MACHINE LEARNING TECHNIQUES IN HEART DISEASE FORECASTING

Shah Samkitkumar Rajnikant[1] , Dr. Harsh Lohiya[2]

[1]Research Scholar, Dept. of Computer Science & Engineering,
Sri SatyaSai University of Technology and Medical Sciences,
Sehore, Bhopal-Indore Road, Madhya Pradesh, India

[2]Research Guide, Dept. of Computer Science & Engineering,
Sri SatyaSai University of Technology and Medical Sciences,
Sehore, Bhopal-Indore Road, Madhya Pradesh, India

ABSTRACT

Cardiovascular disease (sometimes known simply as heart disease) (CVD),has become the leading world of worldwide in recent decades. This is true not just in India, but everywhere else as well, including the United States. For this reason, a fast, accurate, and hands-on method of diagnosing such diseases is required. Several medical datasets have benefited from the automated inspection of huge and complex data by means of Machine Learning algorithms &methods. Many researchers have recently turned to the use of various machine learning approaches to aid in the early diagnosis of cardiac problems. The human heart is the second most essential organ after the brain. It pumps blood through the body's veins and arteries to keep everything running smoothly. The medical field's work to predict the prediction of heart diseases is substantial. The medical centre benefits from data analytics since it allows for more accurate disease forecasting based on a larger data set. Every month, a mountain of information about data must be updated and stored. The collected information can be used as a benchmark for forecasting the spread of future diseases. Several data mining and machine learning approaches, including Artificial Neural Networks (ANNs), Random Forests, and Support Vector Machines, have been successfully applied to the problem of cardiovascular disease prediction (SVM). Doctors and hospitals in India and throughout the world are facing an increasing difficulty with the prediction and diagnosing of heart disease. Finding an effective and rapid diagnosis method for cardiovascular diseases is crucial for lowering the number of deaths caused by them. Use of data mining methods and machine learning algorithms is critically important here. In order to aid doctors in the prediction and diagnosing of heart disease, researchers are ramping up their efforts to create software using machine learning algorithms. Using machine learning algorithms, the primary goal of this research is to forecast the likelihood that a certain patient would develop cardiac disease.

KEYWORDS: Neural Network, Machine Learning, Supervised learning, Support vector machine, Random forest.

DOI Number: 10.48047/nq.2022.20.9.NQ44843 **NeuroQuantology2022;20(9):7246-7256**

INTRODUCTION

In the last decade, cardiovascular disease has surpassed all other causes of death worldwide. According to the World Health Organization, coronary artery disease and cerebrovascular accident account for almost 80% of all fatalities from cardiovascular disease. Heart disease is caused by both genetic predisposition and

environmental variables in one's personal and professional life. Heart mass index (BMI), blood pressure (BP), cholesterol variables, and the presence of any preexisting heart diseases are only few of the many physiological factors that might play a role in the development of cardiovascular disease. Tobacco use, excessive coffee and alcohol intake, stress, and inactivity are further contributors to the risk. An accurate and early diagnosis of heart disease can be very helpful to the medical field, as it allows for the taking of preventative measures before problems arise. The heart is a very important organ. One of the heart's main functions is to pump blood throughout the body. The brain and other organs will shut down and the person will die within minutes if it stops working properly. Lifestyle factors, occupational stress, and poor food choices all contribute to the rising prevalence of cardiovascular disease.

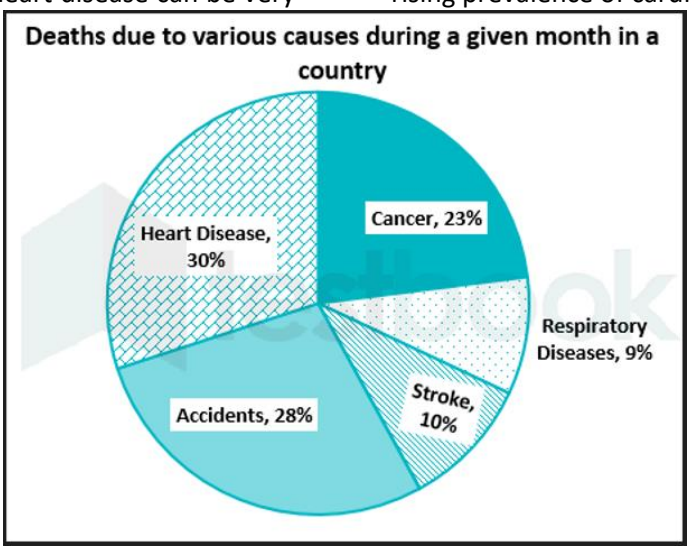

Figure 1.1 Cause of Death in Developing Countries

In recent decades, heart disease has risen to prominence as one of the leading diseases worldwide. The World Health Organization (WHO) reports that heart-related illnesses kill 17.7 million people per year. Indeed, heart disease is now the leading cause of mortality worldwide, including in India. The 2016 Global Burden of Disease Report was released on September 15, 2017. It found that heart disease was responsible for the deaths of 1.7 million Indians that year. Health diseases and lost productivity are both higher when someone has a heart condition. The World Health Organization (WHO) has estimated that between 2005 and 2015, cardiovascular diseases cost India $237 billion. Therefore, it is crucial to have a prediction for reliably predicting cardiovascular diseases. Throughout the world, medical organizations collect information on a wide variety of health-related topics. To derive value from these data, many machine learning techniques can be applied. The acquired data, however, is typically quite noisy and involves a huge volume. Machine learning techniques make quick work of exploring these datasets, which are too large for human minds to handle. Thus, these algorithms have recently shown to be an effective tool for accurately forecasting the presence or absence of heart-related disorders. Technology is being increasingly used in the medical health to help doctors make better decisions. It aids medical doctors in the disease of diseases, the identification of appropriate drugs, and the identification of trends and correlations in patient diagnosis data. Existing procedures miss a significant number of people who would benefit from preventative care for cardiovascular risk, while others receive unnecessary intervention. Machine learning presents a chance to boost precision by taking advantage of complicated connections between risk factors. To determine if cardiovascular risk

prediction can be enhanced by machine learning, we analyzed a dataset.

Machine learning is a technique for automatically analyzing data and extracting patterns and insights that humans would otherwise have to manually identify and infer. Machine learning encompasses an enormous range of topics, and its relevance and use are expanding rapidly. Machine learning includes different classifiers from Supervised Learning, Unsupervised Learning, and Ensemble Learning to make predictions and determine how accurate a dataset is. That information will be invaluable for the HDPS project we're working on. Many different disorders can affect the heart, and the term "cardiovascular disease" describes all of them. An estimated 17.9 million people worldwide lost their deaths to CVDs in 2016, according to the World Health Organization. Among adults, this is the leading cause of death. By analysing a person's past medical records, our project can help determine who is most likely to be diagnosed with heart disease. It identifies people who are experiencing cardiac symptoms like chest pain or high blood pressure and aids in the diagnosis of heart disease with fewer medical tests and more effective therapies. The primary data mining techniques utilized in this project are the Logistic Regression, K-Nearest Neighbor, and Random Forest Classifier. When compared to the prior system, in which only one data mining technique was used, our project's accuracy of 87.5% is a significant improvement. Therefore, the HDPS's accuracy and productivity were enhanced by employing additional data mining techniques. Supervised learning, which includes logistic regression, is one type of machine learning. In logistic regression, only values that fall inside a discrete range are considered. The purpose of this project is to determine if the patient is at risk for developing cardiovascular heart disease based on demographic and medical diseases including gender, age, chest discomfort, fasting glucose level, and other factors. Patients' medical records and associated characteristics are extracted from a dataset stored at UCI. Using this data collection, we are able to make an educated guess as to whether or not a given patient is suffering from heart disease. We use a patient's 14 medical characteristics to predict his or her risk of developing cardiovascular disease and place them into one of four categories. Logistic regression, K-Nearest Neighbor, and Random Forest Classifier are the three algorithms used to hone these healthcare variables. KNN provides the highest accuracy (88.52%) of any of these algorithms. Finally, we use this method to categorize patients according to their risk for developing heart disease; furthermore, it is extremely cost-effective.

LITERATURE REVIEW

Pooja Anbuselvan et al (2020),Numerous illnesses can affect the heart, and they are collectively referred to as heart diseases or cardiovascular diseases. Diseases of the blood vessels, irregular heartbeats, and congenital cardiac abnormalities all fall within this category. As the decades have passed, it has become the leading cause of death around the globe. In order to effectively manage a disease, it is vital to find a precise and trustworthy method for identifying it at an early stage. Processing massive amounts of data is where data science really comes into its own in the medical sciences. In order to accurately forecast cardiac issues, researchers employ a variety of Data Mining and Machine Learning Techniques to examine massive data sets. In order to determine which supervised learning model is the most effective, this work compares and contrasts Logistic Regression, Naive Bayes, Support Vector Machine, K-Nearest Neighbors, Decision Tree, Random Forest, and the ensemble technique XGBoost. When compared to other algorithms, Random Forest is shown to have the highest accuracy (86.89%).

Harshit Jindal et al (2021),The incidence of heart disease seems to be rising at an alarming rate, making it crucial and worrying to try to foresee its onset. It is important to be thorough and accurate while making this diagnosis because it is a challenging process. The primary goal of this study is to determine which patients are at higher risk for developing heart disease

based on a number of clinical characteristics. Based on the patient's medical records, we developed a technique to determine the likelihood that heart disease will be discovered in the patient. We predicted and classified the patient with heart disease using a number of machine learning algorithms, including logistic regression and KNN. The model's potential to improve the accuracy of heart attack prediction in every given individual is governed by a very helpful method. In comparison to other classifiers used in the past, such as naive bayes, the suggested model's combination of KNN and Logistic Regression to predict whether or not an individual had heart disease was quite impressive. Consequently, a sizable burden has been released thanks to the proposed model, which may be used to calculate the classifier's likelihood of making a correct diagnosis of heart disease. Improved health care quality and lower overall costs are also benefits of the Given heart disease prediction system. The data gathered from this study will be very useful in identifying future heart disease patients. It operates on files saved in the.pynb format.

RESEARCH METHODOLOGY& DESIGN

It has been motivated by the relatively substantial quantity of work connected to the use of Machine Learning algorithms in the detection of cardiovascular heart disease. We present a brief survey of the relevant literature here. Predictions of cardiovascular disease have been made using a variety of methods, including Logistic Regression, KNN, Random Forest Classifier, etc. The outcomes demonstrate that each algorithm has its own distinctive potential for detecting the designated aims. The model with IHDPS integrated was able to compute the decision boundary using both the conventional and the modern machine learning and deep learning model. Essential data, such as a family history of heart disease, was made available. Even if their accuracy is far lower than the present IHDPS model, new models are on the horizon, such as one that may diagnose coronary heart disease using an artificial neural network and other machine learning and deep learning methods. A deep Neural Network was built that took into account the provided attributes related to the disease and was able to produce an output that was carried out by the output perception and almost included 120 hidden layers to guarantee an accurate result of having heart disease if we use the model for Test Dataset. Heart disease diagnosis is one of the areas where the monitored network has been recommended. Researchers were able to gauge the model's accuracy by having doctors do tests on data the model had never seen before; the model, having been trained using the previous learning data, correctly predicted the outcome. This research compares many machine learning algorithms that could help doctors and data scientists make more accurate diagnoses of cardiac illness. These algorithms include K nearest neighbours (KNN), logistic regression, and random forest classifiers. Reading up on the most recent papers, publications, and statistics pertaining to cardiovascular illness is a key part of this paper job. The approach describes how the proposed model is organised.

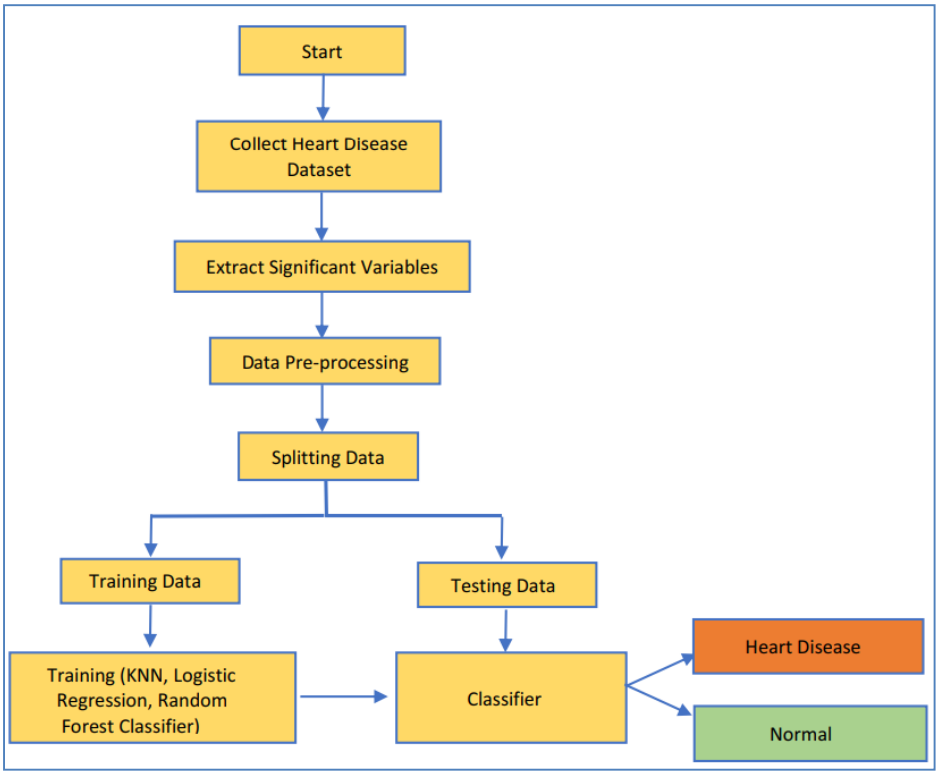

Figure 1.3 Proposed Model

A methodology can be thought of as a procedure with phases that, when followed, transform raw data into easily identifiable data patterns for the benefit of the consumers. Figure 1.3, which lays out the recommended method, shows how the first step is data collection, the second is extracting key values, and the third is the preparation stage, in which we study the data. Depending on the algorithms, data preprocessing may also involve dealing with missing values, cleaning the data, and normalizing the results. The proposed model uses KNN, Logistic Regression, and Random Forest Classifier to classify data after it has been preprocessed. Finally, we implemented the proposed model and used a number of performance indicators to assess our model's accuracy and efficiency. Using this model and a variety of classifiers, a reliable Heart Disease Prediction System (EHDPS) was created. Predictions are made using 13 different medical factors, including chest discomfort, fasting glucose, blood pressure, cholesterol, age, sex, etc.

CLASSIFICATION ALGORITHMS

Predicting future data based on currently available information is the goal of the supervised learning method known as classification. In this paper, we offer a diagnosis that uses classification algorithms to identify heart disease. Individual classifiers are trained on the training dataset before being sent to the test dataset. The dataset was split 80/20. This dataset is used to evaluate the performance of the classifiers. In the following part, I will describe how each of these classifiers works.

1. **Logistic Regression** - Predicting the likelihood of a target variable can be done with logistic regression, which is a supervised learning sorting process. The dependent or goal variable can only take on one of two values: either 0 for failure or 1 for success because of its binary character.
2. **Naive Bayes** - A supervised algorithm such as a Naive Bayes classifier is one option. It's a simple method for categorizing data that uses Bayes' theorem. In this model, qualities are assumed to be unrelated to one

another. The probability can be calculated with the help of Bayes' theorem, which is a mathematical notion. There is no interconnected structure or correlated structure among the predictors. All of its features contribute equally to increasing the likelihood of success. Classifiers based on Naive Bayes theory are commonly employed in many realistic but challenging situations.

$P(X/Y) = P(Y/X) \times P(X)/P(Y)$,

The likelihood, or probability of predictor X given a set of Y's, is denoted by P(Y/X), while the prior probability of class X is P(X) and the prior probability of predictor Y is P(Y).

3. **Support Vector Machine** - In a support vector machine (SVM) model, the various categories are projected onto a multidimensional hyperplane. To reduce the error as much as possible, SVM will generate the hyperplane through a series of iterations. To achieve this purpose, SVM first classifies datasets so that a maximum marginal hyperplane can be located between them (MMH).
4. **K-Nearest Neighbour**- K-Nearest Neighbor is a supervised classification algorithm. Next-door neighbour information is used to determine how an object should be categorized. It's a form of "learning by example." Euclidean distance is used to determine how far apart two attributes are in a set. As a guide, it takes a set of labelled locations and utilizes them to determine where to place yet another label. Information is organized into groups defined by their shared data. The K-Nearest Neighbors algorithm requires no model building or other presumptions to run. You can use this adaptable method for search, regression, or classification. The inclusion of noise and irrelevant variables reduces the effectiveness of even the simplest approach, K-NN.
5. **Random Forest**- The random forest algorithm is a classification for making accurate classifications under supervision. Several trees work together to form a forest in this algorithm. Each individual tree in random forest lets out a class expectation and the class with most votes turns into a model's forecast. The accuracy of the random forest classifier increases as the number of trees grows. Classification is where it really shines, and it also excels at handling missing values, so it's not only a regression tool. Predictions take longer to obtain since more data and more trees are needed, and the findings are not verifiable.
6. **XGBoost**- XGBoost is a distributed gradient model that has been optimized to be fast, adaptable, and easy to transport. To put it simply, it is an ensemble Machine Learning algorithm that makes use of the gradient boosting framework and uses on decision trees. By utilizing regularization, missing-data handling, and parallel processing, it prevents overfitting and biases in the gradient boosting algorithm.

DATA SOURCE

A Dataset of People were Chosen with Medical Conditions and Heart Conditions in Mind. Several different medical problems can have an impact on the heart. World Health Organization (WHO) reports that cardiovascular diseases account for the majority of deaths among people in their middle years. We use a database that includes the health records of 304 people of varying ages.

S.NO	COMMENT	REPORT	STANDARDS
1	Age	Age in Years	Continuous
2	Sex	Sex of Subject	Male/Female
3	CP	Chest Pain	Four Types
4	Trestbps	Resting Blood Pressure	Continuous
5	Chol	Serum Cholesterol	Continuous
6	FBS	Fasting Blood Sugar	<, Or >120 mg/dl
7	Restecg	Resting Electrocardiograph	Five Values
8	Thalach	Maximum Heart Rate Achieved	Continuous
9	Exang	Exercise Induced Angina	Yes/No
10	Thal	Defect Type	Reversible/Normal/Fixed
11	Num(Disorder)	Heart Disease	Not Present/Present in the four major types

Using this dataset, we may ascertain whether or not a patient has been diagnosed with heart disease based on a variety of factors, including the patient's medical features such as age, resting blood pressure, fasting sugar level, etc. With the help of this dataset, which contains 13 different medical variables from 304 patients, we were able to identify and categories those who were at risk for developing heart disease. This Heart Disease dataset was obtained from the UCI repository. Patients at risk of developing heart disease are identified using the dataset to assess patient patterns. Files are separated into two sections: evaluation and practice. One record is represented by one row in this dataset. There are 14 columns and 303 rows in all. A comprehensive list of characteristics is provided in "Table 1".

RESULT AND DISCUSSION

The purpose of this research is to determine which classification algorithm is the most accurate in determining whether a specific patient will develop heart disease. The UCI dataset was used to test the performance of Logistic Regression, Naive Bayes, Support Vector Machine, K-Nearest Neighbor, Decision Tree, Random Forest, and XGBoost. Datasets were split into training and testing sets, models were trained, and accuracy was measured all in Python.

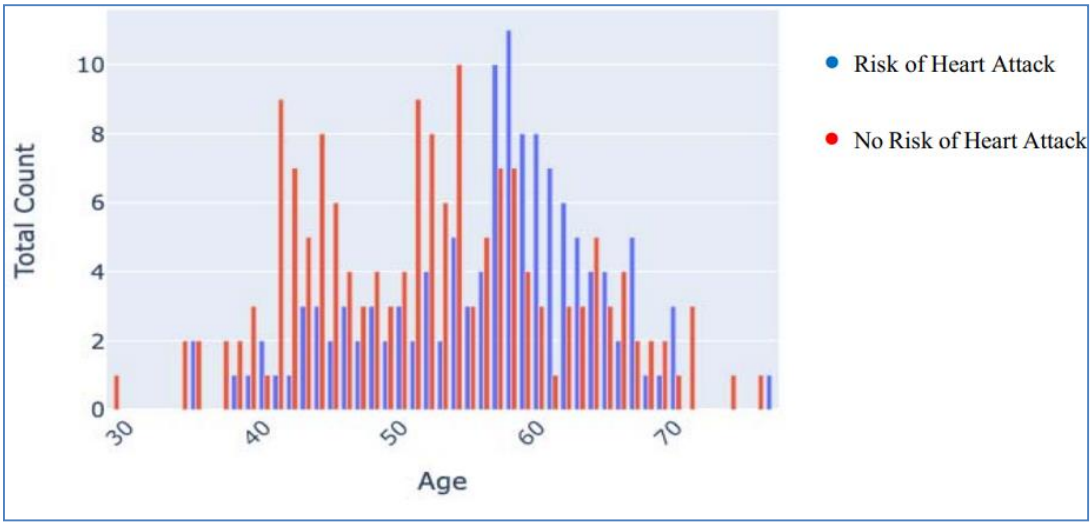

Figure 1.4The Heart Attack Risk Relative to Age.

These findings show that despite the fact that most studies employ diverse algorithms like SVC and Decision tree to identify patients with heart disease, KNN, Random Forest Classifier, and Logistic regression produce superior outcomes.

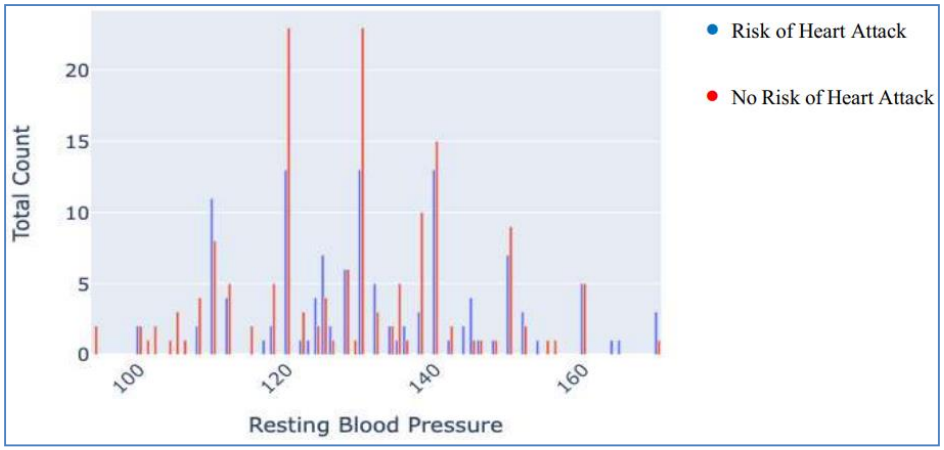

Figure 1.5Predicts the Chance of a Heart Attack Based on Resting BP.

Our algorithms are faster, more efficient with regards to both time and money, and more accurate than those utilized by earlier studies. As an added bonus, the highest levels of accuracy achieved by KNN and Logistic Regression are both equivalent to 88.5%, which is higher than or on par with the accuracy levels achieved by prior studies.

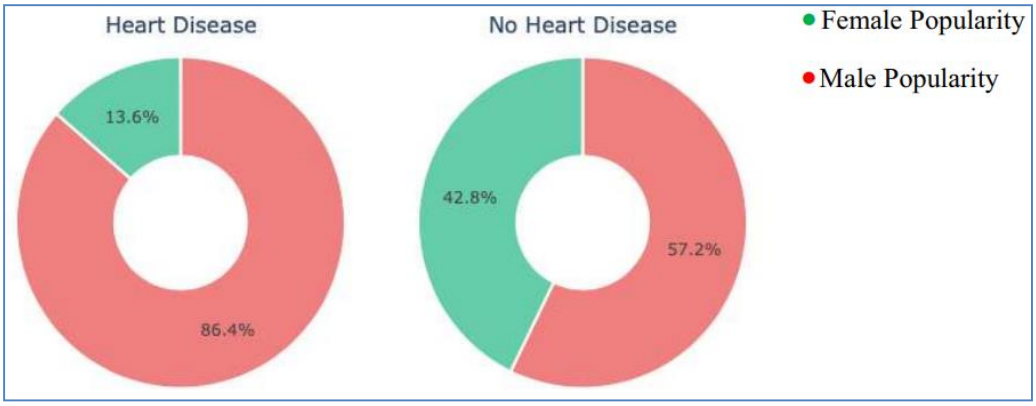

Figure 1.6Shows patients who have or do not have heart disease according on their gender.

As a result, we conclude that the larger number of medical attributes we incorporated into our dataset contributed to a significant boost in our prediction accuracy. Furthermore, the results of our project show that the prediction of a patient with heart disease using Logistic Regression or KNN is superior than that using Random Forest Classifier.

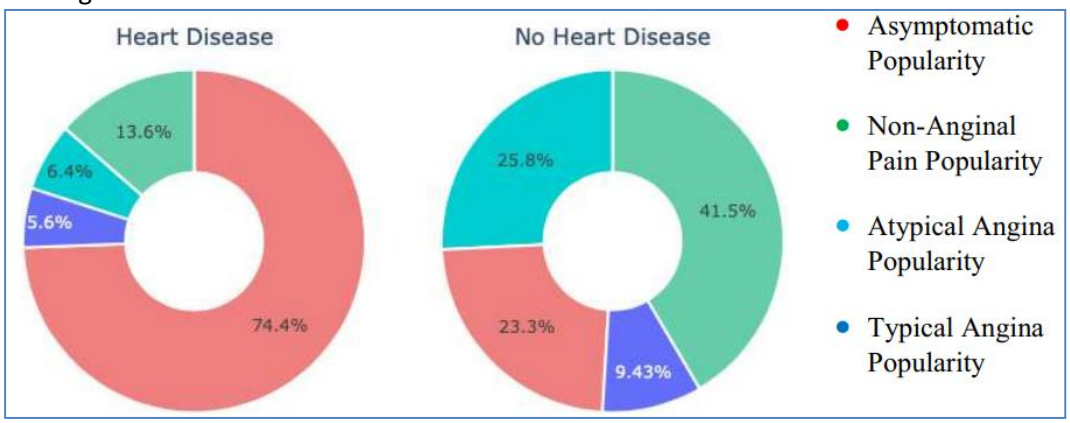

Figure 1.7Displays patients who have or do not have heart disease based on the type of Chest Pain.

KNN and Logistic Regression have been shown to be superior in the diagnosis of cardiac conditions. Figures 2, 3, 4, and 5 shows a scatter plot of the number of patients classified by age group, as predicted by the classifier. Issues with Sexual Function, Sexual Pain, and Resting Blood Pressure.

CONCLUSION

Using three distinct machine learning (ML) classification techniques, a model for identifying cardiovascular disease has been created. A dataset of patients' medical records is mined for the patient medical history that predicts lethal heart disease in order to produce predictions about those patients' future health (including symptoms like chest discomfort, blood sugar levels, blood pressure, and so on). Helpful if the patient has a history of heart disease, this approach can identify it from the patient's clinical data. Logistic regression, Random Forest Classifier, and the K-Nearest Neighbor algorithm were used to construct the existing model. The model we've developed has an accuracy of 87.5%. More training data improves the model's likelihood of making a correct diagnosis of heart disease in a given individual. These computer-aided methods allow for faster, more accurate patient predictions at significantly lower costs. Since these Machine Learning methods outperform humans and can make more accurate predictions, they may be used to a wide variety of medical databases, which benefits both patients and doctors. We were

able to improve upon earlier models' 85% accuracy by cleaning the dataset and applying logistic regression and KNN, allowing us to make more informed predictions about the patients who will ultimately be diagnosed with heart diseases thanks to this project. Furthermore, we find that KNN's accuracy of 88.52% is the greatest of the three algorithms we tested. Figure 1.6 illustrates that 44% of those included in the dataset have some form of heart disease..

REFERENCES

[1] Soni J, Ansari U, Sharma D &Soni S (2011). Predictive data mining for medical diagnosis an overview of heart disease prediction.International Journal of Computer Applications.

[2] Dangare C S &Apte S S (2012). Improved study of heart disease prediction system using data mining classification techniques.International Journal of Computer Applications.

[3] Ordonez C (2006). Association rule discovery with the train and test approach for heart disease prediction.IEEE Transactions on Information Technology in Biomedicine.

[4] Shinde R, Arjun S, Patil P &Waghmare J (2015). An intelligent heart disease prediction system using k-means clustering and Naïve Bayes algorithm.International Journal of Computer Science and Information Technologies.

[5] Bashir S, Qamar U &Javed M Y (2014, November). An ensemble-based decision support framework for intelligent heart disease diagnosis.In International Conference on Information Society.

[6] Jee S H, Jang Y, Oh D J, Oh B H, Lee S H, Park S W & Yun Y D (2014). A coronary heart disease prediction model: the Korean Heart Study. BMJ open, 4(5), e005025.

[7] Ganna A, Magnusson P K, Pedersen N L, de Faire U, Reilly M, Ärnlöv J &Ingelsson E (2013). Multilocus genetic risk scores for coronary heart disease prediction. Arteriosclerosis, thrombosis, and vascular biology, 33(9), 2267-72.

[8] Jabbar M A, Deekshatulu B L & Chandra P (2013, March). Heart disease prediction using lazy associative classification.In 2013 International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s) (pp. 40- 6).

[9] DangareChaitrali S and Sulabha S Apte. "Improved study of heart disease prediction system using data mining classification techniques." International Journal of Computer Applications 47.10 (2012): 44-8.

[10] SoniJyoti. "Predictive data mining for medical diagnosis: An overview of heart disease prediction." International Journal of Computer Applications 17.8 (2011): 43-8.

[11] Chen A H, Huang S Y, Hong P S, Cheng C H & Lin E J (2011, September). HDPS: Heart disease prediction system. In 2011 Computing in Cardiology (pp. 557-60).

[12] Parthiban, Latha and R Subramanian. "Intelligent heart disease prediction system using CANFIS and genetic algorithm."International Journal of Biological, Biomedical and Medical Sciences 3.3 (2008).

[13] Wolgast G, Ehrenborg C, Israelsson A, Helander J, Johansson E &Manefjord H(2016). Wireless body area network for heart attack detection. IEEE antennas and propagation magazine, 58(5), 84-92.

[14] Seckeler MD, Hoke TR. The worldwide epidemiology of acute rheumatic fever and rheumatic heart disease.ClinEpidemiol.

[15] Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine learning improve cardiovascular risk prediction using routine clinical data? PLoS ONE. 2017;12(4):e0174944.

[16] Ramalingam VV, Dandapath A, Raja MK. Heart disease prediction using machine learning techniques: a survey. Int J Eng Technol. 2018;7(2.8):684–7.

[17] Pouriyeh S, Vahid S, Sannino G, De Pietro G, Arabnia H, Gutierrez J. A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease. In: 2017 IEEE symposium on computers and communications (ISCC). IEEE. p. 204–207.

[18] Bouali H, Akaichi J. Comparative study of different classification techniques: heart disease

use case. In: 2014 13th international conference on machine learning and applications. IEEE.p. 482–86.

[19] Chauhan R, Bajaj P, Choudhary K, Gigras Y. Framework to predict health diseases using attribute selection mechanism. In: 2015 2nd international conference on computing for sustainable global development (INDIACom). IEEE. p. 1880–84.

[20] T.Nagamani, S.Logeswari, B.Gomathy, Heart Disease Prediction using Data Mining with Mapreduce Algorithm, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-3, January 2019.