History of Medicine, 2021, 7(2): 183-193

DOI: 10.17720/2409-5834.v7.2.2021.10

# A SUPERVISED LEARNING-BASED APPROACH FOR PREDICTING CARDIOMYOPATHY DISEASE IN HEART PATIENTS' CARDIOVASCULAR HEALTH PREDICTION

**R Ramadevi,Ramesh Thokala,G Ashwini**

Assistance Professor Assistance Professor Assistant Professor

Department of CSE,

Sree Dattha Institute of Engineering and Science,

**ABSTRACT**

Cardiomyopathy is a chronic and often progressive heart condition characterized by abnormalities in the structure or function of the heart muscle. Early detection of cardiomyopathy is essential for effective management and treatment, as it can lead to life-threatening complications such as heart failure, arrhythmias, and sudden cardiac death. Conventional diagnostic methods primarily rely on clinical assessments, electrocardiograms (ECGs), and echocardiography, which may not always provide accurate predictions or early warnings. Moreover, these approaches tend to overlook the potential influence of genetic and lifestyle factors, which are increasingly recognized as critical contributors to cardiomyopathy risk. The conventional diagnostic system for cardiomyopathy suffers from several limitations. Clinical assessments, while valuable, often rely on subjective judgments and may not detect subtle changes in heart function until the disease has progressed significantly. ECGs and echocardiography can be more objective but may miss early-stage cardiomyopathy. Furthermore, these approaches typically do not consider genetic predispositions or lifestyle factors, which can significantly affect disease risk.

**Keywords:** Random Forest, ETC, Classification, Cardiomyopathy

## 1. INTRODUCTION

One American dies every 36 seconds due to CVD. More than.665 million people die due to heart disease which 1 in every 4 deaths. Cardiovascular disease costs a lot to the US healthcare system. In the years 2014 and 2015, it cost about $219 billion per year in terms of healthcare services, medicine, and lost productivity due to death. Early diagnosis can also help to prevent heart failure which can lead to the death of a person. Angiography is considered as the most precise and accurate method for the prediction of cardiac artery disease (CAD), but it is very costly which makes it less accessible to low-income families. It is difficult to identify heart disease because of several contributory risk factors such as diabetes, high blood pressure, high cholesterol, abnormal pulse rate and many other factors. Heart disease is a collection of diseases impacting the heart and veins of human beings. Cardiac disease symptoms vary depending on the specific type of cardiac disease. Detecting and diagnosing the cardiovascular disease is an on-going job that can be achieved with enough experience and knowledge by a qualified professional [1]. There are many factors including age, diabetes, smoking, overweight, junk foods diet and so on. Several factors/parameters have been identified that cause heart disease or increase cardiac disease. Most hospitals have management software for monitoring

their clinical and/or patient data. It is popular now and Such systems produce enormous amounts of patient information. These data are seldom used for clinical decision-making support [2]. These data are valuable, and information is kept largely unused in these data. It is an extremely difficult task to turn the accumulated clinical data into useful information that can make intelligent systems support decision-making for healthcare practitioners. This factor led to research on the processing of medical data due to the lack of experts and the number of cases incorrectly diagnosed, a rapid and efficient automated detection system was required. The main purpose is to classify the key features of the medical data using the classifier model and use the models for the early prediction of cardiac disease.

According to WHO, Heart Diseases are a leading cause of death worldwide. It is quite difficult to identify the cardiovascular disease (CVD) [3] because of some contributory factors which contribute to CVD like high blood pressure, cholesterol level, diabetics, abnormal pulse rate, and many other factors. Sometimes CVD symptoms may vary for different genders. For example, a male patient is more likely to have chest pain while a female patient has some other symptoms with chest pain like chest discomfort: such as nausea, extreme fatigue, and shortness of breath. Researchers have been exploring a wide range of techniques to predict heart diseases but the disease prediction at an early stage is not very efficient due to many factors [4], including but not limited to complexity, execution time, and accuracy of the approach. As such, proper treatment and diagnosis can save many lives.

Cardiomyopathy is a complex and multifaceted medical condition characterized by structural and functional abnormalities in the heart muscle. Cardiomyopathy is a significant public health concern worldwide. It can lead to heart failure, arrhythmias, and other life-threatening complications [5]. Understanding the underlying causes, mechanisms, and potential treatments for cardiomyopathy is crucial to improve the overall health and well-being of affected individuals and reduce the burden on healthcare systems.

Cardiomyopathy can result from various causes, including genetic mutations, infections, autoimmune diseases, metabolic disorders, and exposure to toxins or drugs. Researchers are motivated to study these different etiologies to develop targeted therapies and interventions based on the specific underlying factors. Genetic factors play a significant role in the development of certain types of cardiomyopathy [6], such as hypertrophic cardiomyopathy and dilated cardiomyopathy. Identifying the genetic mutations responsible for these conditions can lead to better risk assessment, early diagnosis, and potentially gene-based therapies. Advances in medical imaging, genetics, and molecular biology have improved our ability to diagnose and monitor cardiomyopathy. Researchers are motivated to explore the potential of these technologies to enhance early detection and personalized treatment strategies.

Improving the quality of life for individuals living with cardiomyopathy is a primary goal [7]. Research in this area seeks to address not only the physical aspects of the disease but also its psychosocial and emotional impacts on patients and their families. Understanding the risk factors and preventive measures for cardiomyopathy is essential. Lifestyle modifications, such as diet and exercise, as well as pharmacological interventions, can help reduce the risk of developing cardiomyopathy. Researchers are motivated to identify effective prevention strategies.

Cardiomyopathy affects people of all ages and backgrounds globally. The advocacy efforts of patient and caregiver organizations play a significant role in motivating research into cardiomyopathy [8].

## 2.LITERATURE SURVEY

Rani et. al [11] proposed a hybrid decision support system that can assist in the early detection of heart disease based on the clinical parameters of the patient. Authors have used multivariate imputation by chained equations algorithm to handle the missing values. A hybridized feature

selection algorithm combining the Genetic Algorithm (GA) and recursive feature elimination has been used for the selection of suitable features from the available dataset. Further for pre-processing of data, SMOTE (Synthetic Minority Oversampling Technique) and standard scalar methods have been used. In the last step of the development of the proposed hybrid system, authors have used support vector machine, naive bayes, logistic regression, random forest, and Adaboost classifiers. It has been found that the system has given the most accurate results with random forest classifier. The proposed hybrid system was tested in the simulation environment developed using Python. It was tested on the Cleveland heart disease dataset available at UCI (University of California, Irvine) machine learning repository. It has achieved an accuracy of 86.6%, which is superior to some of the existing heart disease prediction systems found in the literature.

Kavitha et. al [12] proposed a novel machine learning approach to predict heart disease. The proposed study used the Cleveland heart disease dataset, and data mining techniques such as regression and classification are used. Machine learning techniques Random Forest and Decision Tree are applied. The novel technique of the machine learning model is designed. In implementation, 3 machine learning algorithms are used, they are 1. Random Forest, 2. Decision Tree and 3. Hybrid model (Hybrid of random forest and decision tree). Experimental results show an accuracy level of 88.7% through the heart disease prediction model with the hybrid model. The interface is designed to get the user's input parameter to predict the heart disease, for which they used a hybrid model of Decision Tree and Random Forest.

Mohan et. al [13] proposed a novel method that aims at finding significant features by applying machine learning techniques resulting in improving the accuracy in the prediction of cardiovascular disease. The prediction model is introduced with different combinations of features and several known classification techniques. They produce an enhanced performance level with an accuracy level of 88.7% through the prediction model for heart disease with the hybrid random forest with a linear model (HRFLM).

Shah et. al [14] presents various attributes related to heart disease, and the model on basis of supervised learning algorithms as Naïve Bayes, decision tree, K-nearest neighbor, and random forest algorithm. It uses the existing dataset from the Cleveland database of UCI repository of heart disease patients. The dataset comprises 303 instances and 76 attributes. Of these 76 attributes, only 14 attributes are considered for testing, important to substantiate the performance of different algorithms. This research paper aims to envision the probability of developing heart disease in the patients. The results portray that the highest accuracy score is achieved with K-nearest neighbor.

Guoet. Al [15] proposed Recursion enhanced random forest with an improved linear model (RFRF-ILM) to detect heart disease. This paper aims to find the key features of the prediction of cardiovascular diseases through the use of machine learning techniques. The prediction model is adding various combinations of features and various established methods of classification. it produces a better level of performance with precision through the heart disease prediction model.

## 3. PROPOSED SYSTEM

### 3.1 Overview

Addressing critical issues through multidisciplinary research efforts will not only advance our understanding of cardiomyopathy but also lead to the development of more effective diagnostic and therapeutic strategies, ultimately improving the lives of individuals affected by this challenging cardiac condition. Figure 1 shows the block diagram of proposed system. The detailed operation is illustrated as follows:

Proposed k-Nearest Neighbours (KNN) Classifier: Implement a KNN classifier using the pre-processed training data. Experiment with different values of k (number of neighbours) to find the optimal value through cross-validation. Evaluate the KNN classifier's performance on the test dataset, employing the same evaluation metrics used for DTC and RFC.

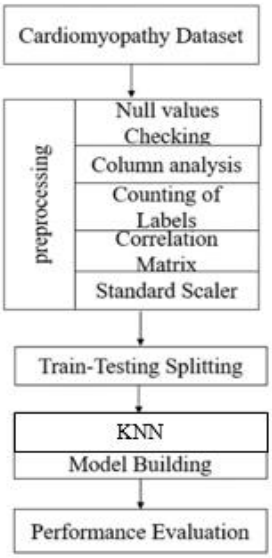


Figure 1: Proposed Block Diagram

**3.2 KNN**

K-Nearest Neighbours (KNN) is a simple yet powerful supervised machine learning algorithm used for classification and regression tasks. It's based on the idea that data points with similar features tend to belong to the same class or have similar values in the case of regression. KNN is a distance-based classification algorithm. It assigns a new data point to the majority class of its k-nearest neighbours. The choice of 'k' (the number of neighbours) is a crucial hyperparameter that impacts the model's performance. KNN is an instance-based learning method, meaning it doesn't build a model during training. Instead, it memorizes the entire training dataset and uses it for predictions.

**Working Principle:**

Step 1: Distance Metric:

- KNN uses a distance metric (typically Euclidean distance, but others like Manhattan, Minkowski, etc., are also possible) to measure the similarity between data points. The algorithm finds the 'k' nearest neighbours with the smallest distances to the new data point.

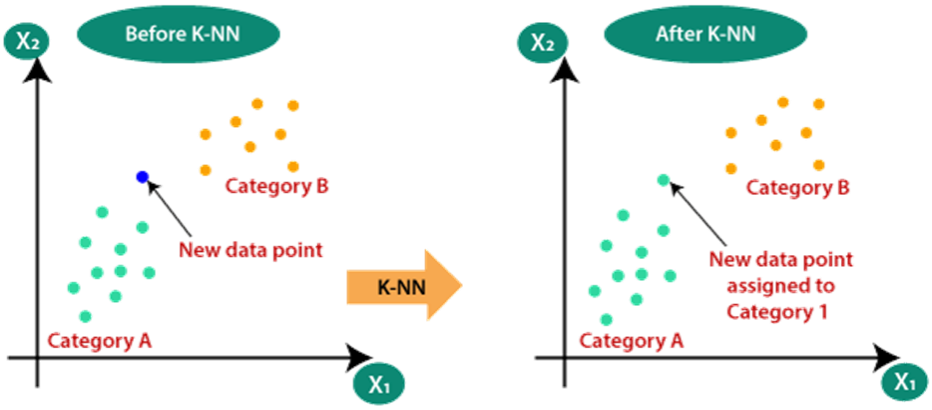


Fig. 2: KNN initialization

- Voting Mechanism: For classification, KNN uses a majority voting mechanism among its neighbours. The class that occurs most frequently among the neighbours is assigned to the new data point. For regression, it takes the mean (or median) value of the 'k' nearest neighbours as the prediction.

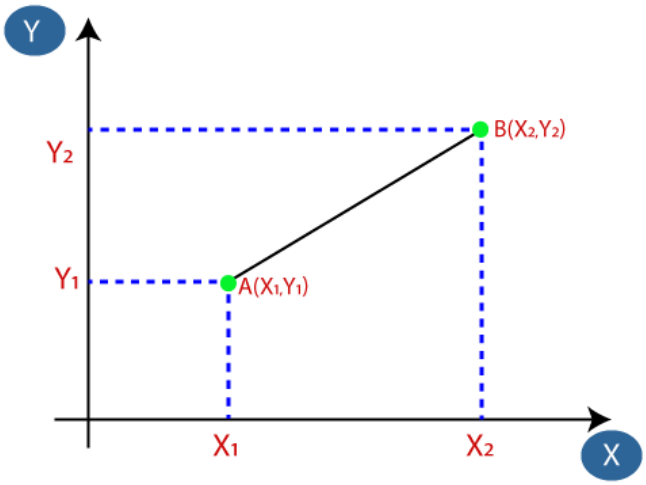


Fig. 3: Distance measurement in KNN

Step 2. Hyperparameter 'k':

- Choosing the Right 'k': The choice of 'k' is crucial. A small 'k' makes the model sensitive to noise and outliers but may capture local patterns well. A large 'k' smooths out local variations but can make the model less accurate.
- Methods for Choosing 'k': Cross-validation, grid search, and domain knowledge are common approaches to determine the optimal 'k' value.
- Simplicity: KNN is easy to understand and implement, making it a suitable choice for beginners.
- No Training Phase: It doesn't require a training phase since it memorizes the data, making it suitable for online learning and non-stationary data.

- Non-Parametric: KNN is non-parametric, meaning it makes no assumptions about the underlying data distribution.
- Works for Multiclass Problems: KNN naturally handles multi-class classification problems.

Step 3. Variants:

- Weighted KNN: Assigns different weights to neighbors based on their distance. Closer neighbors have a greater influence on the prediction.
- KNN with Feature Scaling: Feature scaling is essential when using KNN, as it's distance-based. Standardization (scaling features to have mean=0 and standard deviation=1) is often applied.

KD-Tree and Ball-Tree: These data structures can be used to speed up KNN search for large datasets.

## 4. RESULTS

This figure 4 displays the dataset uploaded by the user within the Cardiomyopathy Disease GUI. It provides a visual representation of the raw data, showcasing the structure and format of the dataset. Figure 5 presents a count plot depicting the distribution of categories within the dataset. Each category's frequency is represented by the height of the corresponding bar, offering insights into the class distribution before preprocessing.

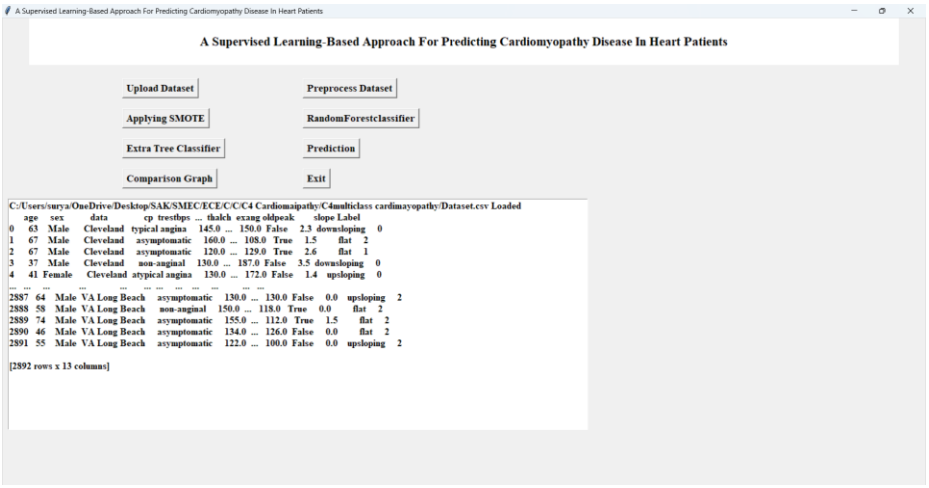


Figure 4: Displays the uploaded dataset in Cardiomyopathy Disease GUI.

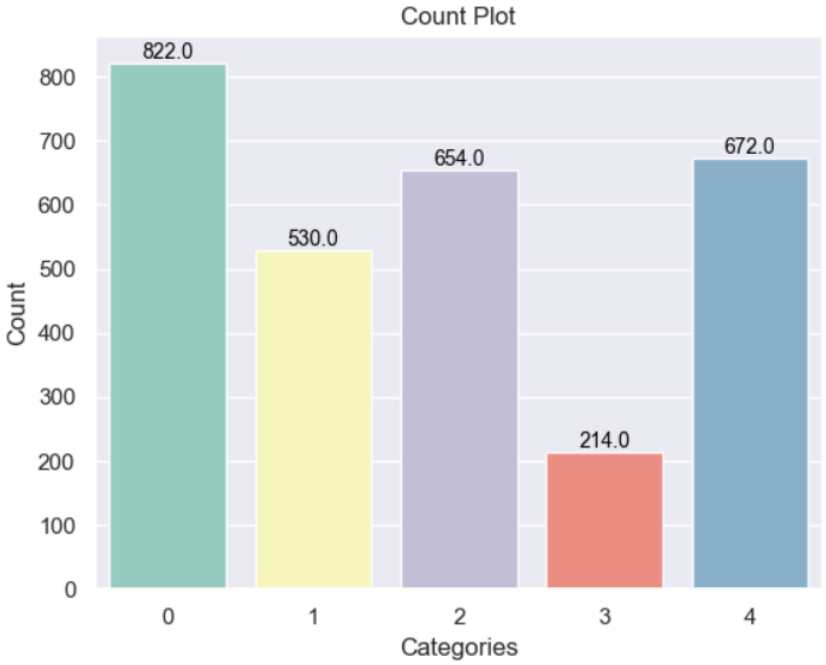


Figure 5: Presents the count plot of Categories.

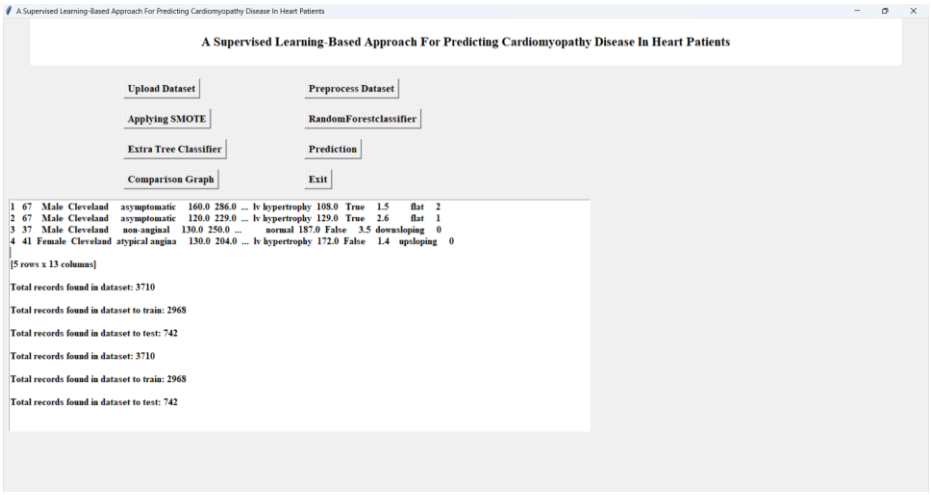


Figure 6: Shows the Preprocessing of the uploaded dataset.

This figure 6 illustrates the preprocessing steps applied to the uploaded dataset. It includes handling missing values, encoding categorical variables using LabelEncoder, and applying SMOTE to address class imbalance. The processed dataset is prepared for further analysis and model training. Figure 7 showcases the balanced count of categories within the dataset after applying SMOTE. By addressing class imbalance, SMOTE ensures that each category has a comparable representation in the dataset, thereby enhancing model performance and generalization.

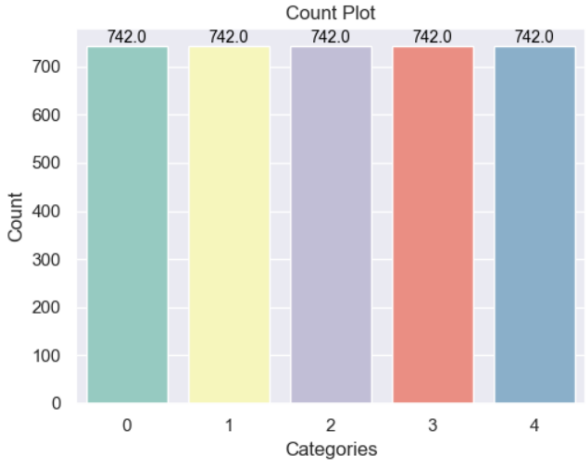


Figure 7: Presents the balanced count of Categories after smote.

Figure 8 presents the confusion matrix generated for the Random Forest Classifier model. The matrix illustrates the classifier's performance by depicting the true positive, true negative, false positive, and false negative predictions. It provides a detailed assessment of the model's accuracy and error rates. Figure 9 showcases the confusion matrix generated for the Extra Tree Classifier model.

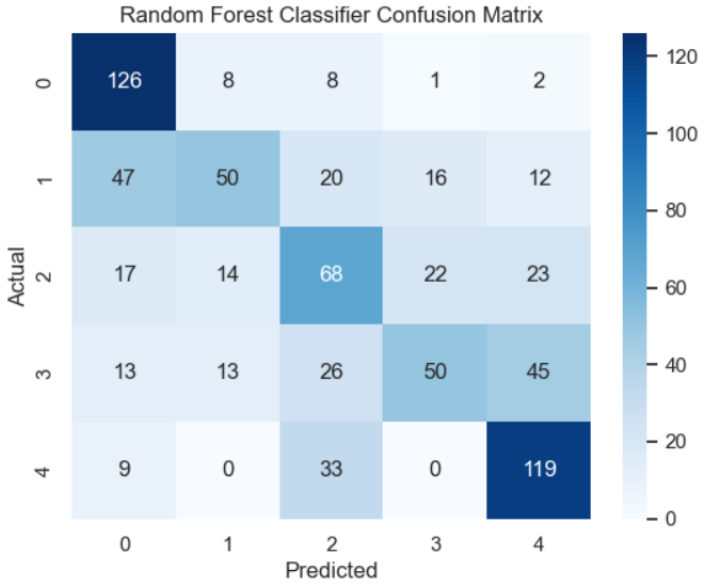


Figure 8: Confusion matrix of Random Forest Classifier model.

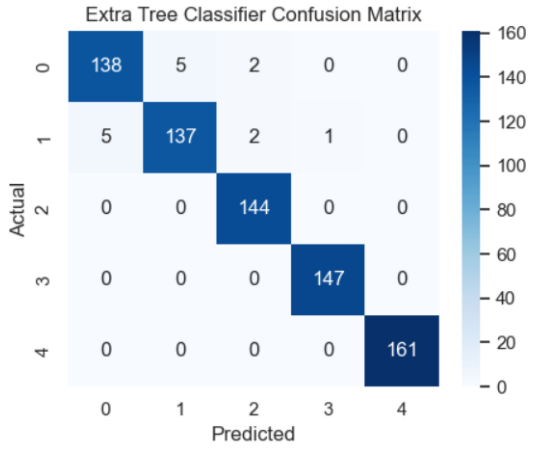


Figure 9: Confusion matrix of Extra Tree Classifier model.

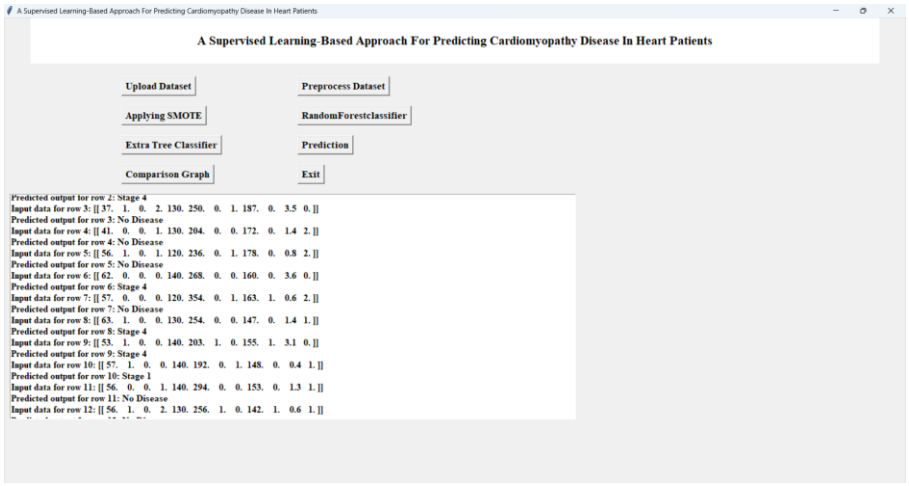


Figure 10: Proposed Model Prediction on test data.

Figure 10 displays the proposed model's predictions on test data. It showcases the model's ability to classify instances into different categories, aiding in the identification of potential cases of cardiomyopathy disease. Each prediction provides valuable information for medical practitioners and researchers in making informed decisions.

Table 1: Performance Evaluation of ALL models.

| **Metric** | **Random Forest Classifier** | **Extra Tree Classifier** |
|---|---|---|
| Precision | 55.50% | 97.92% |
| Recall | 55.31% | 97.93% |
| F1 Score | 53.53% | 97.92% |
| Accuracy | 55.66% | 97.98% |

**Precision**:

- Precision measures the proportion of true positive predictions among all positive predictions made by the classifier. For the Random Forest Classifier, it stands at 55.50%, indicating that when it predicts a positive class (cardiomyopathy disease), it is correct about 55.50% of the

time. The Extra Tree Classifier demonstrates significantly higher precision at 97.92%, indicating a high accuracy in identifying positive cases.

**Recall**:

- Recall, also known as sensitivity, calculates the proportion of true positive predictions among all actual positive instances in the dataset. The Random Forest Classifier achieves a recall rate of 55.31%, indicating that it correctly identifies 55.31% of all actual positive cases. Similarly, the Extra Tree Classifier achieves a recall rate of 97.93%, suggesting a high ability to capture positive cases.

**F1 Score**:

- The F1 score is the harmonic mean of precision and recall, providing a balance between the two metrics. For the Random Forest Classifier, the F1 score is 53.53%, indicating a balanced performance in terms of precision and recall. On the other hand, the Extra Tree Classifier demonstrates an F1 score of 97.92%, reflecting a high level of precision and recall combined.

**Accuracy**:

- Accuracy represents the proportion of correctly classified instances (both true positives and true negatives) among all instances in the dataset. The Random Forest Classifier achieves an accuracy rate of 55.66%, indicating the overall correctness of its predictions. In contrast, the Extra Tree Classifier exhibits a significantly higher accuracy rate of 97.98%, underscoring its effectiveness in classifying instances accurately.

## 5. CONCLUSION

In summary, the multifaceted analysis conducted in this study, which included in-depth Exploratory Data Analysis (EDA), meticulous data preprocessing, and a thorough evaluation of the Decision Tree Classifier (DTC), Random Forest Classifier (RFC), and the novel k-Nearest Neighbors (KNN) Classifier, was instrumental in unraveling the intricacies of the cardiomyopathy dataset. This comprehensive approach provided invaluable insights into the dataset's inherent characteristics, the nuanced performance of each classifier, and their respective strengths and limitations. The outcomes of this study not only advanced our understanding of how these machine learning models can be applied to cardiomyopathy classification but also illuminated potential avenues for further research and refinement. These findings have the potential to shape the future of cardiac health diagnostics by guiding the development of more accurate and effective tools for the early detection and management of cardiomyopathy, ultimately improving patient outcomes and healthcare practices in this critical domain.

## REFERENCES

[1] Bakar, Wan Aezwani Wan Abu, et al. "A Review: Heart Disease Prediction in Machine Learning & Deep Learning." *2023 19th IEEE International Colloquium on Signal Processing & Its Applications (CSPA)*. IEEE, 2023.

[2] Dileep, P., et al. "An automatic heart disease prediction using cluster-based bi-directional LSTM (C-BiLSTM) algorithm." *Neural Computing and Applications* 35.10 (2023): 7253-7266.

[3] Mishra, Nilamadhab, et al. "Visual Analysis of Cardiac Arrest Prediction Using Machine Learning Algorithms: A Health Education Awareness Initiative." *Handbook of Research on Instructional Technologies in Health Education and Allied Disciplines*. IGI Global, 2023. 331-363.

[4] Guo, Saidi, et al. "Survival prediction of heart failure patients using motion-based analysis method." *Computer Methods and Programs in Biomedicine* 236 (2023): 107547.

[5] Nandy, Sudarshan, et al. "An intelligent heart disease prediction system based on swarm-artificial neural network." *Neural Computing and Applications* 35.20 (2023): 14723-14737.

[6] Pant, Aman, et al. "Heart disease prediction using image segmentation Through the CNN model." *2023 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. IEEE, 2023.

[7] Nandy, Sudarshan, et al. "An intelligent heart disease prediction system based on swarm-artificial neural network." *Neural Computing and Applications* 35.20 (2023): 14723-14737.

[8] Rani, P., Kumar, R., Ahmed, N.M.O.S. et al. A decision support system for heart disease prediction based upon machine learning. J Reliable Intell Environ 7, 263–275 (2021). https://doi.org/10.1007/s40860-021-00133-6

[9] M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai and R. S. Suraj, "Heart Disease Prediction using Hybrid machine Learning Model," 2021 6th International Conference on Inventive Computation Technologies (ICICT), 2021, pp. 1329-1333, doi: 10.1109/ICICT50816.2021.9358597.

[10] S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[11] Shah, D., Patel, S. & Bharti, S.K. Heart Disease Prediction using Machine Learning Techniques. SN COMPUT. SCI. 1, 345 (2020). https://doi.org/10.1007/s42979-020-00365-y
C. Guo, J. Zhang, Y. Liu, Y. Xie, Z. Han and J. Yu, "Recursion Enhanced Random Forest with an Improved Linear Model (RERF-ILM) for Heart Disease Detection on the Internet of Medical Things Platform," in IEEE Access, vol. 8, pp. 59247-59256, 2020, doi: 10.1109/ACCESS.2020.2981159.

[12] Hager Ahmed, Eman M.G. Younis, Abdeltawab Hendawi, Abdelmgeid A. Ali, Heart disease identification from patients' social posts, machine learning solution on Spark, Future Generation Computer Systems, Volume 111, 2020, Pages 714-722, ISSN 0167-739X, https://doi.org/10.1016/j.future.2019.09.056.

[13] Katarya, R., Meena, S.K. Machine Learning Techniques for Heart Disease Prediction: A Comparative Study and Analysis. Health Technol. 11, 87–97 (2021). https://doi.org/10.1007/s12553-020-00505-7

[14] Kannan, R., Vasanthi, V. (2019). Machine Learning Algorithms with ROC Curve for Predicting and Diagnosing the heart disease. In: Soft Computing and Medical Bioinformatics. SpringerBriefs in Applied Sciences and Technology (). Springer, Singapore. https://doi.org/10.1007/978-981-13-0059-2_8