Beat-by-beat Classification of ECG Signals with Machine Learning Algorithm for Cardiac Episodes

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*—Heart failure (HF) is a common clinical syndrome of cardiac episode leading to a variety of cardiac diseases. Detecting these cardiac episodes from electrocardiogram (ECG or EKG) data and classifying these large data automatically with high accuracy in real-time is critical for useful application of wearables targeting cardiac disease monitoring. With this motivation, in this study, we used the BIDMC Congestive Heart Failure (CHF) datasets (from PhysioNet database). A total of 15 patient records was analyzed, which have NYHA Class level III and IV patients from the database. Simultaneous measurements of the 2 leads of ECG were stored in the record. The captured data was sampled at 250 Hz. The extracted features were for three categories: temporal, spectral, and statistical. In total, we extracted 28 features out of which 7 were of amplitude types, 6 were based on frequency, and the remaining 15 were statistical features. Machine learning models explored include SVM, KNN, ensemble tree, neural network, decision tree, naive bayes, and logistic regression. We evaluated different model performance in each patient data and combined patient data. In our analysis, neural network was the best performer in terms of accuracy for cardiac patients. We further studied neural network to test sensitivity, specificity, accuracy, precision, f1-score to evaluate the best performer statistics. Neural network has 99.5% overall accuracy for interpatient data classification, and was also among the best performers. In interpatient classification, the performance was: sensitivity 99.80%, specificity 99.0%. accuracy 99.42%, precision 99.80%, and F1 score 99.64%. Accurate detection of ECG beat classes using this approach can allow real-time cardiac disease monitoring.

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

I. INTRODUCTION

Heart failure (HF) is a common clinical syndrome that can be caused by a variety of cardiac diseases [1]. The current commonly used assessment method to classification severity HF is the New York Heart Association (NYHA) cardiac function classification. Cardiac function classification is a clinical method to assess the degree of cardiac impairment. The grading of cardiac function in patients with heart disease can generally reflect the severity of the disease and has practical value in the selection of treatment measures, the evaluation of labor ability, and the judgment of prognosis.

Heart failure and other cardiac diseases can be detected using electrocardiogram (ECG or EKG) signals [2-11]. Analyzing these bio-electrical signals of each heartbeat, the movements produced by different special heart tissues, cardiologists can detect some of these abnormalities [2]. However, manual scrutiny of continuous ECG signals for long duration for each patient is not practical and expensive. Automated detection can be performed with trained machine learning (ML) models.

Stationary Wavelet Transform (SWT) and Support Vector Machine (SVM) are some of the common methods used for such detection. Some of these ML methods eliminate detection of P-peak or R-peak, and therefore do not depend on ECG beat detection for performance [3]. Other methods include extraction of labeled features from ECG and Heart Rate Variability (HRV) signals. The key features useful for such ML classifications can be spectral features, bispectral features, and nonlinear features including sample entropy and Poincaré graph extraction features [4]. The ability to classify according to Heart Failure with Preserved Ejection Fraction (HFPEF) and Heart Failure with Reduced Ejection Fraction (HFREF) can be useful for cardiac patient progression monitoring. Furthermore, early detection of cardiac patient can be possible by comparing the probability of the presence of HFPEF with that of traditional logistic regression [5]. Signature serum creatinine and ejection fraction can be performed by employing traditional biostatistical tests, and these results can be compared with those provided by machine learning algorithms for performance evaluation [6].

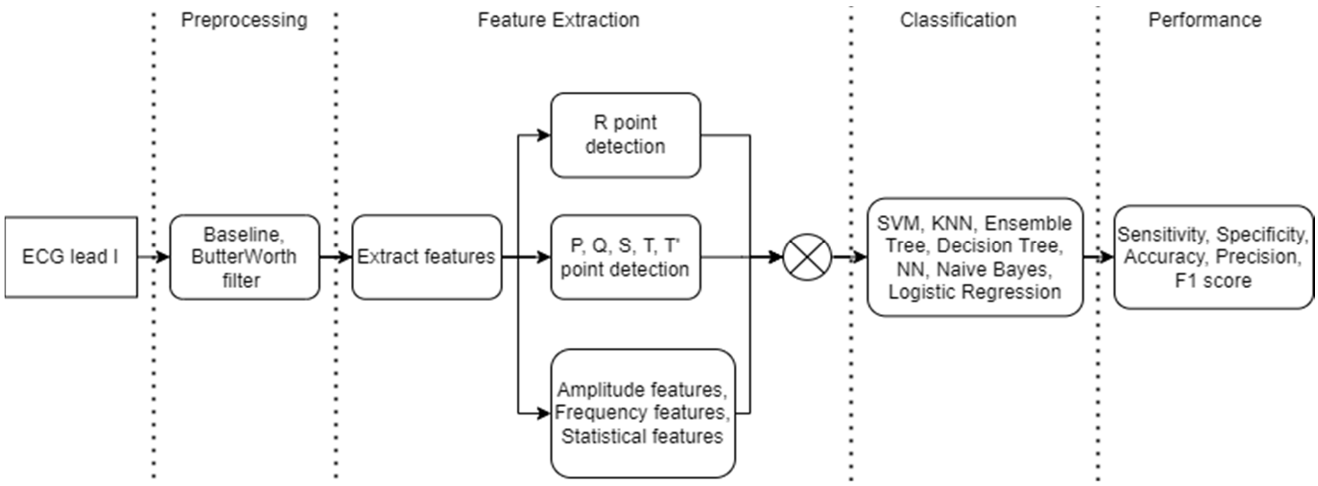

Fig. 1: Process flowchart for processing, analyzing, and classifying beat-by-beat ECG signals using machine learning algorithms.

Various machine learning models, both unsupervised and supervised, have been used to classify HF subtypes within the ventricular volume domain [7]. ECG classification method using machine learning were demonstrated with ML-libs and Scala language on Apache Spark framework based on multiple ECG features [8]. Deep learning (DL) techniques were also demonstrated by analyzing using five classes of ECG datasets and examining the constructed models [9]. A framework for processing of ECG signals to determine HF onset was developed and good performance was achieved using Long Short-Term Memory (LSTM) [10]. In another work, the interval and magnitude features were extracted using single-channel ECG data and classified with Bagging Tree [11].

The focus of this work is to detect cardiac episodes from single-lead ECG data by analyzing each beat at a time, a unique approach, and test various types of classifiers to find the best performer ML algorithm, and also to measure statistical performance metrics for optimal performer. From the ECG signal of each lead, we used features in amplitude, amplitude difference, time interval, and statistical domain to form a feature vector for classification. Our unique approach of beat-by-beat classification can lead to real-time classification of streaming ECG data suitable for implementation on wearable cardiac monitors.

II. METHODOLOGY

In this work, we used the BIDMC Congestive Heart Failure (CHF) datasets from publicly available PhysioNet database [12]. A total of 15 patient records, which have NYHA Class level III and IV patients, were used for this analysis. The simultaneously measured signals of 2 ECG leads were stored in the record. The captured data sampling rate was 250 Hz. All the patients are in the condition of NYHA class level III and IV. The process flowchart for HF is given in Fig. 1.

A. ECG signal preprocessing

The ECG signal is easily contaminated by high-frequency and low-frequency noises. Common interferences are radio frequency interference, Electromyography (EMG), power line interference, mechanical force acting on electrodes, breathing, patient movement, and equipment movement interference. In order to filter the noise, a band-pass filter is used. The characteristic of using this Butterworth filter is that the frequency response curve in the transparency is integrated, there is no characteristic, and the resistance range gradually decreases and appears.

B. R peak detection

The detection of the R peaks is very important for ECG beat identification. The detection of point R affects the correctness of prominent points such as P, Q, R, S, T, and T′. We used the *pan_tompkin* algorithms to detect R peaks [13]. The code is from the MATLAB (MathWorks Inc.) file exchange, which uses the Pan Tompkin algorithm to realize the detection of the R peaks. The document in MATLAB exchange was provided by Hooman Sedghamiz.

C. P, Q, S, T, T′ detection

After detecting the R peak point, we calculate the RR interval and the average RR interval of each data set, and use this information to calculate Q, P, S, T, and T'. The specification to determine these are provided in Table I.

Table I: Formula for detect P, Q, S, T, T' peak.

Peak	Formula
P	Before every R peak to Q peak, from 3/8 of RR maximum
Q	Before every RR interval, minimum between ⅛ of each R peak to R peak
S	Before every R peak, minimum between every R peak to 1/4 of RR interval
T	After 1/4 of R peak to R peak to the 3/8 of R peak to R peak in maximum
T'	After 1/4 of R peak to R peak to the 3/8 of R peak to R peak in minimum

D. Feature extraction

The feature extraction process is executed for every two ECG beats, and then slides one beat (i.e., next beat) every time the processing is complete. The extracted features were for three categories: temporal, spectral, and statistical. Fig. 2 shows some of the key features of ECG signals. The features with three properties of amplitude, frequency, and statistics. Amplitude features include QRS position, peak of Q, S, T, R, and length of QT, RR. Frequency features include sampling frequency of Q, R, S, T, heart rate, and instance heart rate. Statistical features have maximum frequency, mean peak of R, Q, S, T and QT, mean of LF and HF, the ratio of LF/HF, very low frequency, an index of sympathetic and parasympathetic modulation of the autonomic nervous system, the proportion of NN50 divided by the total number of NN (R-R) intervals, root mean square of successive differences, and the stationary wavelet transform. Thus, the total number of features were 28.

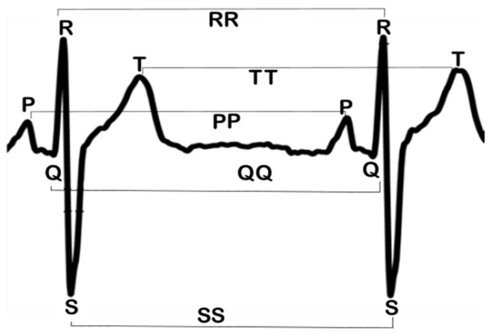

Fig. 2: Key features of temporal representation of ECG signals (x-axis represents time and y-axis represents voltage, arbitrary scale).

E. ML classification

The ML classification method used were from MATLAB Classification Learner Toolbox. In order to train the model, a 10-fold cross-validation technique is used. Explored ML models include Support Vector Machine (SVM), k-nearest neighbors (KNN), ensemble tree, neural network, decision tree, naive bayes, and logistic regression. For each patient's classification, we pick up 80% of ECG beat data for training, and 20% of ECG beat data for testing. For interpatient data, we divided 15 patients' data to training and testing. Here, we used the first 12 patients' data is for training, and the rest of the 3 patients' data is for testing. This approach is called "interpatient classification" which is more representative for clinical application of ML algorithms (developed with patient datasets and applied to real patients). The feature vector of interpatient data is 64,026 x 29. We mark the abnormal beat as 1 and the normal beat as 0. We tested how different ML models perform in each patient data and interpatient data.

F. Performance Evaluation

For performance evaluation, we used multiple statistical metrics to reflect model effects. In addition to accuracy, we also determined sensitivity, specificity, precision, and F1-score to evaluate the best performing ML method. A confusion matrix is a specific matrix used to represent the performance of a ML algorithm, where each column represents the predicted value, and each row represents the actual category, as represented in Fig. 3.

		Actual Value	
		Good = 1	Bad = 0
Predicted Value	Good = 1	TP	FP
	Bad = 0	FN	TN

TP: True Positive *FP: False Postive*
FN: False Negative *TN: True Negative*

Fig. 3: Confusion matrix for performance evaluation of ML models.

Sensitivity is a measure of which actual positives are correctly identified. Specificity measures the proportion of actual negatives correctly identified. Precision is a measure of what proportion of positive identifications are correct. Accuracy means that the measurement results are close to the true value. Accuracy calculates the number of correct classifications for each class and then adds up to calculate the percentage. F1-score is a measure of test accuracy. It considers both the precision and recall of the test to calculate the score.

III. Results

For the peak detection, we detected R peak point first and then P, Q, S, T points. We used the Pan-Tompkin algorithm to detect R-points. The extraction of R peak process runs every two beats, moving one beat each time processing is complete (i.e., sliding window technique). As an illustrative instance, we pick up sample interval from 1200 to 1400 to detect R peaks as depicted in Fig. 4 that shows one beat of R peak from 1200 sample interval to 1400 sample interval. There is a pink circle mark at the top of R peak. The R peak is located at 1338 sample interval in this example. After we found out R peak successfully, the next step is to detect P, Q, S, T points. Using Table 1 to calculate Q, P, S, T, and T' location range. Fig. 5 depicts the detection of P, Q, R, S, T points for running every 5 beats. There are 5 different marks showing where the different peaks are. The red circle represents R peak location, the green triangle represents Q peak location, the yellow square means S peak, the pink square represents P peak location, and the blue triangle means T peak location. The sample interval is from 1800 sample interval to 3400 sample interval. The algorithm detects P, Q, R, S, T points.

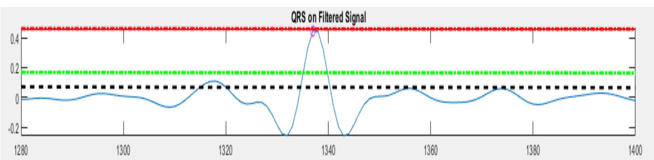

Fig. 4: One beat for R point in an ECG signal

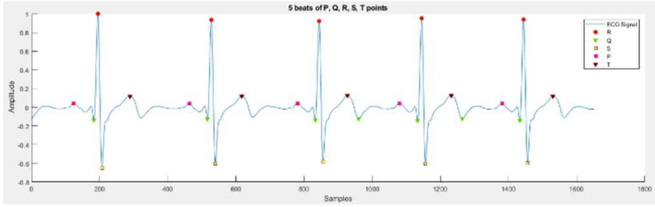

Fig. 5: P, Q, R, S, T peaks detected for 5 ECG beats

In order to know the accuracy of different ML models, we test the performance of each model with 15 patients. Table II shows the summary of different ML model performance for each patient data. We note the highest accuracy in each model. For the SVM, Patient 8 and 11 have accuracy of 99.7% each. For KNN, patient 14 has 99.8% accuracy. In ensemble tree, patient 10 has 99.8% accuracy. For neural network, Patient 1 and 15 have accuracy of 99.9% and 99.8%, respectfully. In decision tree model, Patient 4, 5, 6, and 13 have the best accuracy. Patient 4, 5, and 13 have accuracy of 99.8%, while patient 6 has accuracy of 99.9%. Using naïve bayes model, Patient 3, 12, and 14 have accuracy of 99.8%, 99.7%, and 99.9%, respectfully. For logistic regression model, Patient 7 and 10 have 99.6% and 99.8% accuracy, respectfully. We also assemble all patient's data to test the accuracy of different models for interpatient classification. Table III shows that neural network has 99.5% accuracy, and it is one of the best performances than other models.

Table. II: Summary of different machine learning model accuracy for each patient data.

Patient	SVM (%)	KNN (%)	Ensemble Tree (%)	Neural Network (%)	Decision Tree (%)	Naïve Bayes (%)	Logistic Regression (%)
1	99.1	93.6	99.1	99.9	99.5	98.6	99.6
2	98.1	99.0	94.5	96.3	93.7	93.7	98.7
3	95.0	95.7	94.7	96.6	94.6	96.8	97.8
4	94.7	97.8	95.8	94.1	99.8	98.4	94.2
5	99.0	99.4	94.2	96.8	95.8	98.7	95.2
6	94.3	99.0	95.7	95.7	99.9	96.2	95.0
7	97.3	97.5	95.7	99.0	94.5	93.7	99.6
8	95.7	94.5	99.5	94.7	98.7	95.7	97.3
9	99.2	98.7	96.9	95.7	97.7	93.6	95.7
10	96.9	97.3	99.8	94.5	99.1	96.9	99.8
11	99.7	99.1	93.6	97.7	97.3	94.6	97.9
12	95.0	97.2	97.7	95.7	94.7	94.7	99.0
13	94.5	96.5	94.3	97.2	95.8	99.5	97.7
14	96.6	99.8	94.5	99.6	97.7	99.9	94.7
15	95.7	98.7	95.6	99.8	98.7	94.3	95.7

Table. III: Summary of different machine learning model accuracy for interpatient data.

Patient	SVM (%)	KNN (%)	Ensemble Tree (%)	Neural Network (%)	Decision Tree (%)	Naïve Bayes (%)	Logistic Regression (%)
All	99.4	99.7	98.2	99.5	99.4	98.7	99.2

The neural network classifier showed the best performance than other models within each patient dataset. So we further analyze this ML model to test each patients, and measure performance of sensitivity, specificity, accuracy, precision, and F1-score. Table IV presents the summary of neural network performance in each patient. Here, patient 12 has the highest sensitivity of 97.8%, patient 1 has the highest specificity of 99.8%, patient 10 has the best accuracy of 98.57%, Patient 1 and 6 have the best precision of 99.8%, and patient 7 has the best F1-score of 99.64%. We also combined all patient data for interpatient classification, and used neural network classifier to classify the combined data. As Table V presents, we reached the sensitivity of 98.2%, specificity of 99.1%, accuracy of 98.64%, precision of 99.3%, and F1-score of 99.42%.

Table. IV: Summary of neural network performance for each patient.

Patient	Sensitivity (%)	Specificity (%)	Accuracy (%)	Precision (%)	F1 Score (%)
1	94.3	99.8	97.05	99.8	96.97
2	95.2	97.5	96.23	98.7	97.26
3	93.7	98.7	95.78	99.5	98.37
4	94.5	99.2	98.42	96.9	94.65
5	96.2	96.8	97.37	97.3	94.91
6	94.6	99.7	95.07	99.8	98.57
7	94.9	97.9	97.12	97.7	99.64
8	95.3	99.0	96.25	96.2	97.93
9	94.1	96.6	97.98	94.5	95.75
10	93.0	95.7	98.57	95.6	93.70
11	96.7	97.8	96.50	96.4	97.67
12	97.8	94.5	96.34	97.3	96.24
13	95.0	93.9	98.21	94.7	96.64
14	94.2	97.0	97.69	95.8	94.71
15	94.6	99.4	96.98	99.0	96.88

Table. V: Summary of neural network performance for interpatient data.

Patient	Sensitivity (%)	Specificity (%)	Accuracy (%)	Precision (%)	F1 Score (%)
All	98.2	99.1	98.64	99.3	99.42

IV. Conclusions

This work is aimed to detect cardiac condition of HF patients using machine learning algorithms from single-lead ECG data. Heart failure is a serious global public health problem and is the final stage in the development of most cardiovascular diseases. It is important to be detected in early stage. We used publicly available BIDMC CHF datasets for this analysis. From the ECG signals, we used amplitude, amplitude difference, frequency interval, and statistical features to form a feature vector. We test various types of classifiers to obtain the highest accuracy ML model by analyzing ECG data based on each beat, and compare the performance between each patient's data as well as interpatient data. We further analyzed the model which had highest accuracy to test the performance such as sensitivity, specificity, accuracy, precision, and F1-score. The neural network classifier showed the best accuracy in this work. We choose neural network classifier to test the performance between each patient's data and interpatient data. Thus, the neural network classifier can help us develop efficient algorithm for training and testing in real-time wearable cardiac monitoring device as this algorithm can deal with multiple calculation quickly. To improve this further, we can increase dataset for training and testing models for more accurate results. The classification accuracy is more appropriate to real world performance as the size of the training data increases. More features can also help to improve the accuracy and performance of the models. We can also employ feature selection to find the optimal subset of features. Feature ranking and top feature selection can eliminate irrelevant or redundant features, thereby reducing the number of features, improving model accuracy, and reducing running time. On the other hand, the selection of truly relevant features simplifies the model and assists in understanding the process of data generation. One limitation of this study is that the data for training and testing for the normal beat and abnormal beat are not balanced. Because we have unbalanced training dataset and testing dataset, the result of machine learning classifier might be skewed. We will further analyze this in future with balanced dataset.

Acknowledgement

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

References

[1] Ho KK, Pinsky JL, Kannel WB, Levy D. The epidemiology of heart failure: the Framingham Study. J Am Coll Cardiol 1993; 22:6A.

[2] E.J.d.S. Luz, et al., “ECG-based heartbeat classification for arrhythmia detection: A survey,” Computer methods and programs in biomedicine, vol. 127, 2016, pp. 144-164.

[3] 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.

[4] Mohebbi M, Ghassemian H. Prediction of paroxysmal atrial fibrillation based on non-linear analysis and spectrum and bispectrum features of the heart rate variability signal. Comput Methods Programs Biomed. 2012 Jan;105(1):40-9. doi: 10.1016/j.cmpb.2010.07.011. Epub 2010 Aug 21. PMID: 20732724.

[5] Austin PC, Tu JV, Ho JE, Levy D, Lee DS. Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. J Clin Epidemiol. 2013 Apr;66(4):398-407. doi: 10.1016/j.jclinepi.2012.11.008. Epub 2013 Feb 4. PMID: 23384592; PMCID: PMC4322906.

[6] Chicco, D., Jurman, G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Med Inform Decis Mak 20, 16 (2020). https://doi.org/10.1186/s12911-020-1023-5

[7] Alonso-Betanzos A, Bolón-Canedo V, Heyndrickx GR, Kerkhof PL. Exploring Guidelines for Classification of Major Heart Failure Subtypes by Using Machine Learning. Clin Med Insights Cardiol. 2015 May 21;9(Suppl 1):57-71. doi: 10.4137/CMC.S18746. PMID: 26052231; PMCID: PMC4441365.

[8] 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

[9] Murat F, Yildirim O, Talo M, Baloglu UB, Demir Y, Acharya UR. Application of deep learning techniques for heartbeats detection using ECG signals-analysis and review. Comput Biol Med. 2020 May;120:103726. doi: 10.1016/j.compbiomed.2020.103726. Epub 2020 Apr 8. PMID: 32421643.

[10] 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.

[11] 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.

[12] A.L. Goldberger, et al., “PhysioBank, PhysioToolkit, and PhysioNet - Components of a new research resource for complex physiologic signals,” Circulation, vol. 101, no. 23, 2000, pp. E215-E220.

[13] 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.