International Journal of Applied Engineering Research ISSN 0973-4562 Volume 18, Number 2 (2023) pp. 136-139

Identification of Heart Disease in Beginning Stage using Gradient Boosting Classification

Dr. P. Senthil Pandian[1], Dr. R. RubeshSelvaKumar[2], Dr. R. Muneeswaran[3] and Mrs. S. Valli Mayil[4]

[1]Associate Professor, Department of CSE, Solamalai College of Engineering, Madurai, Tamilnadu.
[2]Associate Professor, Department of CSE, Sethu Institute of Technology, Virudhunagar, Tamilnadu.
[3]Associate Professor and Head, Department of Mechanical Engineering,
Solamalai College of Engineering, Madurai, Tamilnadu
[4]Assistant Professor, Department of CSE, Solamalai College of Engineering, Madurai, Tamilnadu.
E-Mail: psenthilpandian@gmail.com, dr.rubeshrubesh1972@gmail.com.

Abstract:
Heart illnesses are growing more common, according to a recent WHO (World Health Organisation) report. Every year, this results in the deaths of 17.9 million people. As the population increases, it becomes more challenging to diagnose an illness and start treatment at an early stage. Recent technology developments and machine learning techniques have accelerated various studies in the health field. Using the pertinent parameters that this work has taken, the objective of this particular technique is to construct a machine learning model for early heart disease prediction. The Cleveland dataset, which includes 14 different key parameters for study, has been taken for the UCI (University of California Irvine machine learning repository) heart disease prediction. Machine learning techniques such as Gradient Boosting (RF), Extreme Learning Machine (ELM), K-Nearest Neighbour (KNN), and Decision Tree (DT) were used in the model's construction. This work has sought to find connections between the many variables contained in the dataset using these traditional machine learning approaches in an effort to predict early cardiac disease as accurately as possible. The proposed strategy illustrates the results of the three algorithms, ELM, Gradient Boosting, and Decision Tree. To ascertain whether the patient is at risk for heart disease or not, our current work has incorporated all three methodologies. By using this method, the final accuracy was 89%.

Keywords: Humans Healthcare, Logistic Regression, Machine Learning Algorithms, Extreme Learning Machine, Gradient Boosting Classification.

INTRODUCTION

Humans are primarily concerned with their health. WHO guidelines state that everyone has a fundamental right to good health. For regular health checkups, it is thought that sufficient health care services should be available. Over 31% of all fatalities globally are caused by heart-related conditions. Early diagnosis and treatment of different cardiac ailments is particularly difficult, especially in developing nations, due to the lack of diagnostic facilities, qualified doctors, and other resources that affect the precise prognosis of heart disease. To help with the early diagnosis of cardiac disease, medical assistance software is currently being developed using computer technology and machine learning methods. By identifying any heart-related disorders in their early stages, the chance of death can be reduced. Numerous ML approaches are used in the field of medicine to analyze the patterns in the data and make predictions from them.

Healthcare data typically has huge volumes and intricate architecture. ML algorithms may be able to handle big data, which can subsequently be mined for insightful data. Machine learning algorithms produce predictions based on input from the present and past. By encouraging cardiologists to act more rapidly so that more patients can receive treatment in a shorter amount of time, this kind of ML framework for coronary sickness expectancy could possibly save a considerable number of lives. Machine learning, a branch of AI study, is a rapidly expanding area of data science [2]. Machine learning algorithms are designed to perform a variety of tasks, such as prediction, classification, and decision-making. Learning the ML algorithms requires training data. A model is produced by the ML algorithm and is thought of as its output during the learning phase. The model is subsequently tested and validated using a collection of redundant real-time test datasets. The model's overall accuracy in predicting the outcome is validated by comparing the model's final accuracy to the actual number.

RELATED WORKS

The Cleveland heart disease database, which is freely accessible online at a UCI data mining repository, has been the subject of several studies to assess the classification accuracy of different machine learning techniques. On this dataset, the authors of [6] were able to achieve a prediction accuracy of 77% by utilizing the logistic regression approach. In this study, authors [7] enhanced their work and noticed improved prediction accuracy by contrasting various global evolutionary computation algorithms. Authors Bayu Adhi Tama, et al. [8] suggested a study on the application of ML approaches to the diagnosis of diabetes in their publication.

This condition was thought to be very significant to ML. Around 285 million people worldwide have diabetes, according to studies by the International Diabetes Federation (IDF). Although early-stage type 2 diabetes is difficult to identify, the authors' research—which included data mining because they thought it would yield the best results—helped to unlock information from publicly accessible data. In their research, they employed ELMs to gather pertinent information from old medical records. Early type 2 diabetes detection aided patients

in receiving the right therapy and lowered the chance of complications.

A number of applications examined by Yu-Xuan Wang et al. [9] have demonstrated the importance of ML approaches in a variety of domains. They proposed a fresh approach to creating a practical framework. The plan made use of numerous machine learning methods. When the data miner yielded the desired outcome, all the data acquired from the structure was reviewed. The various tests showed that the suggested technique delivered outstanding results. Zhiqiang Ge et al. offered an earlier study on applications for analytics and data mining in 2017. For a variety of reasons, these procedures were applied in the corporate sector. Ten supervised learning algorithms and eight unsupervised learning algorithms have each been studied here [10]. In their study, they showed how semi-supervised type learning algorithms can be used. According to industry approaches, supervised and unsupervised machine learning techniques were deployed in between 90% and 95% of applications. It was thus recommended that the design of various distinctive applications in domains like industry and medical services requires the use of machine learning techniques.

MACHINE LEARNING APPROACHES

In this paper it is employees three popular ML techniques to create the heart disease prediction model. The specifics of these tactics are as follows:

Extreme Learning Machine (ELM):

Extreme Learning Machine [11] classification approach is used to analyze data and discover patterns for classification and regression analysis. When the data is categorized as a two-class problem, ELM is typically taken into account. This technique locates the optimal hyper plane that isolates each data point from one class to the other. The larger the edge or gap between the two classes, the better the model is taken into account. The data points that are close to the margin's edge are known as the Extreme Learning's. ELM is essentially based on mathematical methods for generating complex problems in the real world.

Its helps to chose to use ELM for this research since the dataset, the Cleveland Heart Disease Dataset (CHDD), comprises several classifications that can be predicted based on various attributes. To map training data in ELM, a kernel function is used (Kernels of ELM). Linear kernels, quadratic kernels, polynomial kernels, radial basis function kernels, multilayer perceptron kernels, etc. are a few examples of kernels. There are a few more methods available in addition to the ELM kernel characteristics, such as least squares, sequential minimal optimization, and quadratic programming. Choosing the kernel and approach to avoid overfitting and underfitting issues is the most difficult component of utilizing ELM to construct a model. Because our dataset has a huge number of factors and cases. So, we have a choice between the RBF and the linear kernel. Therefore, it is necessary to compare the final ELM model to actual data.

Decision Tree:

Classification models are developed using the Decision Tree approach in machine learning [12]. This method of categorization is based on a structure that looks like a tree. Given that the intended result is already known, this falls under the heading of supervised learning. Both categorical and numerical data can be used with the decision tree technique. A decision tree is made up of the root node, branches, and leaf nodes. The basis for judging the data is the traversal path from the root to a leaf node. The CHDD dataset's 283 tuples in total were examined as they moved down the decision tree. They might have come to a positive or negative conclusion regarding the prognosis of cardiac disease. These were compared to the actual parameters to make sure there weren't any false positives or false negatives. This displays the model's precision, specificity, and sensitivity.

Gradient Boosting Classification:

Gradient Boosting [15] is a collection of unpruned classification-based trees. Given that it is useless against noise in many real-world circumstances, it performs superbly. Both the dataset and the overfitting risk are very small. In comparison to many other tree-based algorithms, it operates more quickly and often improves data accuracy for testing and validation. Gradient Boosting are created by combining the forecasts from different decision tree algorithms. There are numerous ways to modify the Gradient Boosting's efficiency while creating a random tree..

METHODOLOGY

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

Collection

The Cleveland Heart Disease Dataset, which is online at the UCI Repository [16], is the suggested system in this study. Table 1 displays the 8 attributes considered.

Table 1: Heart Disease Dataset

S.No	Attribute	Desc.	Mean Value
1.	trstbps	Resting Blood Pressure in mm hg	131.693
2.	chol	Serum Cholesterol in mg/dl	247.35
3.	fbs	fasting blood sugar-1 if >120 mg/dl, 0 if	0.144
4.	restecg	Electrocardiographic Results	0.996
5.	thalach	Maximum Heart Rate observed	149.59
6.	oldpeak	ST depression induced through exercise	1.055
7.	slope	slope of the ST segment	0.602
8.	thal	Number of major vessels ranging from 0-3 color by fluoroscopy	0.835

The missing values in the dataset are handled via data preprocessing. Class Value 0, which denotes "tested negative for the disease," is equivalent to Class Value 1, which denotes "tested positive for the disease." With training data making up 80% of the dataset and testing data the remaining 20%, the dataset was divided into distinct percentages.

B. Data Preprocessing

The original dataset has many attributes with missing values, which could lead to erroneous findings and reduce the model's accuracy. The "mean of column" approach is the most effective strategy to replace the missing numbers in this situation. With this method, 0 is replaced with either the neighborhood's mean value or its average value [17]. Then, using the just discovered value, the 0 value is updated. The dataset's values were then changed from being numerical to being nominal so that they could be used with the ML approaches being used and shown in Tables 2 and 3.

Table 2: 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 3: 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

To create the model, Weka Data Mining Tool is employed. The Waikato Environment for Knowledge Analysis (WEKA) [18] is an open-source machine learning programme developed by Waikato University. The Waikato University in New Zealand. Numerous typical data mining tasks, including feature selection, data pre-processing, clustering, classification, and regression, are easily handled by the software. It offers a simple environment for loading data from databases, URLs, or files. The Attribute Relation File Format (ARFF) [19], CSV, C4.5, and Lib ELMs file formats are all supported by the software.

The confusion matrix, true positive, accuracy, recall, false negative, etc. are all easily analysed and visualised. It is open source software that is portable, GUI-based, platform neutral, and packed with a wide range of cutting-edge machine learning techniques, such as deep learning algorithms for image processing.

When contrasting the three models, the following four accuracy metrics were considered:

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 percentage of all right predictions divided by the total number of cases represents a classifier's accuracy.

Accuracy is calculated as = [Number of True Positives plus True Negatives] / [Total Instances].

RESULTS AND DISCUSSION

At the conclusion of our study, the decision tree model's results demonstrate greater accuracy than those of the ELM and Gradient Boosting. Decision tree provides 3% more than the ELM and Gradient Boosting in compared to these models. Following ensemble, the model's accuracy will rise to 97.7%, 13% higher than the observed comparative values shown in Table 4.

Table 4: Performance Measure of Models

Models	Accuracy	Precision	Sensitivity recall
Extreme Learning Machine	0.81	0.890	0.896
Decision Tree	0.842	0.916	0.910
Gradient Boosting	0.816	0.816	0.902

Ensemble model accuracy = 97.78

Ensemble model loss = 2.22

CONCLUSION

By analyzing the various machine learning algorithms, we tried to determine whether or not a particular person would experience cardiac sickness based on several personal traits and indicators. The major goals of our report were to assess the precision and investigate the reasons behind algorithmic deviations. The data were separated into training and testing datasets using a percent split utilizing the 1189 case Cleveland dataset for cardiac diseases. To evaluate the accuracy, we used 14 different attributes and four different algorithms. We discovered that the ensemble model comprising ELM, Gradient Boosting, and decision tree provides 97.78% after the implementation phase is complete. We have discovered that this outcome works well in our case, despite the possibility that another strategy would be more useful for different scenarios and datasets. Additionally, if we expand the training data, we might be able to get more accurate answers, but processing time would be longer, and the system would be slower than it is now

since it would have to deal with more data and be more sophisticated. We chose this course of action because, after considering these potential factors, it is simpler for us to work with.

References:

[1] 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.

[2] 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.

[3] 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.

[4] 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, 2010.

[5] Rodiyatul FS, "Detection of Type 2 Diabetes Mellitus with Data Mining Approach Using Extreme Learning Machine", Vol. 11, issue 3, pp. 12-23, 2018.

[6] 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.

[7] Ding Biao Huang, "Data Mining and Analytics in the Process Industry: The Role of Machine Learning", 2017 IEEE Translations and content mining are permitted for academic research only, vol. 5, pp. 20590-20616, 2017.

[8] https://en.wikipedia.org/wiki/Support_vector_machine

[9] https://en.wikipedia.org/wiki/Decision_tree_learning

[10] https://en.wikipedia.org/wiki/Bayes27_theorem

[11] https://en.wikipedia.org/wiki/Naive_Bayes_classifier

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

[13] https://archive.ics.uci.edu/ml/datasets/heart+disease

[14] https://wekatutorial.com/

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

[16] https://www.cs.waikato.ac.nz/ml/weka/arff.html