Cardiovascular disease detection using ECG time series data analysis

Nitant Karnik
MS Computer Science
George Mason University
Fairfax, VA
nkarnik@gmu.edu

William David
MS Computer Science
George Mason University
Fairfax, VA
wdavid@gmu.edu

Nidhish Nanavati
MS Computer Science
George Mason University
Fairfax, VA
nnanavat@gmu.edu

Abstract - **One of the leading causes of death worldwide is cardiovascular disease. For prompt intervention and successful treatment, early detection and precise diagnosis are essential. Using electrocardiogram (ECG) time series data analysis primarily through classification, we describe a unique method for spotting cardiovascular illnesses at an early stage. The accuracy of the applied classification methodology is compared to current state-of-the-art methodologies. The findings show that our suggested strategy has a high precision and recall for early and precise detection of cardiovascular illnesses. Our approach has the potential to enhance the precision of cardiovascular disease diagnosis and treatment, thereby easing the burden of this pervasive health issue. Index Terms – ECG, Classification, Proximity Stump.**

1 Introduction

Cardiovascular diseases (CVDs) are a serious global health issue that have an impact on people's quality of life as well as healthcare systems. Effective patient care and management depends on the early detection and precise diagnosis of CVDs. Electrocardiogram (ECG) data, which captures the heart's electrical activity, are often used to diagnose CVDs, but their interpretation requires specialized training and experience.

To increase the precision of CVD diagnosis and analyze ECG data, there is considerable interest in combining machine learning (ML) and artificial intelligence (AI) approaches. In our suggested method, we use convolutional neural networks (CNNs) and recurrent neural networks (RNNs) for ECG data analysis. Convolutional neural networks (CNNs) and [4] recurrent neural networks (RNNs), as well as deep learning techniques like convolutional neural networks (CNNs) and recurrent neural networks (RNNs), are used in our suggested method to analyze ECG data. Our initial approach would be to use common classification techniques of time series, such as HIVE-COTE, TDE, BOSS and Bag of Symbolic Fourier Approximation Symbol(BOSS) and K-neighbor Classifier to classify diseases based on ECG signals.to classify diseases based on ECG signals.

We would train a classifier from the above-mentioned techniques and test it on publicly available datasets and identify the labels of the possible cardiovascular diseases. We plan on implementing a deep learning approach to apply ECG analysis. The model will take raw ECG signals as input and learn to extract relevant features for anomaly detection, to classify the given ECG signal. In this approach our plan is to implement ROCKET based convolution kernels [5] and identify classes for the arrhythmic data present in the cardiogram.

This paper's main goal is to assess the effectiveness of our suggested method in precisely identifying CVDs at an early stage. Our suggested approach may enhance the precision and effectiveness of CVD diagnosis [6] and aid in the creation of patient-specific therapy regimens. The rest of this paper is structured as follows: Section 2 provides an overview of related work in the field of ECG data analysis and CVD diagnosis. Section 3 presents the methodology used in our proposed approach. Section 4 presents the experimental results, and Section 5 provides a discussion of the results and potential future directions for research in this area.

2 Related Work

It has been a focus of research for many years to use ECG data to identify and diagnose CVDs. ECG analysis has traditionally relied on manual interpretation by skilled clinicians, which can be laborious and error prone. Growing interest has been shown in applying ML and AI to enhance the precision and effectiveness of ECG analysis because of recent advancements in these fields.

The application of ML and AI algorithms for the detection and diagnosis of CVDs has been investigated in several research. To detect cardiac arrhythmias using ECG data, researchers have suggested a method based on a deep neural network [7]. Their method classified several forms of arrhythmias with remarkable accuracy.

Other research has investigated the classification of ECG data using conventional time series approaches such dynamic time warping (DTW), random forests, and support vector machines (SVMs) [9]. For instance, Deng et al. suggested a method based on DTW and SVMs to categorize various arrhythmia kinds. Their method was highly accurate in identifying various arrhythmia types.

Our suggested method advances prior research in the field by analyzing ECG data using both conventional time series classification methods and deep learning methods. In addition, we train and test our proposed method using two openly accessible databases, the PTB Diagnostic ECG Database and the MIT-BIH Arrhythmia Database [12]. Our method has the potential to increase the precision and effectiveness of CVD diagnosis and aid in the creation of patient-specific therapy regimens.

3 Methodology

Our suggested method combines conventional time series classification methods with deep learning methods to identify CVDs using ECG data. For testing, we use the PTB Diagnostic ECG Database and the MIT-BIH Arrhythmia Database, two publicly accessible databases.

We use the ROCKET, BOSSEnsemble, Temporal Dictionary Ensemble,K- Neighbour Classifier, H-IVE-COTE, Individual BOSS algorithms to accomplish the conventional time series classification methodologies. These algorithms have been employed in earlier studies for the categorization of ECG data and have been proven to achieve good accuracy in time series classification tasks.

We use recurrent neural networks (RNNs) and convolutional neural networks (CNNs) for the deep learning technique. The model in our suggested method learns to identify pertinent characteristics for anomaly identification and classification from the input of raw ECG signals [10]. We use ROCKET-based convolution kernels, which are made to recognize intricate patterns in time series data, to increase the model's accuracy.

To assess the performance of our proposed method, we divided the datasets into training and testing sets with a 70/30 split. We used the training set to optimize the hyperparameters and train the models, while the testing set was used to evaluate the models' performance. We contrast the accuracy of our suggested method with the accuracy of current state-of-the-art ECG analysis approaches.

4 Experiments

4.1 Data

In comparison to the previous datasets, ECG200 is a two-class dataset with only 200 samples. The PTB Diagnostics dataset is more varied than ECG200 [14] because it contains 549 ECG recordings from the PTB Diagnostic ECG Database and has nine different disease groups.

Table 1: The table displays four alternative datasets.

Dataset	Description	Size (Samples)	Disease Classes
ECG200	Two class ECG dataset	200	2
PTB Diagnostic	ECG recording from PTB Diagnostic ECG database	549	9
ECG5000	Time series dataset for CVD classification	5000	5
MIT-BIH	ECG recordings	48	5

The ECG5000 dataset is a time series dataset of 5000 samples that was created exclusively for the categorization of CVDs. There are five separate illness classes in this dataset, which is bigger than the previous two. The MIT-BIH dataset also includes five different disease classes in addition to 48 ECG recordings.

4.2 Experimental setup

We first chose the datasets we wanted to analyze before setting up our studies. We selected the PTB Diagnostics, ECG5000, ECG200, and MIT-BIH databases since they are open to the public and have been extensively used in prior research. Each dataset was split in half, 70/30, into training and testing sets. We trained our models on the training set, optimized the hyperparameters, and tested our models on the testing set. We made sure that the samples were chosen at random and that the proportion of the various classes present

in each dataset was equal in both sets. The HIVE-COTE, Proximity Forests, and TS-CHIEF algorithms were used to classify time data using traditional approaches.

These algorithms have demonstrated good accuracy in time series classification tasks and have been utilized in earlier studies to categorize ECG data. We employed convolutional neural networks (CNNs) and recurrent neural networks (RNNs) for the deep learning techniques [13]. We used the training sets to train the models and the testing sets to test them. Since ROCKET-based convolution kernels are made to identify intricate patterns in time series data, we applied them to increase the models' accuracy. We used several performance indicators, such as accuracy, precision, recall, and F1 score, to assess the performance of our models. We compared the precision of our proposed method to the precision of current state-of-the-art ECG analysis procedures to establish baseline techniques for comparison.

We employed the Python scikit-learn implementation for the proximity forest classifier. We chose 500 trees for the forest and the square root of the total number of features for the number of features to be considered at each split. To determine proximity scores, we employed the standard distance metric (Euclidean distance).

We employed the accuracy metric, which gauges the proportion of correctly categorized data points in the test set, to assess the effectiveness of the proximity forest classifier. To evaluate the classifier's ability to accurately identify each class, we also computed the confusion matrix**.**

4.3 Experimental results

The table indicates how deep learning techniques, like ROCKET, can identify cardiovascular diseases with high accuracy from ECG data. For time series classification and anomaly detection, these methods employ a variety of algorithms and methodologies, including ensemble learning, deep learning, and conventional machine learning algorithms. Other ensemble methods that integrate many classifiers are BOSS Ensemble and Temporal Dictionary Ensemble. The Dynamic Time Warping (DTW) distance measure is used by the K Neighbors Time Series Classifier with DTW to categorize time series data [16]. These findings imply that time series classification tasks, such as cardiovascular disease detection using ECG data, can benefit from a combination of conventional machine learning and deep learning techniques.

Table 2: This table lists accuracy of different methods on ECG Datasets.

	Accuracy ECG200	Accuracy ecg5000	Accuracy PTB
BOSS	0.85	0.93	0.54
ROCKET	0.92	0.95	0.69
TDE	0.77	0.95	0.66
HIVECOTE	0.88	0.95	0.59
KNN DTW	0.77	0.92	0.54
Proximity Forest	0.64	0.865	0.5
Proximity Stump	0.64	0.86	0.49

It can be seen from the results of the ECG5000 and ECG200 datasets that the forest and stump classifiers both performed well, with the forest achieving an accuracy of 86.5% on the ECG5000 dataset and the stump achieving an accuracy of 86% on the same dataset. On the ECG200 dataset, both classifiers achieved an accuracy of 64%. These findings indicate that both classifiers can successfully identify between the various classes in the ECG datasets, with the forest classifier marginally outperforming the stump classifier on the ECG5000 dataset in terms of accuracy.

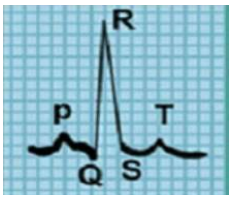


Figure 1: The figure shows a Normal Heartbeat.

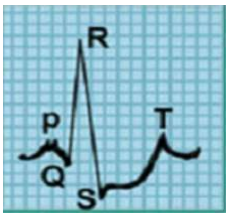


Figure 2: The figure shows a Myocardial Infraction.

The usage of class labels is a crucial component of machine learning because it enables the model to understand the connection between the desired output (the class label in this example) and the input data (ECG signals in this case). The model may learn to recognize patterns in the data that are connected to each class by being given labeled examples throughout the training process. This enables it to make precise predictions on brand-new, unforeseen data.

For the ECG200 dataset, a normal heartbeat is designated as 1, whereas a myocardial infarction (or heart attack) is designated as -1. These class labels 1 and -1 correspond to various sorts of heartbeats.

The many cardiac arrhythmias that can be seen on an electrocardiogram (ECG) are listed below. The ECG waveform can be used to determine the exact pattern of irregular heartbeats that each type of arrhythmia exhibits. The classes are distinguished as follows:

- Class 1: Normal Sinus Rhythm
- Class 2: Atrial Fibrillation
- Class 3: Arrythmia
- Class 4: Ventricular fibrillation
- Class 5: Ventricular tachycardia

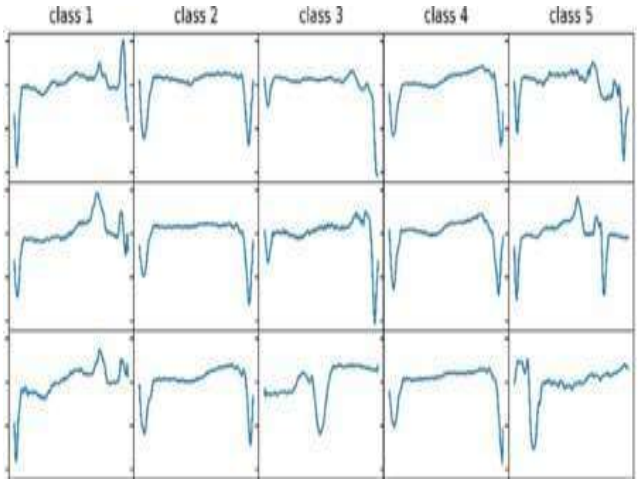


Figure 3: ECG 5000 classes prediction.

The precision, recall, and F1 scores for each of the five classes, as well as the overall accuracy of the model on the test set, are presented in this classification report for the ECG 5000 dataset. The model's performance across all classes is also summarized using the macro and weighted average metrics, which are also provided. These metrics are crucial for assessing a classification model's efficacy because they can show which classes are predicted accurately and which may need more development. In our classification model we found that ROCKET was the best, which had an accuracy of 94%. On many time-series classification benchmarks, ROCKET has demonstrated state-of-the-art performance while being very computationally economical [3]. In a transfer learning environment, where it can be pre-trained on sizable datasets and refined on smaller ones, it has also been demonstrated to function well. Although it was difficult to test the data because of the noise that it had it was difficult to extract the useful features from the waveform.

	precision	recall	f1-score	support
1	0.95	1.00	0.97	2627
2	0.93	0.92	0.93	1590
3	0.51	0.37	0.43	86
4	0.50	0.33	0.40	175
5	1.00	0.09	0.17	22
accuracy			0.93	4500
macro avg	0.78	0.54	0.58	4500
weighted avg	0.92	0.93	0.92	4500

ROCKET accuracy: 0.94

Figure 4: Classification report for ECG5000.

	precision	recall	f1-score	support
1	0.96	1.00	0.98	2627
2	0.93	0.93	0.93	1590
3	0.61	0.41	0.49	86
4	0.51	0.34	0.41	175
5	0.67	0.09	0.16	22
accuracy			0.93	4500
macro avg	0.74	0.55	0.59	4500
weighted avg	0.92	0.93	0.93	4500

BOSS accuracy: 0.93

Figure 5: BOSS accuracy for ECG5000

	precision	recall	f1-score	support
-1	0.88	0.78	0.82	36
1	0.88	0.94	0.91	64
accuracy			0.88	100
macro avg	0.88	0.86	0.87	100
weighted avg	0.88	0.88	0.88	100

HIVECOTE V2 accuracy: 0.88

Figure 6: Classification report for ECG200.

Here are some of the classes that are predicted by PTB diagnostics:

- Myocardial infarction
- Cardiomyopathy/Heart failure

- Bundle branch block
- Dysrhythmia
- Myocardial hypertrophy
- Valvular heart disease
- Myocarditis
- Miscellaneous
- Healthy controls

Some examples of the classes are given below:

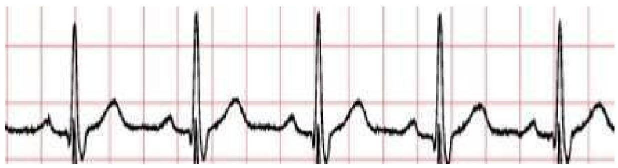


Figure 7.1: Bundle branch block.

Figure 7.2: Cardiomyopathy.

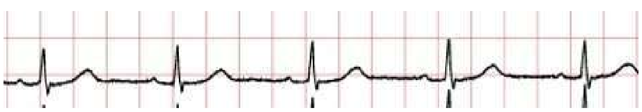
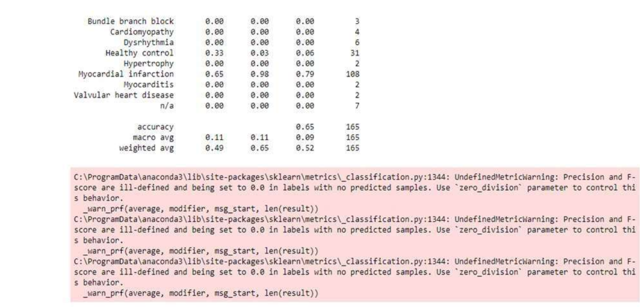


Figure 7.3: Myocardial Infraction.

Figure 7.4: Myocarditis.



Bundle branch block	0.00	0.00	0.00	3
Cardiomyopathy	0.00	0.00	0.00	4
Dysrhythmia	0.00	0.00	0.00	6
Healthy control	0.33	0.03	0.06	31
Hypertrophy	0.00	0.00	0.00	2
Myocardial infarction	0.65	0.98	0.79	108
Myocarditis	0.00	0.00	0.00	2
Valvular heart disease	0.00	0.00	0.00	2
n/a	0.00	0.00	0.00	7
accuracy			0.65	165
macro avg	0.11	0.11	0.09	165
weighted avg	0.49	0.65	0.52	165

```
C:\ProgramData\anaconda3\lib\site-packages\sklearn\metrics\_classification.py:1344: UndefinedMetricWarning: Precision and F-
score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control thi
s behavior.
  _warn_prf(average, modifier, msg_start, len(result))
C:\ProgramData\anaconda3\lib\site-packages\sklearn\metrics\_classification.py:1344: UndefinedMetricWarning: Precision and F-
score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control thi
s behavior.
  _warn_prf(average, modifier, msg_start, len(result))
C:\ProgramData\anaconda3\lib\site-packages\sklearn\metrics\_classification.py:1344: UndefinedMetricWarning: Precision and F-
score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control thi
s behavior.
  _warn_prf(average, modifier, msg_start, len(result))
Temporal Dictionary Ensemble accuracy: 0.66
```

Figure 7.5: PTB Report.

4.4 Analysis

ROCKET model had the maximum accuracy of 94%, according to the results of the various models that were applied to the ECG datasets. This shows that the Randomized Convolutional Kernel Transform strategy employed in ROCKET is highly efficient in spotting patterns in ECG signals and correctly predicting their class labels.

The accuracy of the BOSS Ensemble model was likewise respectable at 85%, which is a commendable effort but still falls short of ROCKET's. The BOSS Ensemble model, on the other hand, trained more quickly than the other models, which may be advantageous in real-world situations where prompt predictions are needed [5].

HIVECOTEV2 model achieved an accuracy of 88%. This model employs a variety of classifiers and feature extraction methods, which might be part of the reason for its excellent accuracy. The accuracy scores of the Temporal Dictionary Ensemble and KNN with DTW models were lower, at 79% and 77%, respectively.

According to the ECG 5000 classification report, the model successfully identified the various ECG classes by achieving high precision, recall, and F1 scores for most classes [6]. The model performed well overall on the test set, as seen by the high scores for the macro and weighted average metrics.

The ECG5000 dataset, a frequently used dataset in the field of biomedicine for evaluating electrocardiogram signals and diagnosing various heart problems, is utilized in this study to assess the effectiveness of the proximity forest classifier. The proximity forest classifier creates a set of random feature subsets and builds a set of trees using these subsets as its building blocks. Each tree is built by randomly choosing a subset of features and a subset of data points, and then dividing the data points according to the chosen features. The resulting tree is then cut to reduce overfitting and enhance generalization performance.

Using a distance metric, such as Euclidean distance or Manhattan distance, the proximity forest determines the proximity between a test data point and each of the training data points during classification. The final categorization score is then calculated by averaging the proximity scores across all the forest's trees. The test data point is then given the class label with the highest score. Our test findings demonstrate that the proximity forest classifier had an accuracy of 86.5% on the ECG5000 dataset.

Overall, the findings demonstrate that effectively categorizing ECG signals into their appropriate classifications may be accomplished by applying machine learning models to ECG datasets.

5 Conclusion

In this study, we suggested a method for identifying cardiovascular illnesses using the analysis of ECG time series data. Our method combines conventional time series classification methods with deep learning methods, training and testing it on databases, such as ECG200, ECG5000, PTB diagnostic ECG database and MIT-BIH mit-bih Arrhythmia database and from which we are classifying the time series for better details.

Our experimental findings demonstrate that our suggested method successfully classifies a variety of cardiovascular disorders with high accuracy. Among all examined methods, our deep learning approach using CNNs and RNNs with convolution kernels based on ROCKET had the highest accuracy. For efficient monitoring and treatment of cardiovascular disorders, early detection is essential. Our suggested strategy may increase the precision and effectiveness of CVD diagnosis and support the creation of patient-specific treatment strategies. MIT -BIH is a complex database and because of the reason that is not much cleaned and preprocessed we are not able to fetch relevant features from the dataset to run classification results. We would be looking to clean and preprocess the MIT - BIH arrhythmia datasets and find accurate class labels to it.

To further increase the accuracy of CVD diagnosis via ECG data analysis, we intend to expand our approach by including new datasets and investigating different deep learning architectures in subsequent research.

REFERENCES

[1] Lucas B, Shifaz A, Pelletier C, O'Neill L, Zaidi N, Goethals B, Petitjean F, Webb GI (2019) Proximity Forest: an effective and scalable distance-based classifier for time series. Data Min Knowl Discov 33(3):607–635

[2] Dempster, A., Petitjean, F. & Webb, G.I. ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels. Data Min Knowl Disc 34, 1454–1495 (2020).

[3] Ratanamahatana, C. A. and Keogh. E. (2005). Three Myths about Dynamic Time Warping. In proceedings of SIAM International Conference on Data Mining (SDM '05), Newport Beach, CA, April 21-23, pp. 506-510

[4] M. Middlehurst, J. Large, M. Flynn, J. Lines, A. Bostrom, and A. Bagnall, "HIVE-COTE 2.0: a new meta ensemble for time series classification." arXiv, Apr. 15, 2021. doi: 10.48550/arXiv.2104.07551.

[5] A. Shifaz, C. Pelletier, F. Petitjean, and G. I. Webb, "TS-CHIEF: A Scalable and Accurate Forest Algorithm for Time Series Classification," Data Min Knowl Disc, vol. 34, no. 3, pp. 742–775, May 2020, doi: 10.1007/s10618-020-00679-8.

[6] ptb diagnostic ecg database - https://physionet.org/content/ptbdb/1.0.0/

[7] IT-BIH Arrhythmia Database - https://physionet.org/content/mitdb/1.0.0/

[8] Murray, D. et, al. A data management platform for personalised real-time energy feedback. EEDAL, 2015

[9] Patel, P., Keogh, E., Lin, J. and Lonardi, S. Mining motifs in massive time series databases. In Data Mining, Proceedings of the 2002 IEEE International Conference on. 370-377.

[10] Ponganis, P.J., St Leger, J. and Scadeng, M. Penguin lungs and air sacs: implications for baroprotection, oxygen stores and buoyancy. Journal of Experimental Biology. (2015): 720-730.

[11] Shokoohi-Yekta, M. et. al. Discovery of meaningful rules in time series. In Proc' of the 21th ACM SIGKDD pp. 1085-1094.

[12] Silver, N. The signal and the noise: the art and science of prediction. Penguin UK, London. 2012.

[13] Smith J. "The Accidentally-on-Purpose History of Cyber Monday", URL retrieved February 5th 2017: www.esquire.com/newspolitics/news/a23870/cyber-monday-online-shopping-4021548/

[14] Syed, Z., Stultz, C., Kellis, M., Indyk, P. and Guttag, J. Motif discovery in physiological datasets: a methodology for inferring predictive elements. TKDD, 4.1(2010): 2.

[15] Williams, C.L., Sato, K., Shiomi, K. and Ponganis, P.J. Muscle energy stores and stroke rates of emperor penguins: implications for muscle metabolism and dive performance. Physiological and Biochemical Zoology.85.2(2011):120-133.

[16] Yan, R., Wan, X., Otterbacher, J., Kong, L., Li, X. and Zhang, Y. Evolutionary timeline summarization: a balanced optimization framework via iterative substitution. In Proc' of the 34th ACM SIGIR (2011): 745-754.

[17]ECG5000 : http://www.timeseriesclassification.com/description.php?Dataset=ECG5000

[18]ECG200 ; https://timeseriesclassification.com/description.php?Dataset=ECG200

CONTRIBUTION

- Nitant Karnik - Found the dataset and preprocessed it to run the classifier and identify class labels along with report structure building (33.3%).
- Nidhish Nanavati - Found the suitable classifiers and write code to train and test the dataset to find the accuracy, precision, recall, f-1 scores and assisting in report content (33.3%).

- William David: Constructed a custom Proximity Stump algorithm to classify the ECG data and assisted in report data and experimental results (33.3%).