most common clinical arrhythmias. Due to its variability and susceptibility, patients may be at risk at any time. When concomitant with heart disease, frequent premature contractions can cause a chaotic and dangerous heart rhythm that can lead to sudden cardiac death. So, it is important for heart patients to detect PVC problems in real time. PVC beats can be detected from electrocardiogram (ECG or EKG) signals [2-7]. Therefore, an intelligent method for automatic heartbeat classification based on ECG recordings is needed, which will be of great help to clinicians in diagnosing heart disease [2]. Automated analysis techniques to identify PVCs can be performed using a trained machine learning (ML) model.

PVC discrimination method based on multi-feature combination and random forest algorithm is proposed to improve the discrimination rate of PVC [3]. Normal, PVC, and other types of ECG heartbeats are classified using different features and evaluate logistic regression (LG), neural network (NN), support vector machine (SVM) with different parameters [4]. Developed sparse Bayesian methods, such as correlation vector machines (RVMs), offer a parsimonious solution compared to support vector machines (SVMs) but show competitive accuracy [5]. A k-nearest neighbor (KNN) classifier computes the distance between samples based on these features to detect PVCs [6]. Random forest is adopted as a supervised algorithm for training the features generated by the autoencoder. Multiple Active Learning Options Uncertainty-based and diversity-based strategies are studied on top of random forests [7]. The PVC recognition features based on deep learning methods are fed into a convolutional neural network (CNN) to find unique patterns and classify them more efficiently [8].

The focus of this work is to analyze each heartbeat through a unique approach and apply in the wearable device. We form the slimier training and test data for normal heartbeats and abnormal heartbeats. Detect PVC beats from single-lead ECG data and test the accuracy of the top-ranked features and compare against all features to find the best performance. Increase the number of features and add feature selection and ranking. We use the random forest classifier and measure statistical performance metrics for top performer. In single-lead ECG signals, we use signal specific features and signal independent features to form feature vectors for classification. Our special beat-by-beat classification method classifies ECG data in real-time for wearable heart monitors. Efficient feature selection minimizes the number of features used and reduces the power consumption of the monitoring device.

II. METHODOLOGY

This study used the AAMI criteria to classify MIT BIH heartbeat types [9]. MIT-BIH has five heartbeat categories for subsequent processing. Each category includes one or more types of heartbeats, as shown in Table I. Class N has normal and bundle branch block heartbeat types. Class S has supraventricular ectopic heartbeat (SVEB). Class V has

Table I: The standard of AAMI classes and labeling of MIT-BIH long-term ECG Database with the full database, and extract training and test dataset.

AAMI	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	600,232	1,500	64,095	2,908	1,117	669,852
TRAINING	266,949	285	37,522	169	218	305,143
TESTING	333,283	1,215	26,573	2,739	899	364,709

The training dataset contain records 14046, 14157, 14184. The testing dataset contain records 14134, 14149, 14172, 15814.

ventricular ectopic beats (VEB). Class F has beats that fusion normal and VEB generated. Class Q has an unknown beat. To achieve our goal, we accessed the PhysioNet database exposed by the public data resource and selected the MIT BIH long-term ECG database (ltdb) [10]. We use PVC beats to classify with normal beats. The process flow diagram of PVC beat detection is shown in Figure 1.

A. ECG data

In our study we used data from the MIT-BIH Long-Term ECG Database [11], which includes many recordings of PVC beats. The database contains 7 records of duration between 14 to 23 hours and containing two ECG lead signals. Data was sampled at 128 Hz and bandpass filtered. The largest category is "normal beat" (N) with over 600,000 samples, and the smallest category is "unknown beat" (Q) with only around 1000 samples. Only normal beat (N) and ventricular ectopic beat (V) were used for the analysis [11]. The training dataset contains ECG data recorded from samples 14046, 14157, 14184, and the test dataset contains ECG data recorded from samples 14134, 14149, 14172, 15814. Both datasets contain approximately 300,000 heartbeats and mix arrhythmia recordings of normal (N) heartbeats and ventricular ectopic (V) beats. Table I also shows a breakdown of each dataset and by heartbeat type.

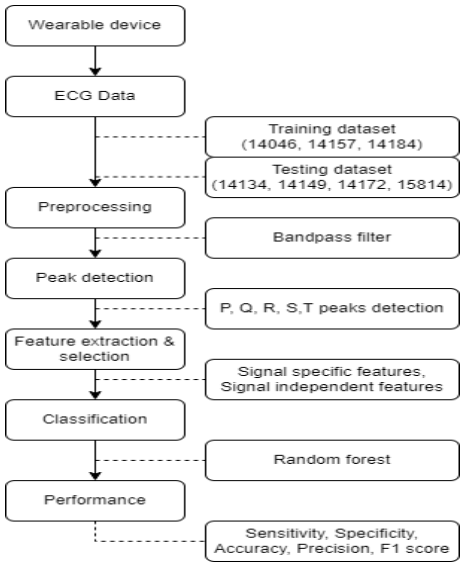

Fig. 1: The process flow diagram for processing, analyzing, and classifying beat-to-beat ECG signals using machine learning algorithms.

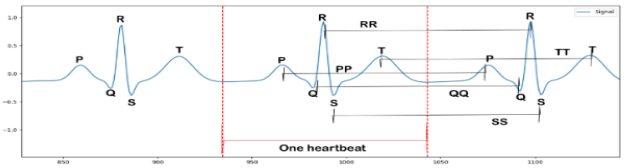

Fig. 2: A representative ECG signal waveform showing 3 beats, where P, Q, R, S, T peaks and inter-beat parameters are shown from one heartbeat.

B. ECG signal preprocessing

In recorded electrocardiogram (ECG) signals, clinical information is masked by multiple noises and distortions. These noises and distortions result in a low signal-to-noise ratio (SNR). The main noise sources are power frequency interference, baseline drift, EMG interference and random noise. After analyzing these interferences, select an appropriate method to remove these noises to obtain a relatively real ECG signal. The frequency of ECG signal is mainly concentrated in 5~20HZ, so choose a low-pass filter to filter out EMG interference.

C. Peak detection

In ECG heartbeat recognition, the detection of R peak is very important. Figure 2 shows P, Q, R, S, T peaks and one heartbeat. The detection of the R peak affects the correct position of the remaining P, Q, S, T, T′ peaks. We use the pan_tompkin algorithm in this work to detect the exact location of the R peak [12]. The open-source code, "The Python Toolbox for Neurophysiological Signal Processing" uses the Pan Tompkin algorithm to detect R peaks. This file is provided by neurokit 2 [13]. After we find the location of the R peak, we can start to calculate the RR interval and the average RR interval. P, Q, S, T and T' peaks can be found using the RR interval and the average RR interval. The middle value of the first R peak and the second R peak as the starting point of a beat. The middle value between the second R peak and the third R peak is regarded as the end point. We use middle of two R-peaks to form a heartbeat

D. Feature extraction and selection

The feature extraction process performs extraction every two ECG beats and slides one beat each time (i.e., the next beat) when processing is complete. We extract the signal specific feature and the signal independent feature and combined them together to do the feature ranking. For the signal specific feature, we extract 7 of amplitude features, 6 of frequency features, and 15 of statistical features. Time series feature extraction library (TSFEL) which is a python package is used to extract 175 signal independent features [17]. We used principal component analysis (PCA), Elastic net regularization, linear and logistic regression coefficients with Lasso (L1) and Ridge (L2) regularization (Elastic net regularization), univariate filter of constant, quasi constant and duplicate feature removal (CQCDFR) and analysis of variance test (ANOVA) to do the feature selection. In PCA, the total of 34 signal independent features be selected. In ENR, the total of 46 signal independent features be selected. In CQCDFR, the total of 34 signal independent features be selected. In ANOVA, the total of 32 signal independent features be selected. We combined signal independent features with the signal specific feature, separately and take top 10 ranking of features to do the classification. Table II and Table III presents the detail of signal independent features and the top 10 ranked features using PCA, ENR, CQCDFR and ANOVA selection.

E. ML classification

We train the model with an ML classification method and use a 10-fold cross-validation technique. We explore random forest classifier to find the best accuracy. Ventricular ectopic beats are marked with 1, and normal heartbeats are marked with 0. Training datasets contain 304,471 heartbeats, and testing datasets contain 359,856 heartbeats. We first classify all features to check the accuracy. In PCA, ENR, CQCDFR, and ANOVA algorithms, we tested the top 10 ranked features for random forest training and testing, separately. The training method is that we select the top-ranked feature for training. Then we train by adding the second-ranked feature to the first-ranked feature until we add 10 ranked features for training.

Table II: Summary of signal independent features and the top 10 ranked features using PCA and ENR selection.

Type of feature	Features
Signal Independent Feature (Time series feature extraction library (Tsfel))	Total 175 features be extracted
Principal Component Analysis (PCA) (34)	fast Fourier transform mean coefficient, power bandwidth, spectral distance, median absolute diff, median diff, spectral entropy, …etc.
Elastic net regularization (ENR; Embedded) (46)	histogram, signal distance, ECDF percentile count, empirical cumulative distribution function (ECDF), root mean square, slope, …etc.
PCA feature ranking of signal specific feature and signal independent feature (Top 10 ranked in order)	fast Fourier transform mean coefficient 22, power bandwidth, histogram 3, signal distance, ECDF Percentile Count, empirical cumulative distribution function (ECDF), histogram 5, root mean square, fast Fourier transform mean coefficient 23, autocorrelation
ENR feature ranking of signal specific feature and signal independent feature (Top 10 ranked in order)	empirical cumulative distribution function (ECDF 9), ECDF 4, fast Fourier transform mean coefficient 23, fast Fourier transform mean coefficient 25, fast Fourier transform mean coefficient 1, wavelet absolute mean, neighborhood peaks, Fourier transform mean coefficient 22, Fourier transform mean coefficient 14, frequency of T

Table III: Summary of signal independent features and the top 10 ranked features using CQCDFR and ANOVA selection.

Type of feature	Features
Signal Independent Feature (Time series feature extraction library (Tsfel))	Total 175 features be extracted
Univariate filter of constant, quasi constant and duplicate feature removal (CQCDFR)(34)	empirical cumulative distribution function (ECDF), root mean square, slope, wavelet energy, wavelet entropy, …etc.
Analysis of variance test (ANOVA) (32)	signal distance, slope, wavelet energy, wavelet entropy, spectral centroid, …etc.
CQCDFR feature ranking of signal specific feature and signal independent feature (Top 10 ranked in order)	root mean square, slope, wavelet energy, wavelet entropy, spectral centroid, autocorrelation, mean absolute deviation, spectral skewness, wavelet standard deviation, wavelet variance
ANOVA feature ranking of signal specific feature and signal independent feature (Top 10 ranked in order)	median frequency, negative turning points, neighborhood peaks, peak to peak distance, wavelet standard deviation, fast Fourier transform mean coefficient, spectral distance, median absolute diff, median diff, spectral entropy

F. Performance

Regarding performance evaluation, we use various statistical metrics to demonstrate model performance. We evaluated the accuracy, sensitivity, specificity, precision, and F1 score for all feature accuracy and best accuracy on the feature selection method. A measure of correctly identifying actual positives can be expressed in terms of sensitivity. A measure of the actual proportion of negatives correctly identified can be expressed in terms of specificity. A measure of recognition accuracy can be expressed in terms of precision. Identifying how close a measurement is to the true value can be expressed in terms of accuracy. A measure of test accuracy can be expressed as an F1 score.

III. Results

We use the Pan-Tompkin algorithm to detect the R point, and find the P, Q, S, T points. We extract R peaks using a sliding window technique. In Figure 3, we represent the top of the R peak with a dashed purple line. Three R peaks are located at 875, 985, and 1095 sample intervals, respectively. When we find the R peak we can detect the P, Q, S, T, T' points. Table II shows the standard of P, Q, S, T, T' peak detection. Figure 4 shows the detection of P, Q, S, T points every 3 runs. The purple dashed line numbered 0 represents the position of the T peak, the green dashed line numbered 1 represents the position of the P peak, the yellow dashed line numbered 2 represents the position of the Q peak, and the red dashed line numbered 3 represents the position of the S peak. We set the sampling interval from 828 sampling interval to 1160 sampling interval.

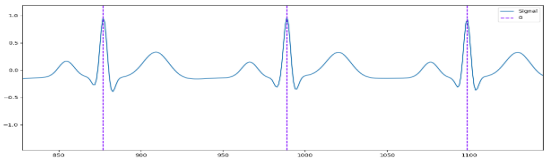

Fig. 3: Three beats for R point detection with Pan-Tompkins's algorithm in an ECG signal

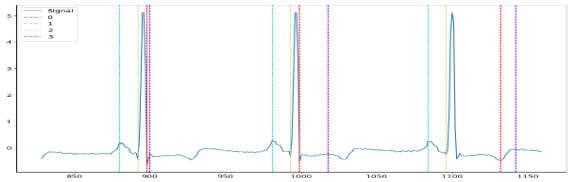

Fig. 4: P, Q, S, T peaks detected for 3 ECG beats after R peak detection

Table IV: Summary of random forest performance for combination of the signal specific feature and the signal independent feature.

Perform(%)	*Sensitivity*	*Specificity*	*Accurace*	*Precision*	*F1*
PCA (62)	70.3	71.1	75.42	71.0	72.6
ENR (74)	81.6	80.3	82.64	80.2	82.3
CQCDFR (62)	78.1	76.8	74	76.7	79.2
ANOVA(60)	78.9	77.2	73	77.3	78.1

We use random forest classifier to test combination of the signal specific feature and the signal independent feature. Table IV shows the performance for combination of the signal specific feature and the signal independent feature. Elastic net regularization (ENR) algorithm got the best performance than the other algorithms. Total of 74 features comes from the signal specific feature and ENR selected signal independent feature. We get the 82.64% accuracy, 81.6% sensitivity, 80.3% specificity, 80.2% precision, and 82.3% F1 score.

We also test accuracy of top ranked features and compare with all features to find the best performance. Figure 5 shows the accuracy of ranking the top 10 ranked features from the signal specific feature and the signal independent feature using the random forest classifier. The blue, orange, gray and yellow bars represent the PCA, ENR, CQCDFR and ANOVA algorithms, respectively. In PCA algorithm, the highest accuracy rate in first 7 features ranked is 90%. In ENR, the highest accuracy in first 8 features ranked is 98%. In CQCDFR, the highest accuracy in first 6 features ranked is 90%. In ANOVA, the highest accuracy in first 7 features ranked is 94%.

We test the performance evaluation of sensitivity, specificity, accuracy, precision, and F1 score on the best accuracy of random forest. As shown in Table V. In PCA, we achieved 91.3% sensitivity, 92.4% specificity, 90.32% accuracy, 92.3% precision, and 91.8% F1 score in top 7 features. In ENR, we achieved 94.7% sensitivity, 99.6% specificity, 97.83% accuracy, 99.6% precision, and 97.1% F1

score in top 8 features. In CQCDFR, we achieved 93.3% sensitivity, 91.4% specificity, 90% accuracy, 91.3% precision, and 91.6% F1 score in top 6 features. In ANOVA test, we achieved 94.9% sensitivity, 99.7% specificity, 97% accuracy, 99.6% precision, and 97% F1 score in top 7 features.

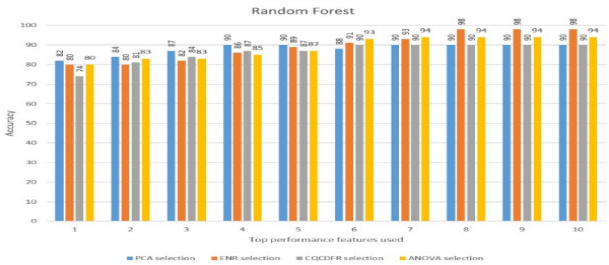

Fig. 5: Comparison of the accuracy for top 10 ranked of the signal specific feature and the signal independent feature.

Table V: Summary of random forest performance for top ranked features.

Perform (%)	*Sensitivity*	*Specificity*	*Accurace*	*Precision*	*F1*
PCA					
top 6 features	90.4	89.3	88	89.4	87.2
top 7 features	**91.3**	**92.4**	**90**	**92.3**	**91.8**
top 8 features	91.3	92.4	90	92.3	91.8
ENR					
top 7 features	90.5	96.2	93	96.3	92.3
top 8 features	**94.7**	**99.6**	**98**	**99.6**	**97.1**
top 9 features	94.7	99.6	98	99.6	97.1
CQCDFR					
top 5 features	89.7	88.9	86	90.2	88.9
top 6 features	**93.3**	**91.4**	**90**	**91.3**	**91.6**
top 7 features	93.3	91.4	90	91.3	91.6
ANOVA Test					
top 6 features	92.7	90.4	91	90.5	93.1
top 7 features	**94.9**	**99.7**	**94**	**99.6**	**97**
top 8 features	94.9	99.7	94	99.6	97

IV. Conclusions

This work aims to detect cardiac conditions in patients from a long-term ECG database using machine learning algorithms from single-lead ECG data and apply in the wearable device. A PVC beat is a premature heartbeat that originates in the ventricles and disrupts the heart's normal rhythm. Early detection and monitoring are important, which can provide early predictor for worsening cardiac health condition. We performed this analysis using the publicly available MIT BIH long-term ECG dataset. From the ECG signal, we use signal-specific features and signal-independent features to form feature vectors. We use PCA, ENR, CQCDFR, and ANOVA algorithms for feature extraction and selection. We test random forest classifier to obtain the best feature selection method by analyzing ECG data based on different feature extraction and selection and compare the performance of different features. We test accuracy of top ranked features and compare with all features to find the best performance. Based on the different algorithms used for doing feature selection, the result of selected features is different. In my work, Elastic net regularization algorithms are always higher than the other algorithms. We further analyzed the most accurate feature selection methods to test performance. ENR algorithm showed the best accuracy in this work as it is suitable for multivariate data structure. Since both PCA and ANOVA are univariate methods based on SVD, they do not take into account the potential multivariate nature of the data structure, whereas ENR has the ability to tackle that issue, hence leading to better performance as demonstrated in this paper. Therefore, ENR selection of embedded method can help us select effective features for training and testing of real-time wearable heart monitoring devices. Effective feature selection can minima the numbers of feature using and reduce the power consumption of monitoring devices. More features help improve model accuracy and performance. With the increase in the amount of training data, the classification accuracy and features are more suitable for the needs of early predictive monitoring of cardiac health.

Acknowledgement

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

References

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

[2] Tiantian Xie, Runchuan Li, Shengya Shen, Xingjin Zhang, Bing Zhou, Zongmin Wang, "Intelligent Analysis of Premature Ventricular Contraction Based on Features and Random Forest", Journal of Healthcare Engineering, vol. 2019, Article ID 5787582, 10 pages, 2019. https://doi.org/10.1155/2019/5787582.

[3] M. M. Casas, R. L. Avitia, M. A. Reyna and A. Cárdenas, "Evaluation of three machine learning algorithms as classifiers of premature ventricular contractions on ECG beats," 2016 Global Medical Engineering Physics Exchanges/Pan American Health Care Exchanges (GMEPE/PAHCE), 2016, pp. 1-6, doi: 10.1109/GMEPE-PAHCE.2016.7504615.

[4] Ribeiro, Bernardete & Marques, Amândio & Henriques, Jorge & Antunes, Manuel. (2007). Choosing Real-Time Predictors for Ventricular Arrhythmia Detection.. IJPRAI. 21. 1249-1263. 10.1142/S0218001407005934.

[5] Yu J, Wang X, Chen X, Guo J. Automatic Premature Ventricular Contraction Detection Using Deep Metric Learning and KNN. Biosensors (Basel). 2021 Mar 4;11(3):69. doi: 10.3390/bios11030069. PMID: 33806367; PMCID: PMC8000997.

[6] Xianrong Zhang, Muhammad Shafiq, Guijun Zheng, Junping Wan, Zhe Sun, "Premature Ventricular Contractions' Detection Based on Active Learning", Scientific Programming, vol. 2021, Article ID 5556011, 14 pages, 2021. https://doi.org/10.1155/2021/5556011.

[7] Sarshar NT, Mirzaei M. Premature Ventricular Contraction Recognition Based on a Deep Learning Approach. J Healthc Eng. 2022 Mar 26;2022:1450723. doi: 10.1155/2022/1450723. PMID: 35378947; PMCID: PMC8976634.

[8] Sarshar NT, Mirzaei M. Premature Ventricular Contraction Recognition Based on a Deep Learning Approach. J Healthc Eng. 2022 Mar 26;2022:1450723. doi: 10.1155/2022/1450723. PMID: 35378947; PMCID: PMC8976634.

[9] 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.

[10] 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).

[11] 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.

[12] 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.

[13] 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.