WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Comparative Study of Machine Learning Algorithms for Heart Disease Prediction.

RUPIKA.V[1] , SASMITHA.R[2] , LATHIKA.R[3]

[1] Student, Department of Software Systems and AIML, Sri Krishna Arts and Science College, Coimbatore-641008, Tamil Nadu, India
[2] Student , Department of Software Systems and AIML , Sri Krishna Arts and Science College, Coimbatore-641008, Tamil Nadu, India
[3] Student , Department of Software Systems and AIML , Sri Krishna Arts and Science College, Coimbatore-641008, Tamil Nadu, India

ABSTRACT:

Cardiovascular disease remains among the top causes of death across the globe, with a life lost to it nearly every minute in countries such as the United States. The increasing accessibility of patient health records has made it possible to use machine learning and data mining algorithms to aid in early and correct diagnosis. These methods can minimize the amount of medical tests needed while achieving quicker and more accurate predictions. This paper compares three decision tree–based classifiers, J48, Logistic Model Trees (LMT), and Random Forest, on the Cleveland heart disease dataset from the UCI Machine Learning Repository, which comprises 303 patient records with 76 attributes. The aim is to determine the best algorithm for heart disease presence prediction. Through the revelation of concealed data patterns, the research seeks to assist physicians in determining patient risk, from none to having a high chance of disease. The research illustrates the promise of decision tree–based models as decision-support tools in medicine.

Keywords: Machine Learning, Data Mining, Decision Tree, Heart Disease Prediction, Healthcare Analytics

I. Introduction

Cardiovascular diseases continue to be the leading cause of death globally in the last decade (WHO, 2007). European Public Health Alliance reports (2010) indicate that heart attacks, strokes, and other circulatory disorders contribute to close to 41% of all global deaths. The problems with such diseases are that they come in a vast array of symptoms, making timely and correct diagnosis hard to obtain. Historically, doctors depend on their personal knowledge and clinical experience to balance various patient characteristics—age, blood pressure, cholesterol level, family history, and lifestyle factors—when deciding on diagnosis. But as healthcare systems produce large amounts of electronic health records, there is scope to use this information for better insights. Through such data mining, one can identify concealed patterns and leverage them to help doctors diagnose heart disease more accurately. Data mining is a branch of machine learning that deals with extracting implicit, significant

knowledge and not-so-obvious relationships from large datasets (Lee et al., 2000). In medicine, it has proven quite promising in aiding policy-making, cutting down on errors, enhancing preventive treatment, and helping with early detection. When put into practice clinically, data-based prediction models may enhance medical know-how, increase precision, and curtail inappropriate testing. This research concentrates on decision tree classifiers, namely J48, Logistic Model Trees, and Random Forest, for the prediction of heart disease probability. Employing the Cleveland dataset of the UCI Machine Learning Repository and the WEKA tool for experimentation, this work will test the efficiency of these algorithms and assess the best method of aiding medical decision-making in actual healthcare practice.

A. LITERATURE REVIEW

Heart disease prediction using data mining and machine learning has been a prominent area of research for the last two decades. Numerous classification methods—e.g., Decision Trees, Naïve Bayes, Neural Networks, Support Vector Machines (SVM), Kernel Density Estimation, Bagging, and other ensemble methods—have been tried on patient data with varying degrees of success (Yan et al., 2003; Das et al., 2009; Raj Kumar & Reena, 2010). A very prominent difference among these studies lies in the selection of parameters and datasets utilized. Sitair -Taut et al., for instance, used the Naïve Bayes and J48 decision tree algorithms of WEKA to classify coronary heart disease. Bagging techniques in WEKA were used by Tu et al. and compared with J48, with significant improvements in prediction reported. Other researchers showed that Random Forest classifiers are very good for medical diagnosis, even equalling or outperforming Bayesian approaches at times in terms of accuracy (Sitar-Taut et al., 2009). Vijiya rani et al. compared various tree-based algorithms such as Decision Stump, Random Forest, and Logistic Model Trees on heart disease datasets in 2013. Their study demonstrated that the performance of the algorithm can be very diverse based on the set of attributes and data type used. Overall, existing research indicates that although many machine learning algorithms are applicable to cardiovascular diagnosis, decision tree models are unique because they are interpretable,

easy to use, and robust in their performance using various datasets. They are therefore a trusted option for assisting with clinical decision-making in medical settings. .[1][3][6][9]

II. BACKGROUND

Heart disease strikes millions of individuals every year and is still the most prevalent cause of mortality among men and women worldwide. Cardiovascular diseases, as reported by the World Health Organization (WHO), cause around 12 million deaths every year, with one death from heart disease occurring every 34 seconds globally. Precise diagnosis is an important factor in prevention and treatment, yet it tends to be complicated, time-consuming, and expensive. To minimize the reliance on large-scale clinical tests, computer decision support systems are being more widely implemented. These systems use data mining and machine learning approaches to reveal latent patterns in patient histories, thus helping medical professionals to more accurately predict disease risk. Research has pointed to a number of major risk factors closely linked to heart disease, including age, hypertension, increased cholesterol, diabetes, obesity, family history, physical inactivity, and abnormal fasting glucose. Through these characteristics, machine learning models can separate high-risk subjects from those unlikely to develop cardiovascular diseases. Of the numerous techniques used, Decision Trees have proven especially well-suited to medical prediction tasks. They can efficiently process categorical and numerical data, deal with missing values, and provide easily interpretable models. Continuous variables, however, usually have to be re-expressed as discrete categories first. For the purposes of reliability, techniques like reduced-error pruning are used to avoid overfitting and improve accuracy.

For this research, the Cleveland Heart Disease dataset from the UCI archive was employed. It contains 303 patient records with 76 features. Following preprocessing, only 13 most important features—age, chest pain type, cholesterol, fasting blood sugar, maximum heart rate, ST depression, etc.—were kept in analysis. Those chosen features are the foundation upon which the predictive capability of decision tree algorithms for heart disease diagnosis is measured.[4]

III. APPROACH AND METHODOLOGY

The heart disease prediction model development has two principal objectives:

1) 1. The system must not be based on pre-existing knowledge concerning patient data.
2) 2. It should be scalable in order to efficiently work with large medical datasets.

Tool Selection

The model is deployed on WEKA, which is an open-source machine learning environment. WEKA offers in-built methods for preprocessing, classification, clustering, association rules, and visualization. Decision tree classifiers were used in this research through 10-fold cross-validation to facilitate unbiased
performance **assessment**

Dataset

The experiments employ the Cleveland Heart Disease database in the UCI Machine Learning Repository. The database initially has 303 patient records with 76 features.

After preprocessing, only 13 major features were retained for analysis because they were identified as the most important for prediction.

Selected Attributes:

Attribute	**Type**	**Description**
Age	**Continuous**	**Age of patient(years)**
Sex	Discrete	0=female,1=male
Cp	Discrete	Chest pain type (1=typical anginal,2=atypical angina,3=nonanginal,4=asymptomatic)1
Trestbps	Continuous	Resting blood pressure (mmHg)
Chol	Continuous	Serum cholesterol(mg/dl)
Fbs	Discrete	Fasting blood sugar>120mg/dl(1=true,0=false)
Restecg	Discrete	Resting electrocardiographic results
Thalach	Continuous	Maximum heart rate achieved
Exang	Discrete	Exercise-induced angina(1=yes,0=no)

Old peak	Continuous	ST depression induced by exercise relative to rest
slope	discrete	Slope of the peak exercise ST segment (1=upsloping,2=flat,3= down sloping)
Ca	Continuous	Number of major vessels coloured by fluoroscopy (0-3)
Thal	Discrete	3-normal,6=fixed defect,7=reversible defect

The target attribute is the Class, which represents the presence or absence of heart disease, ranging from 0 = no presence to 4 = high likelihood.

Methodology Steps

1. Import dataset into WEKA and convert it to ARFF format.
2. Preprocess by removing irrelevant attributes and retaining the 13 significant ones.
3. Apply decision tree algorithms (J48, Logistic Model Trees, Random Forest).
4. Train and test the models using 10-fold cross-validation.
5. Evaluate performance using confusion matrices and metrics: accuracy, sensitivity, specificity, precision, recall, F-measure, and ROC area.

Classification Tree Algorithms Used

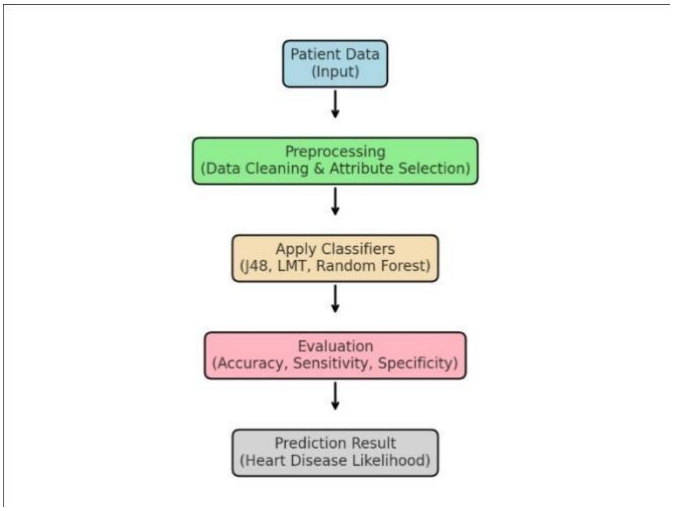

Figure 1: Workflow diagram showing the steps from patient data collection to heart disease prediction using classifiers.

In this study, three decision tree–based classifiers were used to classify the existence of heart disease. They were selected due to their established performance in medical data mining but vary in model construction and dealing with data complexity.

1. J48 (C4.5 Implementation)

J48 is WEKA's version of the popular C4.5 algorithm, an extension of the earlier ID3 decision tree learner. J48 recursively splits the dataset on the attribute that has the maximum information gain in each step. This continues until the data can be split no more and the leaf nodes contain classification outcomes.

One of the problems with decision trees is that they get overly complicated and fit the training data exactly, hence overfitting. J48 addresses this by using reduced-error pruning, removing branches that don't enhance prediction accuracy. This reduces the tree's size, increases processing speed, and improves its ability to deal with new test data.

Strengths of J48:

- Easy to interpret and visualize.
- Handles categorical as well as numerical attributes.

- Offers rules that are directly interpretable by medical professionals

Limitations:

- Illicit to noisy data and noise attributes.
- Trees can still become large for complex datasets**.**

For heart disease prediction, J48 is attractive because it produces clear and transparent rules, making it easier for doctors to validate the system's decisions.

2. Logistic Model Trees (LMT)

Logistic Model Trees merge the tree structure of decision trees with the predictive accuracy of logistic regression. Rather than putting a single class label on the leaves, every leaf node holds a logistic regression model. This enables the classifier to take advantage of linear associations between input features and the target variable, yet utilize a decision tree to manage nonlinearity.
The tree is built employing the Logit Boost algorithm, which iteratively refines the logistic regression models by minimizing classification errors. The model is then pruned using cost-complexity pruning so that unnecessary branches are eliminated.

Strengths of LMT:

- Averages interpretability of trees with statistical rigor of regression.
- Works well with both categorical and continuous attributes.
- Builds smaller and frequently more accurate trees than J48.

Limitations:

- Averages interpretability of trees with statistical rigor of regression.
- Works well with both categorical and continuous attributes.

LMT is especially useful for medical datasets like heart disease because it can learn fine relationships between risk factors like cholesterol, blood pressure, and age, and still give a tree structure for interpretability.

3. Random Forest

Random Forest is a type of ensemble learning that creates many decisions trees and takes their predictions. Each tree is built on a new bootstrap sample of the dataset, and at each node,
a random subset of features is chosen to find the best split. The prediction is made by majority voting among all trees. This approach lowers the danger of overfitting, which is prevalent in individual decision trees, and typically enhances predictive accuracy

Strengths of Random Forest:

- Resistant to noise and missing values.
- Prevents overfitting relative to a single tree.
- Efficient in handling high-dimensional data.
- Allows variable importance insights.

Limitations:

- Less interpretable than the use of a single tree such as J48.
- More computational intensive.

In the case of heart disease prediction, Random Forest is able to provide high accuracy and stability at the cost of decreased interpretability, which could be a limitation in clinical decision-making.[7]

IV. Results

The classification experiments were performed with J48,
Logistic Model Trees, and Random Forest algorithms on the Cleveland dataset (UCI repository). All the experiments were run in WEKA 3.6.10 on an Intel Core i3 (2.4 GHz CPU) system with 4 GB RAM. The results are presented in a grouped manner by algorithm to facilitate comparison.

A. J48 with Reduced-Error Pruning

The J48 algorithm was used with reduced-error pruning, and the resulting confusion matrix obtained for the five class labels is displayed below:

----------Confusion Matrix (J48) ----------

	a	b	c	d	e
a	152	7	2	3	0
b	34	4	10	5	2
c	10	11	7	7	1
d	5	11	12	5	2
e	1	5	2	3	2

Performance of J48:

Algorithm	Training error	Test error
Random forest	0.0	0.2

The J48 model provides a fair trade-off between interpretability and accuracy. Although the error rate is marginally simpler than more sophisticated models, the transparency of decision rules renders it extremely useful for application in medicine..

B. Logistic Model Tree (LMT)

The Logistic Model Tree classifier was also run on the same data.
Its confusion matrix is shown below:

Confusion Matrix (LMT):

	a	b	c	d	e
a	148	12	2	1	1
b	31	10	6	8	0
c	8	12	4	10	2
d	4	11	11	7	2
e	0	5	2	6	0

Performance of LMT:

Algorithm	Training error	Test error
Logistic model tree	0.1156	0.1379

The LMT algorithm had lower training and testing errors than J48, generating more elegant models. But its computational expense is substantially greater, which makes it slower when dealing with larger datasets.

C. Random Forest

The Random Forest classifier, which constructs many decision trees using randomly selected subsets of features, was also used.
The results are outlined below:

Confusion Matrix (Random Forest):

	a	b	c	d	e
a	146	8	4	6	0

b	31	9	9	6	0
c	9	5	13	8	1
d	11	7	10	4	3
e	2	5	3	3	0

Performance of Random Forest:

algorithm	Training error	Test error
J48	0.1423	0.1666

Random Forest displayed zero training error, reflecting its flexibility and ability to fit the training data completely. However, the test error of 0.20 suggests slight overfitting on the given dataset. [2].

V. COMPARISON OF METHODOLOGIES

All three classifiers, J48, Logistic Model Tree (LMT), and Random Forest, were compared based on training and test error measures

Comparison of Algorithm Results

It can be seen from the results that J48 using reduced-error pruning had the best overall performance, with higher sensitivity and balanced accuracy than LMT and Random Forest. LMT, though more specific, could not generalize as well as J48. Random Forest, though powerful on the training set, indicated overfitting with a higher test error rate.

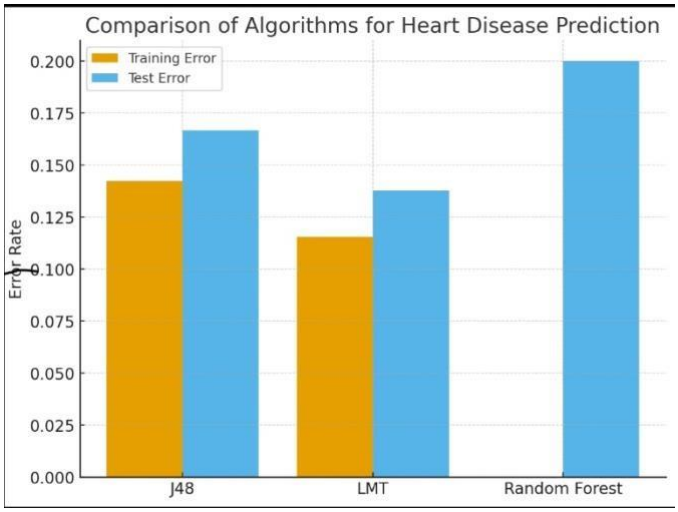

Figure 2: Comparison of training and test errors for J48, Logistic Model Tree,

and Random Forest classifiers

J48 also generated more shallow models through pruning, with fewer trees. LMT resulted in smaller trees but took longer to construct, possibly constraining its scalability. Random Forest, though very robust, took several trees and had lower accuracy with this dataset.
In general, J48 stands out as the most consistent decision tree– based classifier of the three, being most simple, efficient, and accurate.[2]

VI. CONCLUTION

Experimental analysis indicates that the J48 decision tree algorithm with reduced-error pruning is the best-fitting model for heart disease prediction among the classifiers evaluated. It had the highest accuracy (56.76%) and the briefest modelbuilding time of 0.04 seconds, making it both efficient and effective. The Logistic Model Tree (LMT) algorithm, on the other hand, had the lowest accuracy (55.77%) and took additional time (0.39 seconds) to build its model. The Random Forest algorithm, although very flexible and accurate on training data, proved to generalize poorly with a test error of 0.20. These results affirm that methods based on decision trees can be applied usefully in medical prediction problems, especially when interpretability and efficiency are critical. Nevertheless, the results indicate that existing models still achieve only modest gains in predictive accuracy. This identifies the necessity of hybrid and more complex algorithms to enhance nearly heart disease detection, thus enabling improved healthcare results.

VII. FUTURE WORK

Even though this study proves the effectiveness of decision tree algorithms for heart disease prediction, future work has many ways of increasing precision and scalability: Exploring other discretization techniques: Future experimentation can use various methods like Information Gain, Gain Ratio, and Gini Index, in

Algorithm	Training error	Test error
J48	0.1423	0.1666
Logistic model tree	0.1657	0.2379
Random forest	0.0000	0.2000

conjunction with voting ensembles, for better classification results. For example, Equal Frequency Discretization with Gain Ratio Decision Trees and several voting systems may be investigated to enhance diagnostic accuracy. Hybrid model development: A fusion of classifiers—e.g., Support Vector Machines, Logistic Regression, and Decision Trees—can produce more robust predictive systems, especially when handling data imbalance challenges. Unsupervised learning methods: Methods like Association Rule Mining, Clustering, and K-means can be used to uncover hidden patterns and reduce prediction systems. Multivariate decision tree utilization: Utilizing multivariate splitting methods on small and large data sets can enhance accuracy and efficiency by retaining more sophisticated patterns. Applying these directions, future research can develop more robust, scalable, and interpretable models that make available to healthcare professionals more solid decision support for near heart disease diagnosis.[5][8]

REFERENCES:

[1] C. S. Dangare and S. S. Apte, "Improved study of heart disease prediction system using data mining classification techniques," International Journal of Computer Applications, vol. 47, no. 10, pp. 44–48, 2012. [2] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," pp.
108–115, 2008.
[3] Y. E. Shao, C.-D. Hou, and C.-C. Chiu, "Hybrid intelligent modelling schemes for heart disease classification," Applied Soft Computing, Vol. 14, pp. 47–52, 2014.
[3] M. Shouman, T. Turner, and R. Stocker, "Using data mining techniques in heart disease diagnosis and treatment," pp. 173–177, 2012.;3
[4] P. V. Ankur Makwana, "Identify the patients at high risk of re-admission in hospital in the next year," International Journal of Science and Research, vol. 4, pp. 2431– 2434, 2015.
[5] J. Nahar, T. Imam, K. S. Tickle, and Y.-P. P. Chen, "Computational intelligence for heart disease diagnosis: A medical Knowledge driven approach," Expert Systems with Applications, vol. 40, no. 1, pp. 96–104,2013. [7] Y. Xing, J. Wang, Z. Zhao, and Y. Gao, "Combination data mining methods with new medical data to predicting outcome of coronary heart disease," pp. 868–872, 2007.
[8] Combination data mining methods with new medical data to predicting outcome of coronary heart disease,"
in Convergence Information Technology, 2007. International Conference on. IEEE, 2007, pp. 868–872.
[8] Y. E. Shao, C.-D. Hou, and C.-C. Chiu, "Hybrid intelligent modelling, schemes for heart disease classification," Applied Soft Computing, vol. 14, pp. 47–52, 2014.