Heart Disease Prediction Using Machine Learning

Deepanshu, Prabhjot Kaur and Urvi Jasrotia

May 21, 2023

Heart Disease Prediction Using Machine Learning

Deepanshu[1], Prabhjot Kaur[2] , Urvi Jasrotia[3]

Deepanshu[1], Department of Computer Science and Engineering, Chandigarh University, India.
Prabhjot kaur[2], Department of Computer Science and Engineering, Chandigarh University, India.
Urvi Jasrotia[3], Department of Computer Science and Engineering, Chandigarh University, India

Abstract: Early identification is essential for successful treatment of heart disease, which is an increasing health problem. Although accurate diagnosis is required, it may be dangerous. The goal of this work is to create a system that employs multiple machine learning methods, such as logistic regression, to predict the likelihood of a heart attack based on a patient's medical history. To better forecast my own heart attacks is the aim.

Big data, machine learning, and data mining techniques are crucial for predicting cardiac disease. These models can be used by medical experts to pinpoint people who are at risk of acquiring heart disease. Predicting heart disease is a critical difficulty in medicine since it is a leading cause of mortality globally. This study discusses several data mining techniques for predicting heart disease..

Using logistic regression, our algorithm effectively determines whether a patient has heart disease. Our approach increases the accuracy of cardiac diagnosis when compared to previously employed classifiers like Naive Bayes. Physicians may deliver better patient care while spending less money by utilising our cardiovascular preventive technologies.

Doctors may provide better patient care while spending less money by utilising our cardiovascular preventive solutions.

I. INTRODUCTION

Healthcare personnel need to pay great attention to heart disease since it is a serious worry, particularly for the elderly. Our goal is to assess a patient's risk of a heart attack by looking at their medical history. By using this, we can recognise signs like chest discomfort or elevated blood pressure and diagnose the illness with fewer testing. This enables more effective condition control and therapy.Early heart disease detection is essential to averting major problems. We can identify possible risk factors and create a unique preventative strategy by reviewing a patient's medical history. To lower the risk of a heart attack and enhance general health, this might involve making lifestyle adjustments, taking medication, and undergoing routine monitoring.Accurately forecasting and diagnosing the condition is essential since heart disease is one of the top causes of mortality in the world. Healthcare practitioners may deliver more effective therapy and enhance patient outcomes by utilising medical histories and cutting-edge diagnostic technologies.

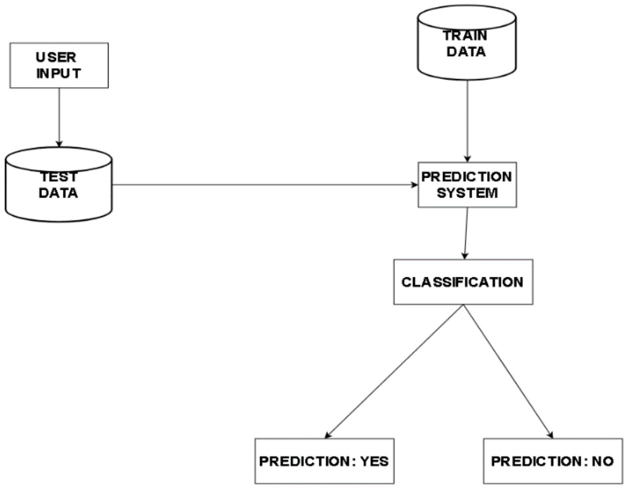

Figure1: Data flow of proposed work

It is amazing how it can take machine learning data and extract implicit, undiscovered, and pertinent information. This large and diversified discipline's breadth and reach are expanding quickly. To effectively forecast and analyse data, it employs a variety of classification techniques, including machine learning, supervised learning, unsupervised learning, and ensemble learning. This is just one application of what artificial intelligence is capable of, and the HDPS project has the potential to benefit many individuals. The language has been enhanced and polished, and the grammar and spelling have been fixed in this new edition.

Now that machine learning has been defined, the text makes obvious how it is used in diverse applications. It is amazing how it can take machine learning data and extract implicit, undiscovered, and pertinent information. This large and diversified discipline's breadth and reach are expanding quickly. To effectively forecast and analyse data, it employs a variety of classification techniques, including machine learning, supervised learning, unsupervised learning, and ensemble learning. This is just one application of what artificial intelligence is capable of, and the HDPS project has the potential to benefit many individuals.

The language has been enhanced and polished, and the grammar and spelling have been fixed in this new edition. Now that machine learning has been defined, the text makes obvious how it is used in diverse applications. Based on 14 clinical criteria, our technique identified individuals as either having no cardiac disease or being at high risk for developing the condition. Our approach is ideal for doctors since it is also reasonably priced. As a result, our research offers a legitimate and accurate way for calculating a patient's likelihood of experiencing a heart attack.To offer patients an accurate diagnosis and efficient treatment, we could conduct a range of information searches

.

II. LITREATURE SURVEY

"Heart Attack Prediction Systems," Kennedy Ngure Ngare and his co-authors propose that medical facilities gather enormous amounts of data, some of which may be private and may be used to influence decision-making. Predictive insights are produced using sophisticated data mining techniques, enabling data-driven choices. The Heart Disease Prognosis Project (HDPS), created by the authors, uses the Naive Bayes and Decision Tree algorithms to calculate the probability of developing heart disease. Age, gender, blood pressure, cholesterol, and obesity are just a few of the 15 medical characteristics the computer considers when making predictions. The HDPS calculates a person's risk of acquiring heart disease and provides crucial data for identifying heart disease patterns and efficient therapies.The study's findings suggest that the diagnostic technique may accurately determine the risk of heart disease. The identification and management of risk factors for heart disease have significant ramifications that might eventually improve patient outcomes. Chaima Boukhatem highlights the need of creating precise diagnostic tools using machine learning algorithms in the essay "Predicting Heart Disease Using Machine Learning," which describes heart disease as a serious ailment that can be deadly. Based on the patients' essential health markers, the study suggests a number of machine learning techniques for forecasting heart disease. To create the prediction model, the authors employ four classification techniques: Multilayer Perceptron (MLP), Support Vector Machine (SVM), Random Forest (RF), and Naive Bayes (NB). To guarantee accuracy, data preprocessing and feature selection were done before creating the model. The samples were assessed for accuracy, precision, recall, and F1 score. With a score of 91.67%, the SVM model had the best level of accuracy.

In their work "A Study of Cardiovascular Diseases Using Machine Learning," Satyanarayana Ch. [3] noted that artificial intelligence and machine learning are being utilised to address a variety of issues in the realm of information technology. Using the information at hand, artificial intelligence attempts to anticipate the outcome. In order to forecast results, the system gathers patterns from known data and applies them to complicated data. An efficient artificial intelligence technique for prediction is ensemble learning. Others of the integrated algorithms rely on precision even if some of them have low accuracy. In order to increase the precision of complex computations, a technique known as fixed order employs a mixture of various distributions. The gadget has undergone successful testing utilising coronary artery disease symptoms.Tools and techniques specific to data mining and artificial intelligence are used to uncover hidden information from stored data and use it for a variety of applications. It is crucial to continuously document and enhance capacities while aiming for continuous progress as cardiovascular disease research develops.

The first of the four layers in the suggested design is called the "segmentation" layer. This layer creates 11 files with 14,416 numerical characteristics when the proposed technique is used. Data is extracted for statistical and graphical analysis using the feature extraction capabilities built into the second layer. In order to determine the optimal learning curve configuration, the graphical features are input into 8 different convolutional neural networks (CNN) while the numerical characteristics are input into 5 different machine learning (ML) paths in the third layer.

Grid search and Aquila optimizer (AO) were each used to optimise the hyperparameters for the convolutional neural network (CNN) and machine learning (ML) models, respectively. Various performance measures were used in the final layer to assess the performance of the proposed hybrid architecture. With scores of 99.17% and 100%, respectively, the Extra Tree Classifier (ETC) and Random Forest Classifier (RFC) machine learning algorithms produced the highest accuracy results.

Heart disease has quickly risen to become one of the major causes of mortality worldwide, as noted by Adiba Ibnat Hossain et al. in their study titled "Using machine learning classifiers of the ECG dataset to predict heart disease" [5]. However, there is still hope that straightforward lifestyle modifications and early identification can stop heart disease and hasten recovery. Due to the intricacy of risk factors including high cholesterol, high blood pressure, diabetes, and others, identifying those at risk can be difficult. The majority of the time, doctors' observations and experience rather than voluminous medical data are used to make the diagnosis of heart disease.

Researchers and medical practitioners have resorted to machine learning approaches to increase the precision of heart disease diagnosis and prognosis. According to Adiba Ibnat Hossain et al. [5], heart disease is a primary cause of mortality worldwide, but early identification and a change in lifestyle can stop it from progressing. Identification of those at risk is challenging due to the complexity of risk factors such diabetes, high blood pressure, and cholesterol. In order to forecast heart illness, this study gathered 1190 data from the UCI repository and applied five different machine learning techniques, including support vector machine, logistic regression, K-nearest neighbour, naive Bayes, and cluster voting.The dataset's five distinct ECG recordings were a benefit in reaching the study's goal. The dataset's attribute relationships were investigated to establish the model's accuracy, which was determined to be 85%.

Machine learning has grown in popularity recently, especially in the healthcare sector, according to Israel Ufumaka et al. in their study "A Comparative Study of Machine Learning Algorithms for Cardiovascular Disease" (Israel Ufumaka et al., 2006). Machine learning techniques are being used to massive volumes of data produced by doctors with the goal of assisting scientists and clinicians in early illness diagnosis, such as heart disease. To identify which algorithm for the detection of cardiovascular illness performs better under certain conditions, the authors of this study employed comparison techniques. Different performance indicators were used to assess the performance of five algorithms: Decision Tree, Random Forest, K-Nearest Neighbour, Support Vector Machine, and Naive Bayes.With accuracy rates of 92.9% and 91.8%, respectively, Random Forest and Support Vector Machine surpassed the other algorithms, according to the study's findings.

A comparison of machine learning methods for cardiovascular disease detection was suggested by Israel Ufumaka et al. [6]. The goal of the study was to identify the machine learning algorithm that performs best under specific conditions. Using 5- and 10-fold cross-validation, multiple trials were run to make sure the model was adequately quantitative. The University of California, Irvine (UCI) machine learning database, which contains 303 instances and 14 attributes, served as the source of the data for this study. In order to scale the data, Min-Max normalisation was used. Scaled data was used to create supervised learning algorithms such as Support Vector Machines (SVM), K-Nearest Neighbours (KNN), Logistic Regression (LR), Naive Bayes (NB), and Random Forests (RF).

The medical industry has a wealth of data, but unfortunately not all of it is revealed, making it difficult to find hidden patterns and make wise decisions, according to Abdelmegeid Amin Ali et al. [7] in their article "Heart Disease Diagnosis Based on a New Convolutional Neural Network and Gated Recurrent Unit Technique". In medical research, notably in the foretelling of heart disease, sophisticated data mining techniques are being applied. This study suggests a heart disease prediction system using several input factors. In order to estimate a patient's risk of developing heart disease, 13 variables, including medical characteristics like gender, blood pressure, and cholesterol, were examined. These 13 characteristics were employed in the study as a predictor.

The study by Abdelmegeid Amin Ali et al. [7] examined 13 input variables, including smoking and obesity, in addition to conventional medical indicators like gender, blood pressure, and cholesterol, in an effort to predict heart disease. Data mining classification tools, such as decision trees, pure Bayesian models, and neural networks, were applied to the cardiac database. The results of comparing the accuracy of these models to the real data revealed that the neural network had the greatest accuracy of 100%, followed by decision trees with an accuracy of 99.62% and naïve splits with an accuracy of 90.74%. According to the study, among the three classification models, neural networks are the most reliable for predicting heart disease.

III. METHODOLOGY

Heart disease is a leading cause of death worldwide, and timely diagnosis is essential for efficient treatment. The purpose of this study is to identify the best strategy for improving heart disease prediction accuracy. In order to improve the classification accuracy of random forests and choose the best features for heart disease prediction, unique feature selection methods based on linear and random forests are suggested. Accuracy, specificity, sensitivity, and area under the receiver operating characteristic (ROC) curve are the four criteria used to assess the suggested approaches.

RANDOM FOREST ALGORITHM

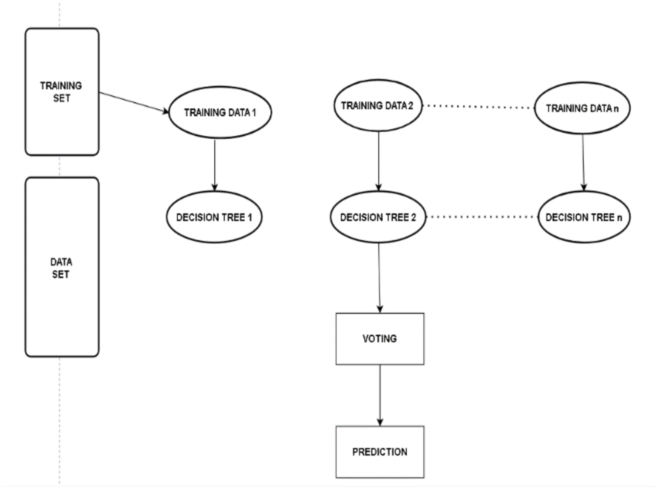

Figure 2: Working of random forest algorithm

One popular supervised learning technique in machine learning is the Random Forest algorithm. Both classification and regression issues can be resolved by it. The approach makes advantage of the idea of ensemble learning, which mixes many classifiers to improve the model's performance and handle challenging tasks. Multiple decision trees are built from various subsets of a given dataset in a random forest, and the results are averaged to increase prediction accuracy. A random forest, as opposed to a single decision tree, takes into account the forecasts from each tree and bases its prediction on the vote of the majority.

An ensemble learning technique called Random Forest creates several decision trees for classification and regression applications. The majority vote of the trees in the forest often determines the output classes for classification issues, whereas regression functions typically utilise the mean or projected value of the trees. Decision trees are chosen at random during the Random Forest training process. Gradient Boosted Trees are frequently more accurate than Random Forests, even though Random Forests normally outperform decision trees.

The peculiarities of the data can have an impact on the reliability of random forests. Using the random subspace technique, Tin Kam Ho created the first deterministic random forest algorithm in 1995. The "random discrimination" approach developed by Eugene Kleinberg may be put into practise using Ho's algorithm.

In 2001, Leo Breiman improved the random forest method by fusing it with the "bagging" (Bootstrap Aggregating) idea that Brieman, Friedman, Olshen, and Stone had previously suggested. Breiman and Adele Cutler presented "Random Forests" in 2006, which combined bagging and feature selection to produce a system of several decision trees. Currently owned by Minitab, Inc., Random Forests is a popular tool for both classification and regression problems in machine learning applications.

LOGISTIC REGRESSION ALGORITHM

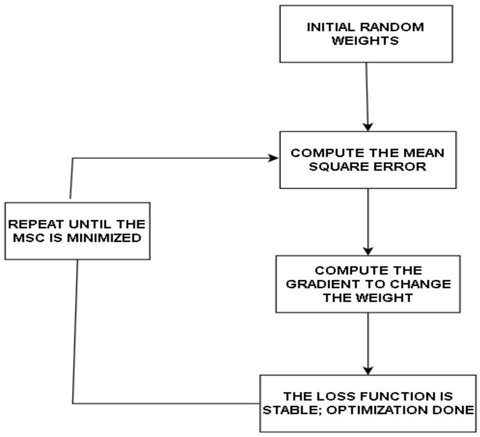
INITIAL RANDOM WEIGHTS
COMPUTE THE MEAN SQUARE ERROR
REPEAT UNTIL THE MSC IS MINIMIZED
COMPUTE THE GRADIENT TO CHANGE THE WEIGHT
THE LOSS FUNCTION IS STABLE; OPTIMIZATION DONE

Figure 3: Working of logistic regression algorithm

In classification issues where the goal is to estimate the likelihood that a certain sample belongs to a certain class, the logistic regression supervised machine learning approach is crucial. The link between a collection of independent variables and a set of binary dependent variables is investigated using this statistical approach. It is a helpful tool for decision-making in a number of areas, such as the categorization of spam emails, where the model determines whether an email is spam or not depending on specific components of the email..

Here are a few words used often in logistic regression:

Independent variables: Features or predictors are used to estimate the variance of the variable.

Dependent variable: The variables we tried to predict using the logistic regression model. The formula used to show how independent and variable the relationship is is called logistic functions. The probability that the variable is 1 or 0 is represented by a value between 0 and 1, which is the result of a logistic function that transforms the input variable.

Odds: It is the ratio of what has been to those who have not. It differs from probability because probability measures the probability of an event occurring relative to all possible outcomes.

Coefficient: The results of the logistic regression model show the relationship between independence and achievement.

The log variable is represented by a constant called a cutoff in the logistic regression model when all arguments are equal to zero.

A method of estimating the coefficients of a logistic regression model that maximizes data for a given model is called maximum likelihood estimation.

IV. CONCLUSION

The authors of this study set out to look into several machine learning methods for detecting cardiac problems. In this work, K closest neighbour (KNN), logistic regression, and random forest classifiers were utilised as approaches. A survey of journals, published studies, and recent data on cardiovascular illness served as the foundation for the approach for this study. The procedure required a number of processes, including data gathering and the translation of redundant data into a format that was user-friendly. The second phase of the procedure is where the study's primary findings were attained.

Preprocessing the data, also known as the third step, entails analysing the data, resolving missing values, cleaning the data, and normalising it in accordance with the business process. Then, we use classifiers like KNN, logistic regression, and random forest classifiers to categorise the data. Finally, we use a variety of performance indicators to assess the performance and correctness of our suggested model. Doctors and forensic experts can utilise our suggested model as a helpful tool to precisely detect heart illness in the cardiovascular system. To sum up, this study provides a thorough analysis of machine learning methods that can be used to identify cardiac disease.

The techniques and models we introduce can predict more heart disease, which can improve patient outcomes. The model uses 13 medical conditions, including high blood pressure, diabetes, and chest pain.

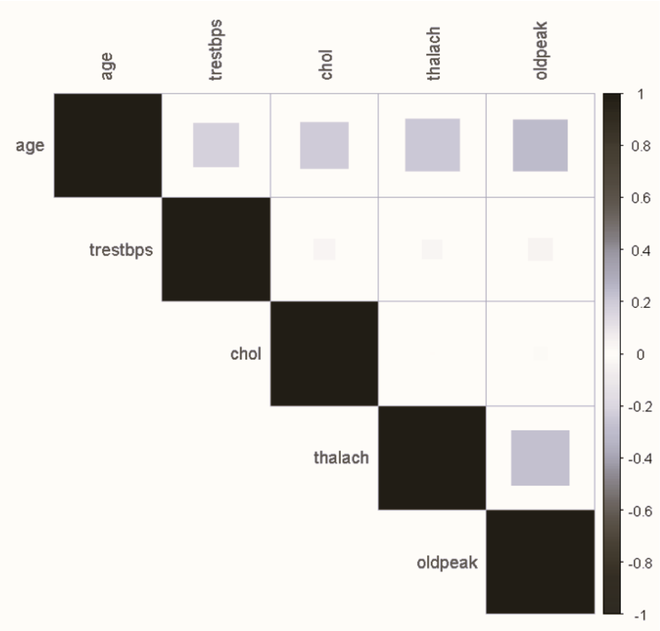

Figure 4: Result obtained from the dataset being used

Methods of machine learning exceed predictions made by humans, making them useful tools for healthcare practitioners. The experiment used data purification, logistic regression, and random forest to correctly forecast heart disease in patients with 98% and 97% accuracy, respectively. We gathered patient medical history information from patients of all ages in order to fully understand this serious health concern. We can identify people with heart disease using this data, which gives us vital details like age, resting blood pressure, fasting glucose, and other medical issues.

V. REFERENCES

[1] Acharya U R, Fujita H, Oh S L, Hagiwara Y, Tan J H & Adam M (2017). Application of deep convolutional neural network for automated detection of myocardial infarction using ECG signals. Information Sciences, 415, 190-8.

[2] 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 (i-Society 2014) (pp. 259-64). IEEE.

[3] Brown N, Young T, Gray D, Skene A M & Hampton J R (1997). Inpatient deaths from acute myocardial infarction, 1982-92: analysis of data in the Nottingham heart attack register. BMJ, 315(7101), 159-64.

[4] Buechler K F & McPherson P H (1999). U.S. Patent No. 5,947,124. Washington, DC: U.S. Patent and Trademark Office.

[5] 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). IEEE.

[6] Dangare Chaitrali 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

[7] Dangare C S & Apte S S (2012). Improved study of heart disease prediction system using data mining classification

techniques. International Journal of Computer Applications, 47(10), 44-8.

[8] Folsom A R, Prineas R J, Kaye S A & Soler J T (1989). Body fat distribution and self-reported prevalence of hypertension, heart attack, and other heart disease in older women. International journal of epidemiology, 18(2), 361-7

[9] 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.

[10] 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). IEEE.

[11] 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.

[12] Kiyasu J Y (1982). U.S. Patent No. 4,338,396. Washington, DC: U.S. Patent and Trademark Office.

[13] Ordonez C (2006). Association rule discovery with the train and test approach for heart disease prediction. IEEE Transactions on Information Technology in Biomedicine, 10(2), 334-43.

[14] 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).

[15] Patel S & Chauhan Y (2014). Heart attack detection and medical attention using motion sensing device -kinect. International Journal of Scientific and Research Publications, 4(1), 1-4.

[16] Piller L B, Davis B R, Cutler J A, Cushman W C, Wright J T, Williamson J D & Haywood L J (2002). Validation of heart failure events in the Antihypertensive and Lipid Lowering Treatment to Prevent Heart Attack Trial (ALLHAT) participants assigned to doxazosin and chlorthalidone. Current controlled trials in cardiovascular medicine, 3(1), 10.

[17] Raihan M, Mondal S, More A, Sagor M O F, Sikder G, Majumder M A & Ghosh K (2016, December). Smartphone based ischemic heart disease (heart attack) risk prediction using clinical data and data mining approaches, a prototype design. In 2016 19th International Conference on Computer and Information Technology (ICCIT) (pp. 299-303). IEEE

[18] 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, 6(1), 637-9.

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

[20] 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, 17(8), 43-8.

[21] Takci H (2018). Improvement of heart attack prediction by the feature selection methods. Turkish Journal of Electrical Engineering & Computer Sciences, 26(1), 1-10.

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

[23] Worthen W J, Evans S M, Winter S C & Balding D (2002). U.S. Patent No. 6,432, 124. Washington, DC: U.S. Patent and Trademark Office.

[24] Zhang Y, Fogoros R, Thompson J, Kenknight B H, Pederson M J, Patangay A & Mazar S T (2011). U.S. Patent No. 8,014,863. Washington, DC: U.S. Patent and Trademark Office.