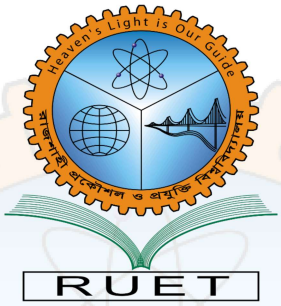
Heaven's Light is Our Guide

RAJSHAHI UNIVERSITY OF ENGINEERING AND TECHNOLOGY



Course Name: Biomedial Engineering Sessional

Course Code: ECE 4144

Project Title:

A simplified approach for Arrhythmia Detection using Automated Machine Learning Techniques

Submitted to	Submitted by
Md Mayenul Islam	Mohtashim Fuad (2010017)
Assistant Professor	MD. Mahfuzul Alam (2010030)
Dept: EEE	Dept: ECE
RUET	RUET

A Simplified Approach for Arrhythmia Detection Using Automated Machine Learning Techniques

Mohtashim Fuad
Department of Electrical & Computer Engineering
Rajshahi University of Engineering & Technology
Email: 2010017@student.ruet.ac.bd

Md. Mahfuzul Alam
Department of Electrical & Computer Engineering
Rajshahi University of Engineering & Technology
Email: 2010030@student.ruet.ac.bd

***Abstract*—This paper is about an automatic system to detect the alarming heart disease arrhythmia using machine learning. Arrhythmia is a type of heart disease which is alarming now a days. As the advancement of technology, diseases are also getting advanced. So it is needed some advanced system to detect diseases. And as a result it is very much important to detect a disease like arrhythmia in a low cost and efficiently. So, here comes an automatic system using machine learning algorithms to detect arrhythmia. The machine learning model used in this paper is random forest classifier model and the dataset is MIT-BIH arrhythmia dataset. From the raw ecg signals some important features were extracted using wfdb library and then applying random forest classifier algorithm on that features we found out that the model is performing impressive for this dataset. We have found more than 98% accuracy with this model. This is important for the future physicians to detect arrhythmia with less effort and efficiently.**

***Index Terms*—Diabetic Retinopathy, Deep Learning, Extreme Learning Machine, Parallel CNN, Medical Imaging, Multi-Class Classification**

I. Introduction

Heart is the one of the most important organs of a human body as it is like a pumping engine which supplies blood to all the other parts of the body by pumping it [1]. According to World Health Organization, heart disease is the main cause of death in the year 2016 globally [2]. About 17.9 million people died due to cardiovascular diseases which reflects 31% of all global deaths. And 67% of them were from comparatively lower and middle income countries [2]. In fact it has also been estimated that cardiovascular disease will be the major cause of death by the year 2035 and this will be more effective in comparatively lower income countries [3]. Even more alarming thing is that cardiovascular disease still is the most common reason for death in developed countries despite of all the advancement of medical field [4]. More than 450,000 people died due to cardiac arrest in United States of America only [4]. So it is very much essential to detect heart disease to protect human lives. On that point, the most popular way to detect or diagnose heart problems is by Electrocardiogram (ECG or EKG) in the present world. Since twentieth century, ECG has been used at the root of cardiovascular disease diagnosis [5] [6]. Basically ECG is the muscular and electrical signals or representation of the contraction and relaxation activity of the heart [6] [7]. These signals are captured by electrodes leads which are placed on the skin. Here the signal is measured by calculating the potential difference between two electrodes leads placed on the skin [8]. During polarisation and depolarization of the articles and ventricles the electrical pulses are captured by those electrodes [5] [6] [7]. The ECG signal morphology consists of a P wave firstly then a QRS complex and finally a T wave [8]. In QRS complex R is the most peak point, as a result it is called as R peak. So here, heart beat can be easily calculated by measuring the R peaks coming in a minute [6].

In the figure 2 below, the waves of ECG signal is shown [6].

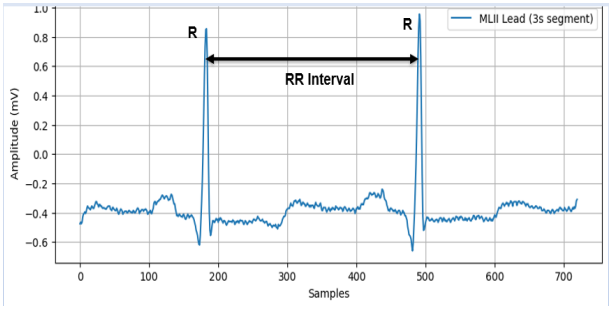


Fig. 1. ECG waveform

And in the figure 1 different intervals and different segments of ECG signal are shown [1].

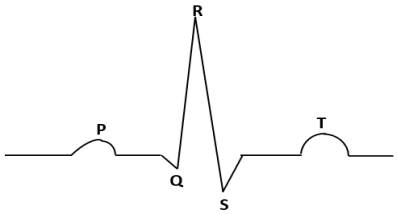


Fig. 2. ECG waveform

Basically arrhythmia is the abnormal condition of heart or mainly the irregular beat of heart [7]. Following that, this disease arrhythmia can be classified into different categories by measuring the beats of the heart. The normal heart beat for a healthy person is around 50-100 bpm. Carrying on that when the bpm of a person becomes less than 50 bpm then it

is referred as bradycardia. On the contrary when it is more than 100 bpm, it is termed as tachycardia [7]. For this ECG facilities it has become comparatively comfortable for the physicians to detect this risky disease. But this manual study is not still enough to overcome with the increasing risks as it is time consuming and also may become inaccurate [3]. For this reason it is really necessary for the automatic detection of this disease. In now a days the open access of ECG database makes it a great opportunity for us to develop systems for automatic detection of arrhythmia and also it has helped to develop many systems like that over the last decades [5]. And almost all of these systems have commonly four major steps. Those are preprocessing, feature extraction, model training and evaluation. [5] So following those steps in this project also the automatic system is developed using random forest algorithms of machine learning was developed.

II. Literature Review

For automatic detection of cardiovascular diseases various types of automatic systems have been developed using machine learning and deep learning algorithms. On that interest machine learning is playing a very essential role in detection of this diseases to reduce life risks or prediction of this life killing arrhythmia [9]. With these also deep learning algorithms are making a great impact in this field specially a famous algorithm called Convolutional Neural Networks (CNNs) [6]. The types of arrhythmia can be classified by measuring different features like QRS interval, bpm, RR interval and segments [10]. And these features are extracted using Discrete Wavelet Transform (DWT) [10]. A signal can be presented as a set of wavelets by using DWT which in future can be scaled according to the required frequency [10].

Here it is described several works done by different machine learning algorithms in this field. Among those, using a hybrid model of SVM, KNN and random forest algorithms an automatic system was developed to detect and diagnose arrhythmia which is offering around 97.6% accuracy [9]. Notably a seperate investigation was done using SVM which gives us 95.92% accuracy [7]. Building upon this momentum a study was done with accuracy of around 92% using a deep learning model named conventional artificial neural network (ANN) [6]. Even more impressively another one work was done using supervised SVM, KNN, Random Forest and ensemble of these three with accuracy of 83% [1]. In addition another work done using SVM with 98% accuracy [11]. Equally strike another work was done on MIMIC-III database with around 97% accuracy performed using random forest classifier [4]. Again another one work was done using a hybrid ensemble of SVM and Random Forest with having around 98.21% accuracy [3]. Also with F1 score of 0.94 a work was done using deep neural network [2]. With that a work with accuracy of around 91% done with SVM [10]. Then using a feature named amplitude difference a work done with random forest achieving around 98.68% accuracy using random forest [12]. Again another work also done with random forest having almost 100% accuracy where signals are obtained from BIDMC Congestive Heart Failure and PTB diagnosis ECG databases [13]. These are some of the works done in this field using models like SVM, KNN, Random Forest, ANN, LightGBM, Echo State Network, deep learning etc. with impressive accuracy. Except these there are also many many studies done in this topic.

III. Methodology

Here in this section the materials and methods used in this project are described. And also the evaluation metrics used to evaluate the performance of the model are discussed here. Finally the results are analyzed in this section.

A. Materials

Some important materials are used in this thesis to develop the automatic arrhythmia detection system. Those are discussed here.

1) Dataset: This dataset is known as MIT-BIH arrhythmia dataset which took five years to be fully prepared and having 46 records collecting from 44 patients sampled for over 30 minutes. And also this dataset offers with 48 half-hour ambulatory recordings [8] [14]. In this dataset recordings of 25 male subject aged between 32 to 89 years and 22 female subject aged between 23 to 89 years have been taken [14]. This dataset is often called as the golden standard dataset for arrhythmia detection [5]. In this dataset signals are taken at a frequency of 360 Hz [5] [6]. It has almost 15 sets of classified heart beats [3] and. each of the ecg signals are recorded using two leads [5] [6].

2) Used Library: Here we have used Python programming language for developing the system as we needed wfdb and scipy libraries of python. Wfdb library is used to extract features from the raw ECG signals. This library is very much popular for processing, analyzing and visualizing the physiological signals. It is also used for reading, writing and processing the PhysioNet's databases [15]. And scipy is used for scientific computing. This library is also used for optimization, integration, interpolation, eigenvalue problems [16].

3) Machine Learning Model: In this project Random Forest Classifier algorithm is used which is a supervised machine learning algorithm [17].

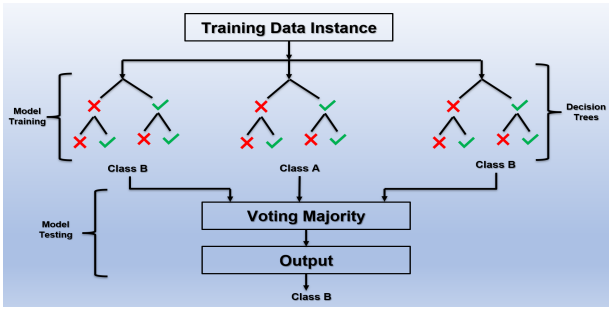


Fig. 3. Random Forest Representation

Basically this algorithm uses many decision trees to predict the output [18] [19]. In figure 3 we can see that how it is voting from all the decision trees and then showing the output [18]. And all the trees are for different parts of the data or for different features or columns and the final decision is taken by averaging or voting of all the trees. This system helps in reducing error and improving accuracy [18] [19]. A large amount of dataset is divided into many small parts by these decision trees to predict the target value [4]. These splitting of data is continued until it reaches the maximum depth or when there is no further division is possible [4].

B. Procedure

Fig. 4. Flowchart of the whole proceudure

This project was conducted following the exact flowchart shown in figure 4 [8]. As MIT-BIH arrhythmia dataset is composed of raw ECG signals, [8] [6] [5] we can't use random forest classifier algorithm here directly as this algorithm works for tabular data not for raw signals [18] [19] [20] [17].

1) Feature Extraction: As the model used for this project is for tabular data, it is must to convert the dataset into a tabular dataset [18] [17]. So we needed to collect or extract features from the dataset and then make a tabular data table with those features [12]. As a result we needed to extract time domain, frequency domain, statistical features & heart rate variablity features [8].

- **Time Domain Features:** RR interval, Heart rate, Mean Amplitude, Standard Amplitude, Maximum Amplitude, Difference between maximum and minimum amplitude, Number of times signal crosses zero [21] [22].
- **Heart Rate Variablity Features:** Root mean square of successive RR differences, Count of successive RR differences, Normalized version of the previous one, Standard deviation of RR intervals [22].
- **Statistical Features:** Skewness, Kurtosis, Slope, Area under the curve [22].
- **Frequency Domain Features:** High frequency power, Low frequency power, Ratio of high and low frequency [21] [22].

Figure 5 is the signal of record 100 from the dataset.

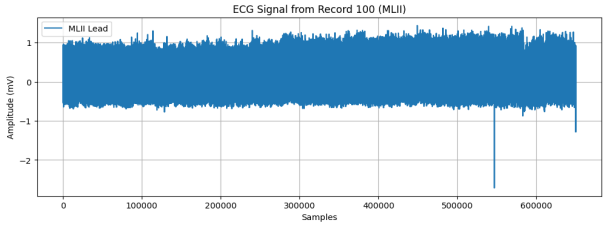


Fig. 5. ECG signal of record 100

Figure 6 is the readable portion of record 100 from the dataset which is the signal recorded for first 10 seconds. And different features from this signal can be extracted, as like RR interval for this portion is 288, Heart Rate is 75, Mean amplitude is -0.31992222222222216 etc.

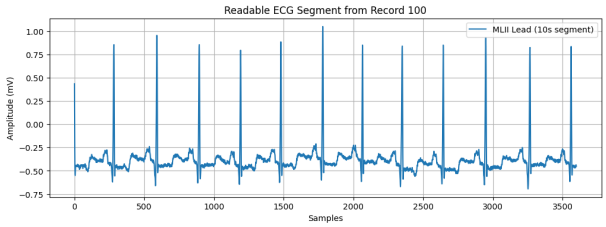


Fig. 6. Readable ECG segment of record 100

After extracting features from the dataset a tabular dataset is formed. Here in figure 7 first six columns of the first five rows of the dataset is shown.

lf_power	hf_power	lf_hf_ratio	slope	auc	peak_to_peak	zero_crossings	label
0.0	0.0	0	-0.001300	32.987083	1.325	3	N
0.0	0.0	0	-0.001825	38.158333	1.475	3	N
0.0	0.0	0	-0.001200	39.338750	1.530	3	N
0.0	0.0	0	-0.001300	40.680000	1.505	3	N
0.0	0.0	0	-0.001775	38.935833	1.385	3	N

Fig. 7. First five row after feature extracion

In the table it is seen that the label column is showing some objects 'N', so we need to label it to make it favourable for random forest classifier algorithm. And we can see figure 8 showing the last seven columns including labels of the first five row of the dataset after labelling. After labelling all the normal labels are taken as '9'.

lf_power	hf_power	lf_hf_ratio	slope	auc	peak_to_peak	zero_crossings	label
0.0	0.0	0	-0.001300	32.987083	1.325	3	9
0.0	0.0	0	-0.001825	38.158333	1.475	3	9
0.0	0.0	0	-0.001200	39.338750	1.530	3	9
0.0	0.0	0	-0.001300	40.680000	1.505	3	9
0.0	0.0	0	-0.001775	38.935833	1.385	3	9

Fig. 8. First five row after labelling

2) Model Training: At first it was needed to split the whole dataset into training set and test set, so for this project 80% of the data are taken for training and rest 20% are for testing puproses. And then we need to apply random forest on those training data.

We know that random forest algorithm combines the output of all the decision trees and then predict the result whether it is

regression or classification [4] [18] [23]. So at first it is needed to know about what a decision tree is and how it predicts output. Decision tree uses a tree which is like a flowchart as shown in figure 9 to predict [23]. This tree is composed of root node, decision node and leaf node [23]. The node from where the division starts is root node, and after the division we get decision node and the last node is leaf node [23].

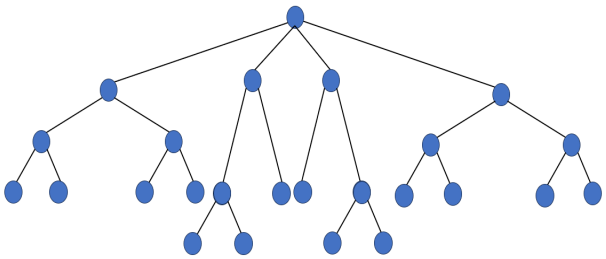


Fig. 9. Decision tree flowchart

And now to know which one is the root node we needed to know about gini impurity, the node which will have less gini impurity will be selected as the root node [4] [23]. Gini impurity is used to calculate the impurity of a dataset [24] [23] [4]. The gini impurity is calculated by using this formulae of gini index:

$$GiniIndex, G = 1 - \sum_{i=1}^{n} (p_i)^2 = 1 - [(p_+)^2 + (p_-)^2]$$

where, n is the total class, p_+ and p_- is the probability of a positive and a negative class respectively [4] [23] [24]. The maximum value of gini index is 0.5 which is considered as most impure and the lowest one is 0 which is the purest [24]. At first for a column the values are arranged in ascending order and then the mid points for all corresponding values are collected and with those mid points gini index is calculated [4] [23].

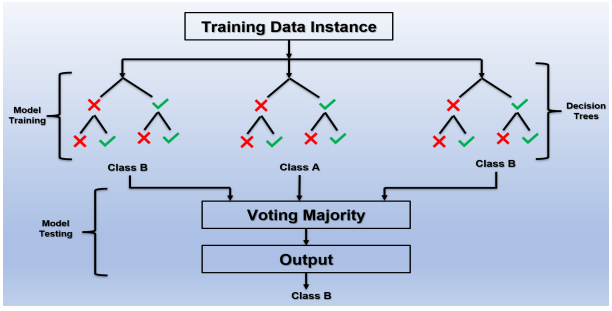


Fig. 10. Random Forest final output process

The split with the lowest gini impurity is taken as the root node and data is splitted according to that value [4]. This process goes on recursively for the other columns also which are taken as decision node and it continues till the leaf node [4]. This way a decision tree is formed for a column or feature and then rest decision trees are made following the same process for the rest of the features [4]. Finally a result come from all of those decision trees and the final output is obtained by majority voting from all those trees as like figure 10 [4] [23].

IV. Evaluation Metrices

Here evaluaiton like F1 score, accuracy, recall etc were calculated to determine the performance of the model used in the project [9] [13].

A. Accuracy

Accuracy defines the capacity of a model to predict different categories perfectly [9].

$$Accuracy = \frac{TN + TP}{TN + TP + FP + FN}$$

B. Recall

Recall is the quality of a model to avoid false negatives [9].

$$Recall = \frac{TP}{TP + FN}$$

C. Precision

Precision is the capacity of a model to avoid false positives [9].

$$Precision = \frac{TP}{TP + FP}$$

D. F1 Score

F1 score is the harmonic mean of recall and precision [9].

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Where, TP is Correctly detected components, TN is correctly rejected components, FP is incorrectly detected components, FN is incorrectly rejected components [9] [10].

V. Results Analysis

The performance of this model is evaluted by calculating the precision, recall, accuracy and F1 score. This model has achieved high accuracy, precision, recall and F1 score indicating it's ability to predict the results accurately.

In figure 11 confusion matrix is shown. This confusion matrix is generated across some arrhythmia classes. This matrix provides a distinct view of model accuracy. It shows high accuracy in detecting diffrent dominant classes. Among them one example is class 9 which detected 14,940 correctly which shows a very high accuracy in the model. In the same way, class 3, 8, 11, 12 have also shown a very high accuracy by predicting respectively 1377, 1602, 1389, 1471 correctly. And also class 2, 4, 6 & 15 have shown impressive work. All of them predicted more than 100 predictions correctly.

Missclassifications were very less for the majority of the classes. And most of the errors occured in between

sturcturally similar type of arrhythmias. So it suggests that further refinement in feature extraction will make it more effective and more accurate.

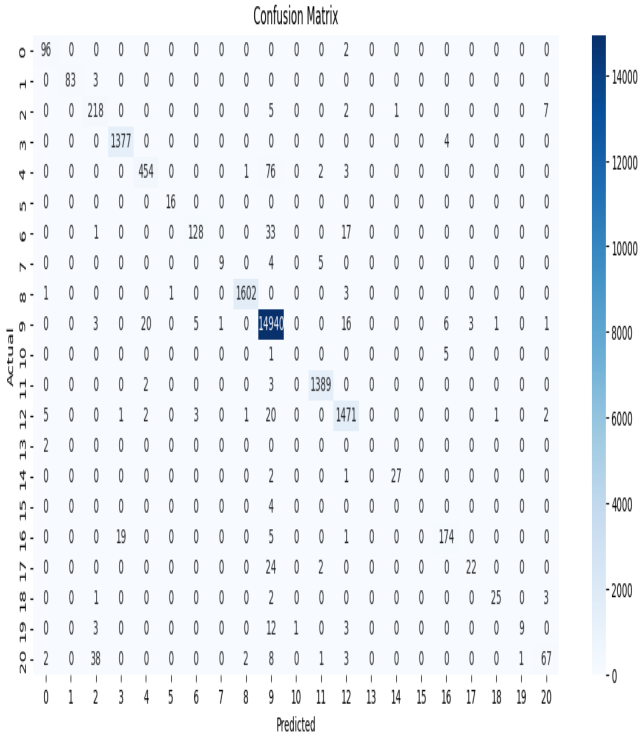


Fig. 11. Confusion Matrix

In the table I shown the precision, accuracy and F1 score for different classes.

Apart from that, this is also obtained that the model has achieved around 98% accuracy, which indicates that the model have predicted most of the predictions correctly almost all of the predictions. And also, precision, recall and F1 score in macro average is respectively 80%, 71% and 74%, which indicates the average performance of all classes and treating each of the classes equally. And the weighted precision, recall and F1 score is all 98% which indicates that the model has shown very impressive performance in the dominant classes.

And also from the table I we can see that class 8 has shown the best performace with precision, recall and F1 score of all 100% and support 1607. And after that class 3, 9 & 11 has shown the best performance with precision, recall and F1 score of 99%, 100&, 99% respectively with support 1381, 14996 and 1394 respectively. Apart from that class 0, 1, 5, 13 and & 16 have shown excellent performance which have precision, recall and F1 score over 90%. And also the dataset was tested with some other models and compared with the random forest model. Surprisingly random forest model performed better than those models like Decision tree, K nearest neighbors, Support Vector Machine, Logistic regression, XGBoost, Light Gradient Boosting Machine. The figure 12 shows the comparison of different machine learning models with random forest with this dataset in a chart.

Class	Precision	Recall	F1-Score	Support
0	0.91	0.98	0.94	98
1	1.00	0.97	0.98	86
2	0.82	0.94	0.87	233
3	0.99	1.00	0.99	1381
4	0.95	0.85	0.90	536
5	0.94	1.00	0.97	16
6	0.94	0.72	0.81	179
7	0.90	0.50	0.64	18
8	1.00	1.00	1.00	1607
9	0.99	1.00	0.99	14996
10	0.00	0.00	0.00	6
11	0.99	1.00	0.99	1394
13	0.97	0.98	0.97	1506
14	0.00	0.00	0.00	2
16	0.96	0.90	0.93	30
17	0.00	0.00	0.00	4
18	0.92	0.87	0.90	199
19	0.88	0.46	0.60	48
20	0.93	0.81	0.86	31
21	0.90	0.32	0.47	28
22	0.84	0.55	0.66	122

TABLE I
CLASSIFICATION REPORT

Here the chart in the figure 12 shows that random forest model performed best with the dataset used in this thesis. The accuracy is 98.2%, while accuracy of the other models are 95.72%, 97.59%, 97.42%, 90.75%, 95.44% and 62.04% for decision tree, k nearest neighbors, svm, logistic regression, xgboost and lightgbm respectively.

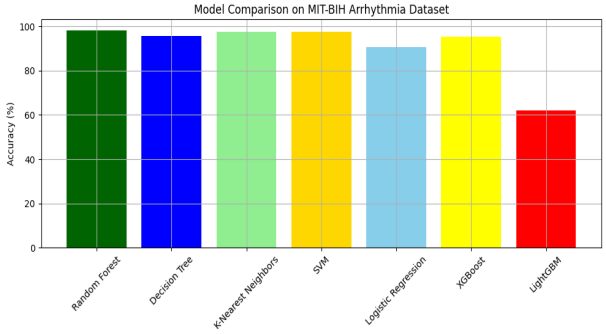


Fig. 12. Comparison of different machine learning model

VI. CONCLUSION

In this paper we discussed about detecting arrhythmia by using MIT-BIH arrhythmia dataset and using model random forest classifier. And this model have performed excellent with this dataset. It has shown around 98% accuracy and with 98% of weighted precision and F1 score. So it is an excellent performace. Still there was some short comings as the macro F1 score is not that much excellent, so there is a room for improvement in this arena. Some of the classes are not performing up to the level. So in later studies it can be improved by extracting features more efficiently.

REFERENCES

[1] M. Sraitih, Y. Jabrane, and A. Hajjam El Hassani, "An automated system for ecg arrhythmia detection using machine learning techniques," *Journal of Clinical Medicine*, vol. 10, no. 22, 2021.

[2] M. Hammad, A. M. Iliyasu, A. Subasi, E. S. L. Ho, and A. A. A. El-Latif, "A multitier deep learning model for arrhythmia detection," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–9, 2021.

[3] S. Bhattacharyya, S. Majumder, P. Debnath, and M. Chanda, "Arrhythmic heartbeat classification using ensemble of random forest and support vector machine algorithm," *IEEE Transactions on Artificial Intelligence*, vol. 2, no. 3, pp. 260–268, 2021.

[4] S. S. Yadav and S. M. Jadhav, "Detection of common risk factors for diagnosis of cardiac arrhythmia using machine learning algorithm," *Expert Systems with Applications*, vol. 163, p. 113807, 2021.

[5] M. Alfaras, M. C. Soriano, and S. Ortín, "A fast machine learning model for ecg-based heartbeat classification and arrhythmia detection," *Frontiers in Physics*, vol. Volume 7 - 2019, 2019.

[6] A. Isin and S. Ozdalili, "Cardiac arrhythmia detection using deep learning," *Procedia Computer Science*, vol. 120, pp. 268–275, 2017. 9th International Conference on Theory and Application of Soft Computing, Computing with Words and Perception, ICSCCW 2017, 22-23 August 2017, Budapest, Hungary.

[7] C. Usha Kumari, A. Sampath Dakshina Murthy, B. Lakshmi Prasanna, M. Pala Prasad Reddy, and A. Kumar Panigrahy, "An automated detection of heart arrhythmias using machine learning technique: Svm," *Materials Today: Proceedings*, vol. 45, pp. 1393–1398, 2021. International Conference on Advances in Materials Research - 2019.

[8] T. Zaharia, G. M. Danciu, I. Ilie, I. E. Nicolae, and S. C. Nechifor, "A simplified approach for accurate arrythmia detection using automated machine learning," in *2023 13th International Symposium on Advanced Topics in Electrical Engineering (ATEE)*, pp. 1–5, 2023.

[9] R. Kumar and Jyoti, "A hybrid approach using svm, knn and random forest for ecg classification," in *2023 11th International Conference on Intelligent Systems and Embedded Design (ISED)*, pp. 1–6, 2023.

[10] K. Subramanian and N. K. Prakash, "Machine learning based cardiac arrhythmia detection from ecg signal," in *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pp. 1137–1141, 2020.

[11] S. Dhyani, A. Kumar, and S. Choudhury, "Analysis of ecg-based arrhythmia detection system using machine learning," *MethodsX*, vol. 10, p. 102195, 2023.

[12] J. Park, S. Lee, and K. Kang, "Arrhythmia detection using amplitude difference features based on random forest," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 5191–5194, 2015.

[13] Z. Masetic and A. Subasi, "Congestive heart failure detection using random forest classifier," *Computer Methods and Programs in Biomedicine*, vol. 130, pp. 54–64, 2016.

[14] V. Singh, S. Tewary, V. Sardana, and H. K. Sardana, "Arrhythmia detection - a machine learning based comparative analysis with mit-bih ecg data," in *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*, pp. 1–5, 2019.

[15] "wfdb: The WFDB python package: tools for reading, writing, and processing physiologic signals and annotations.."

[16] "Introduction to SciPy."

[17] "Random Forest Algorithm - How it works and why it is so effective," 6 2022.

[18] GeeksforGeeks, "Random Forest algorithm in machine learning," 9 2025.

[19] Sruthi, "Random Forest algorithm in machine learning," 5 2025.

[20] "Random forest classification with scikit-learn."

[21] G. SharanYadav, S. Yadav, and P. Prachi, "Time and frequency exploration of ECG signal," vol. 67, no. 4, pp. 5–8.

[22] A. K. Singh and S. Krishnan, "ECG signal feature extraction trends in methods and applications," vol. 22, no. 1, p. 22.

[23] Anshul, "An introduction to random forest algorithm for beginners."

[24] "Splitting decision trees with gini impurity."