Heart Disease Early Prediction Using Machine Learning Approaches

Dr.M.Sengaliappan[1], K.Bharathkumar[2]

[1]Head of the Department, Department of Computer Applications, Nehru College of Management,
Coimbatore, TamilNadu, India
ncmdrsengaliappan@nehrucolleges.com
[2]Student of II MCA, Department of Computer Applications, Nehru College of Management,
Coimbatore, TamilNadu, India bk3006848@gmail.com

Abstract: According to the recent WHO (World Health Organization) report, heart diseases are becoming more prevalent. This causes the results in death of 17.9 million individuals every year. It becomes more difficult to detect and begin therapy at an early stage as the population grows. As a result of recent technological advancements with machine learning approaches have speed up the health sector by several researches. The goal of this particular approach is used to develop the machine learning model for early prediction of heart disease using the relevant parameters which this work has taken. The Cleveland dataset has been taken for UCI (University of California Irvine machine learning repository) heart disease prediction, which includes 14 different major parameters for analysis. The development of the model has made use of machine learning methods includes, Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbour (KNN), and Decision Tree (DT). With the use of these conventional machine learning techniques, this work has attempted to identify correlations between the various features present in the dataset with the purpose of effectively predicting early heart disease with maximum accuracy. The suggested approach depicts the outcomes of the three algorithms namely, SVM, Random Forest, and Decision Tree. This present work has integrated all three methods to determine whether the patient is having possibility of heart disease or not. Through that this approach has given 89% accuracy at the end.

Keywords: SVM; Naive Bayes; Decision Tree; Random Forest; Logistic Regression; Adaboost; XG-boost; python programming; confusion matrix; correlation matrix

I.INTRODUCTION

The primary focus of humans is on healthcare. According to WHO recommendations, everyone has a fundamental right to good health. It is believed that suitable health care services should be accessible for routine health checks. Heart-related illnesses account for over 31% of all fatalities worldwide. Because of the absence of diagnostic facilities, qualified physicians, and other resources that affect the precise prognosis of heart disease, early identification [1] and treatment of various cardiac illnesses is very complicated, especially in poor countries. In view of this worry, medical aid software is currently being created using computer technology and machine learning techniques as a support system for the early diagnosis of cardiac disease. The risk of death can be decreased by detecting any heart-related illnesses in their early stages. In order to analyze the patterns in the data and derive predictions from them, many ML techniques are applied in the field of medicine. In general, healthcare data have enormous volumes and complex structures. Big data may be handled by ML algorithms, which can then be mined for useful information. Algorithms for machine learning make predictions based on real-time input and historical data. This type of ML framework for coronary sickness expectancy can motivate cardiologists to act more quickly so that more patients can receive medications in a shorter amount of time, potentially saving a significant number of lives.

An ever-growing area of data science is machine learning, a subfield of AI research [2]. The algorithms used in machine learning are built to handle a wide range of tasks, including prediction, classification, and decision-making. Training data is needed in order to learn the ML algorithms. A model is created during the learning phase and is regarded as the outcome of the ML algorithm.

A set of redundant real-time test datasets are then used to test and validate this model. The model's ultimate accuracy is then compared to the actual number, which validates the model's overall accuracy in predicting the outcome.

II. RELATED WORKS

A number of investigations have been conducted to evaluate the classification accuracy of various machine learning algorithms using the Cleveland heart disease database, which is freely available online at a UCI data mining repository. By using the logistic regression approach on this dataset, the authors of [6] were able to attain a prediction accuracy of 77%. By comparing different global evolutionary computation algorithms in this study, authors [7] improved their work and saw increased prediction accuracy. In their article, authors Bayu Adhi Tama, et al. [8] proposed a study on the use of ML techniques to identify the diabetic disease. It was believed that this condition was highly important to ML. According to research conducted by the International Diabetes Federation (IDF), around 285 million people worldwide have diabetes. Although it is not easy to detect type 2 diabetes in its early stages, the authors' research, which used data mining because they believed it would produce the best results, assisted in the release of information from readily available data. In their study, they used SVMs to collect related data from past records pertaining to various patients. Early diagnosis of type 2 diabetes helped patients receive appropriate care and reduce the risk of complications.

The significance of ML techniques in numerous fields has been proved by a number of applications studied by Yu-Xuan Wang et al. [9]. They suggested a novel method for developing a functional framework. The strategy employed various machine learning techniques. The entire information gathered from the structure was examined when the data miner produced the correct result. The different testing revealed that the suggested strategy produced excellent outcomes. A prior piece on analytics and data mining applications was proposed by Zhiqiang Ge et al. in 2017. These processes were employed in the business world for a variety of reasons. They have examined 10 supervised learning algorithms and 8 unsupervised learning algorithms here [10]. They demonstrated an application for the semi-supervised type learning algorithms in their research. According to industry methods, between 90% and 95% of applications used both supervised and unsupervised machine learning techniques. As a result, it was suggested that machine learning techniques are essential for the design of several unique applications in fields including industry and medical services.

III. MACHINE LEARNING APPROCHES

To develop the heart disease prediction model, we have used three widely used ML approaches. These strategies' specifics are as follows:

Support Vector Machine:

In order to analyze data and find patterns for classification and regression analysis, support Vector Machine [11] classification technique is utilized. SVM is frequently considered when the data is categorized as a two-class problem. Finding the optimum hyper plane that isolates every data point from one class to the other is how this technique identifies data. The better the model is taken into account, the greater the separation or edge between the two classes. The support vectors are the data points that are located near the margin's edge. SVM is actually built on mathematical techniques for creating challenging real-world issues. Our dataset, the Cleveland Heart Disease Dataset (CHDD), has multiple classes that can be predicted based on different characteristics, thus we decided to apply SVM for this project. A function known as a kernel (Kernels of SVM) is used in SVM to map training data. Examples of kernels include linear kernel, quadratic kernel, polynomial kernel, radial basis function kernel, multilayer perceptron kernel, etc. In addition to the SVM kernel features, a few more techniques are also accessible, including least squares, sequential minimal optimization, and quadratic programming.

The hardest part of using SVM to build a model is choosing the kernel and strategy to avoid overfitting and underfitting problems. Due to the vast number of parameters and cases in our dataset. So, we had the option of choosing either the RBF or the linear kernel. As a result, the final SVM model needs to be evaluated against real data.

Decision Tree:

Machine learning's Decision Tree method [12] is used to create Classification models. This classification approach is based on a structure that resembles a tree. This falls within the supervised learning category because the desired outcome is already known. The decision tree algorithm can be used with category and numerical data. The root node, branches, and leaf nodes make up a decision tree. The traversal path from the root to a leaf node is used as the basis for evaluating the data. A total of 283 tuples from our dataset, CHDD, were evaluated down the decision tree. They could have reached a favorable or unfavorable conclusion on the prognosis of heart disease. To verify for false positives or false negatives, these were compared to the actual parameters. This demonstrates the accuracy, specificity, and sensitivity of the model.

Random Forest Classification:

A collection of unpruned classification-based trees makes up Random Forest [15]. Given that it is ineffective against noise in many real-life situations, it provides excellent performance. The dataset and overfitting risk are both quite low. It operates more quickly than many other tree-based algorithms and typically increases accuracy for testing and validation data. Individual decision tree algorithm forecasts are combined to form random forests. When building a random tree, there are several options for adjusting the performance of the random forest.

IV. METHODOLOGY

The process used to construct the heart disease prediction model is shown in the steps below.

A. Data Collection

We used the Cleveland Heart Disease Dataset, which is available online at the UCI Repository [16].

The 14 attributes taken into account are as follows:

S.No	Attribute	Desc.	Mean Value
1.	age	in years	54.496
2.	Sex	Male, Female	0.6767
3.	cp	Angina, abnang, notang, asympt	2.158
4.	trstbps	Resting Blood Pressure in mm hg	131.693
5.	chol	Serum Cholesterol in mg/dl	247.35
6.	fbs	fasting blood sugar- 1 if >120 mg/dl, 0 if	0.144
7.	restecg	Electrocardiographic Results	0.996
8.	thalach	Maximum Heart Rate observed	149.59
	exang	exercise with angina has occurred	0.326
9.	oldpeak	ST depression induced through exercise	1.055
10.	slope	slope of the ST segment	0.602
11.	thal	Number of major vessels ranging from 0 - 3 color by fluoroscopy	0.835
12.	ca	Heart status	0.67

Hence, a total of 1189 instances were used for this study project. The mean value for each property is displayed in Table 1.

Data preprocessing is used to deal with the dataset's missing values. Class Value 1, which means "tested positive for the disease," and Class Value 0, which means "tested negative for the disease," are both equivalent. The dataset was split into different percentages, with training data accounting for 80% of the data and testing data for the remaining 20%. Sample training data are shown in Table II, while sample testing data are shown in Table III.

B. Data Preprocessing

Many attributes in the original dataset have missing values, which might result in inaccurate results and degrade the model's accuracy. The best way to solve this issue is to replace missing values using the "mean of column" method. This approach substitutes 0 with either the neighborhood's mean value or the neighborhood's average value [17]. The 0 value is then modified in accordance with the newly determined value. After that, the dataset's values were converted from numeric to nominal so that it would work with the ML techniques being applied.

TABLE II. SAMPLE TRAINING DATA

69	1	0	160	234	1	2	131	0	0.1	1	1	0	0
69	0	0	140	239	0	0	151	0	1.8	0	2	0	0
66	0	0	150	226	0	0	114	0	2.6	2	0	0	0
65	1	0	138	282	1	2	174	0	1.4	1	1	0	1
64	1	0	110	211	0	2	144	1	1.8	1	0	0	0
64	1	0	170	227	0	2	155	0	0.6	1	0	2	0
63	1	0	145	233	1	2	150	0	2.3	2	0	1	0
61	1	0	134	234	0	0	145	0	2.6	1	2	0	1
60	0	0	150	240	0	0	171	0	0.9	0	0	0	0
59	1	0	178	270	0	2	145	0	4.2	2	0	2	0

TABLE III. SAMPLE TRAINING DATA

48	1	3	130	256	1	2	150	1	0	0	2	2	1
47	1	3	110	275	0	2	118	1	1	1	1	0	1
47	1	3	112	204	0	0	143	0	0.1	0	0	0	0
46	0	3	138	243	0	2	152	1	0	1	0	0	0
46	1	3	140	311	0	0	120	1	1.8	1	2	2	1
46	1	3	120	249	0	2	144	0	0.8	0	0	2	1
45	1	3	104	208	0	2	148	1	3	1	0	0	0
45	0	3	138	236	0	2	152	1	0.2	1	0	0	0
45	1	3	142	309	0	2	147	1	0	1	3	2	1
45	1	3	115	260	0	2	185	0	0	0	0	0	0

C. Building Model

Weka Data Mining Tool is used to build the model. An open-source machine learning programme called The Waikato Environment for Knowledge Analysis (WEKA) [18] was created by the Waikato University. New Zealand's Waikato University. The software handles a variety of common data mining tasks, such as feature selection, data pre-processing, clustering, classification, and regression, with ease. It provides a straightforward setting for loading data in the form of files, URLs, or databases. The software supports the Attribute Relation File Format (ARFF) [19], CSV, C4.5, and Lib SVMs file formats.

It provides a straightforward analysis and visualisation of the confusion matrix, true positive, accuracy, recall, false negative, etc. It is portable, GUI-based, platform neutral, open source software that is also loaded with a variety of cutting-edge machine learning methods, including deep learning algorithms for image processing, etc.

The following four accuracy measures were taken into consideration while comparing the three models:

Positive Predictive Value or Precision:

Precision is defined as the ratio of true positives to false positives.

Precision = Number of true positives/Number of true positives + False positives.

Recall:

It is the average probability of complete retrieval Recall=

True positives/True positives + False negative. **Accuracy:**

The accuracy of a classifier is given as the percentage of total correct predictions divided by the total number of instances. Accuracy = [Number of True Positives + True Negatives]/[Total Instances]

V. RESULTS:

At the conclusion of our research, the results of decision tree model show more accuracy than the SVM and random forest. In comparison with these models, decision tree gives 3% more than the SVM and random forest. After ensemble, the accuracy of the model will increase to 97.7% which is 13% more the obtained comparison results.

TABLE IV. PERFORMANCE MEASURE OF MODELS

Models	Accuracy	Precision	Sensitivity recall
Support Vector Machine	0.81	0.890	0.896
Decision Tree	0.842	0.916	0.910
Random Forest	0.816	0.816	0.902

Ensemble model accuracy = 97.78

Ensemble model loss = 2.22

VI. CONCLUSION

Through this research, we attempted to analyze the numerous machine learning algorithms and predict whether or not a specific person will develop cardiac illness given various personal characteristics and indications. Our report's main focus was on examining the accuracy and examining the causes of the variations among various algorithms. The Cleveland dataset for cardiac illnesses, which has 1189 cases, was used to divide the data into training and testing datasets using a percent split. To evaluate the accuracy, we used four distinct algorithms and 14 different attributes. After the implementation phase is complete, we have found that ensemble model of SVM, random forest and decision tree provides 97.78%. Although another approach may function more effectively for other situations and datasets, we have found that this result works well in our case. Additionally, if we increase the amount of training data, we might be able to obtain results that are more accurate, but processing time would be longer, and the system would be slower than it is currently since it would have to deal with more data and be more complex. We made this decision since it is easier for us to work with after taking these potential factors into account.

References:

[1] https://www.who.int/hrh/links/en/

[2] https://en.wikipedia.org/wiki/Machine learning

[3] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia and J. Gutierrez, "A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease," 2017 IEEE Symposium on Computers and Communications (ISCC), Heraklion, 2017, pp.204-207, doi: 10.1109/ISCC.2017.8024530.

[4] S. Dhar, K. Roy, T. Dey, P. Datta and A. Biswas, "A Hybrid Machine Learning Approach for Prediction of Heart Diseases," 2018 4th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 2018, pp. 1-6, doi: 10.1109/CCAA.2018.8777531.

[5] C. Raju, E. Philipsy, S. Chacko, L. Padma Suresh and S. Deepa Rajan, "A Survey on Predicting Heart Disease using Data Mining Techniques," 2018 Conference on Emerging Devices and Smart Systems (ICEDSS), Tiruchengode, 2018, pp. 253-255, doi: 10.1109/ICEDSS.2018.8544333.

[6] R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J.-J. Schmid,S. Sandhu, K. H. Guppy, S. Lee, and V. Froelicher, “International application of a new probability algorithm for the diagnosis of coronary artery disease,” The American journal of cardiology, vol. 64, no. 5, pp. 304–310, 1989.

[7] B. Edmonds, “Using localised ’gossip’ to structure distributed learning,” 2005.

[8] Fsdfsdf BayuAdhi Tama,1 Afriyan Firdaus,2 Rodiyatul FS, “Detection of Type 2 Diabetes Mellitus with Data Mining Approach Using Support Vector Machine”, Vol. 11, issue 3, pp. 12-23, 2008.

[9] Yu-Xuan Wang, QiHui Sun, Ting-Ying Chien, Po-Chun Huang, “Using Data Mining and Machine Learning Techniques for System Design Space Exploration and Automatized Optimization”, Proceedings of the 2017 IEEE International Conference on Applied System Innovation, vol. 15, pp. 1079-1082, 2017.

[10] ZhiqiangGe, Zhihuan Song, Steven X. Ding, Biao Huang, “Data Mining and Analytics in the Process Industry: The Role of Machine Learning”, 2017 IEEE Translations and contentmining are permitted for academic research only, vol. 5, pp. 20590-20616, 2017.

[11] https://en.wikipedia.org/wiki/Support_vector_ machine

[12] https://en.wikipedia.org/wiki/Decision_tree_le arning

[13] https://en.wikipedia.org/wiki/Bayes27_theore m

[14] https://en.wikipedia.org/wiki/Naive_Bayes_cla ssifier

[15] https://towardsdatascience.com/understanding-random-forest-58381e0602d2

[16] https://archive.ics.uci.edu/ml/datasets/heart+di sease

[17] https://wekatutorial.com/

[18] https://www.cs.waikato.ac.nz/ml/weka/mooc/dataminingwithweka/slides/Class5DataMiningWithWeka-2013.pdf

[19] https://www.cs.waikato.ac.nz/ml/weka/arff.ht ml