Journal of Theoretical and Applied Information Technology

30th September 2023. Vol.101. No 18

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

COMPARATIVE EVALUATION OF CARDIOVASCULAR DISEASE USING MLR AND RF ALGORITHM WITH SEMANTIC EQUIVALENCE

VINSTON RAJA R [1*], DEEPAK KUMAR A [2+], PRABU SANKAR N [3+], CHIDAMBARATHANU K [4+], THAMARAI I [5+], KRISHNARAJ M [6+], IRIN SHERLY S [7+]

[1*]Assistant Professor, Information Technology, Panimalar Engineering College, Chennai, India.
[2+]Assistant Professor, Computer Science and Engineering, St. Joseph's Institute of Technology, Chennai, India.
[3+]Assistant Professor, Department of Information Technology, Panimalar Engineering College, Chennai, India.
[4+]Professor, Computer Science and Business Systems, R.M.K. Engineering College, RSM Nagar, Kavaraipettai
[5+]Associate Professor, Computer Science and Engineering, Panimalar Engineering College Chennai City Campus
[6+]Assistant Professor, Information Technology, Panimalar Engineering College, Chennai, India.
[7+]Assistant Professor, Information Technology, Panimalar Engineering College, Chennai, India.

[1*]rvinstonraja@gmail.com, [2,+]deepakkumar@stjosephstechnology.ac.in, [3, +] n.prabusankar81@gmail.com, [4, +]kct.it@rmkec.ac.in, [5,+]thamarai.panimalar@gmail.com, [6+]monykrishnaraj@gmail.com, [7+]irinsherly.pit@gmail.com

ABSTRACT

Coronary artery disease is a highly intricate medical condition that affects a significant portion of the global population. It is even being referred to as a silent killer because it results in the death of a person with no obvious symptoms. The timely and accurate detection of heart disease is crucial in the healthcare industry, especially within the cardiology domain, as it enables effective treatment and management of the condition. Based on Machine learning techniques, an accurate and efficient model will be created to diagnosis heart disease. The machine learning models for classification will be developed using Multiple Linear Regression Algorithm and Random Forest Algorithm. Heart datasets were obtained from five countries: Cleveland, Hungary, Swiz, LongBeach, and Statlog, and datasets were analyzed using the Random Forest algorithm, KNN, Naive Bayes, SVM Algorithm and Multiple Linear Regression Algorithm to extract an intelligent pattern for forecasting the risk of heart disease. The accuracy of the MLR and RF models will be tested, and the best model will be deployed in health-care settings to diagnose cardiac disease.

Keywords: *Machine Learning*, *Random Forest (RF), Multiple Linear Regressions (MLR), Data Sampling*

1. INTRODUCTION

Coronary disease is a prevalent ailment that accounts for a significant number of fatalities across the globe. Cardiovascular disease (CVD) is a pathological state characterized by the heart's incapacity to adequately circulate blood to the body's diverse organs, ultimately leading to a cardiovascular collapse. [1-4]. the excellent justification for cardiovascular breakdown is coronary course blockage. During the previous ten years, coronary illness was the leading cause of death on the planet. It includes diseases of the heart muscles, valves, conduction framework, respiratory failure, and other organs. Among the remaining types of heart infections, myocardial dead tissue or coronary episode is the most serious. Heart disease is now found in all socioeconomic classes, as opposed to being a disease of the upper crust.

This paper concentrated and focused on the heart attack. Regardless, respiratory disappointment is included as a

tranquil killer that causes the death of a person without obvious signs. Attempts are being made to anticipate the possibility of this risky disease in the

meantime [15-16]. There are a variety of gadgets and techniques that are regularly tested to meet today's prosperity requirements. Artificial intelligence methods can help in this situation [6-8]. Regardless of the fact that coronary disappointment can occur in various structures, there is a standard course of action of focus hazard factors that influence whether someone will, eventually, be in danger for respiratory disappointment or not.

Early detection and effective management of cardiovascular disease can significantly reduce mortality rates, as prevention is better than cure. Myocardial infarction is commonly known as a heart attack, where damage to the heart muscle occurs due to a blockage in the blood supply to a particular area of the heart. Chest discomfort is a common symptom that can also impact the shoulder, neck, back, or jaw regions, and is often experienced as a sensation of pressure or tightness in the central or left chest area. Authorities aim to discuss a feasible approach for prompt identification of cardiovascular pathology.

Big data is not only about the magnitude of the data but also about its ability to uncover meaningful insights from complex, unstructured, diverse, time-sensitive, and large-scale datasets. Nonetheless, collecting, archiving, querying, distributing, and analyzing data poses significant difficulties. Additionally, there is an increasing trend in utilizing diverse forms of digital social communication. [26]. Data mining is currently a main tool for identifying information in hidden forms among large datas, having established itself as a novel field. This unknown knowledge in the health-care industry can be applied to a variety of application fields, such as heart attack prediction. Employing data analytics and data management techniques can enable the development of novel applications that support medical practitioners and other stakeholders in the healthcare industry to make timely decisions pertaining to heart attack diagnoses in the initial stages [27, 28].

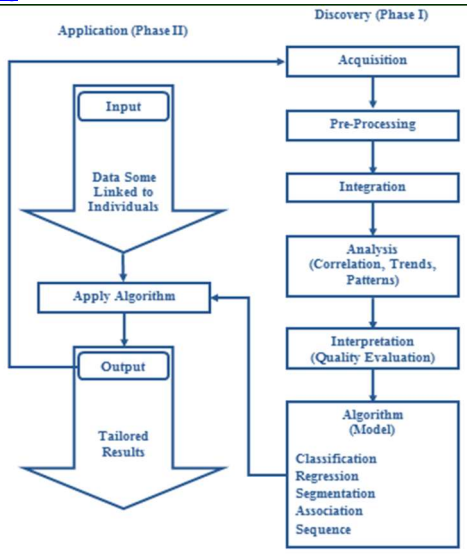

Figure -1 Process of Data Analytics

The data analytics process can be broadly categorized into two approaches, as depicted in Figure-1, which outlines the fundamental steps involved in data analytics [29]. The knowledge and study obtain health care is evolving as a result of big data analytics. This technology is being used by providers more than deliver a more tailored method to wellbeing system.

2. RELATED WORKS

In the study described in reference [1], the author leveraged data analysis algorithms to evaluate the incidence of cardiac ailments. The data was obtained from patients and was weighted based on its contribution to the overall success rate. His method for determining weight coefficient is proposed. Jian Ping Li et al. [2], on the other hand, secured an Intelligent E-Healthcare system and developed his model using Machine Learning Classifiers. In this study, the author discusses more than five algorithms and introduces a novel method called FCMIM to handle the problem of detecting heart illness.

Seyedamin Pouriyeh and colleagues [3] performed a comprehensive study that involved examining and contrasting multiple artificial intelligence methods for identifying cardiovascular pathology. Dakun Lai et al. [5] analysis various parameters of patients and did detail study of all major machine algorithms to find risk and challenges. The research work of [6] used a hybrid model with combining

neural network concept and fuzzy method with an intelligence concept to achieve 87.4 percent of accuracy. Zameer Khan et al.[7] used more than 10 patient factors to develop multiple categorization models. His work includes an overview of the existing algorithm as well as a synopsis of the previous work.S. Nayak et al.[8] underline the importance of early disease diagnosis using mining classification approaches, as well as disease protection at an early phase so that illnesses can be cured and prevented. In their research, Gudadhe and co-authors [9] constructed an architecture utilizing the MLP network and SVM algorithm, which resulted in an 86.41 percent accuracy rate in differentiating between two categories. On the basis of one patient outcome data, Yiwen Meng et al [10] evaluated the random forest classifier with the HMM model for predicting heart disease.

Geweid et al.[11] used an improved SVM-based duality optimization technique to build HD identification techniques. For a better thoughtful of our recommended methodology, the limitations and advantages of the recommended HD diagnosis approaches have been summarized in the past works Therefore, to overcome the challenges associated with cardiovascular pathology diagnosis and develop a non-invasive analysis system, this study employed an expert judgment approach based on Machine Learning (ML) classifiers and artificial fuzzy logic. The outcome of this study was a reduction in mortality rates [12-13].

3. PROPOSED METHODOLOGY

Cardiovascular disease is a significant global health issue, and timely prediction is crucial for appropriate prevention and intervention measures. Machine learning techniques have found widespread applications in the healthcare domain, particularly in the prediction and diagnosis of various medical conditions, including but not limited to cardiovascular diseases. In this study, the performance of five different machine learning algorithms, namely k-nearest neighbor, naive Bayes classifier, support vector machine (SVM), multiple linear regression (MLR), and random forest (RF), were evaluated for predicting cardiovascular pathology. The dataset used in the study is the Cleveland heart disease dataset, comprising 303 instances and 14 features such as age, gender, blood pressure, and cholesterol levels. The dataset was originally obtained from the Cleveland Clinic Foundation and subsequently donated to the University of California, Irvine, where it has been extensively employed in studies related to cardiovascular disease prediction.

For the purposes of this investigation, the dataset was segregated into two distinct sets: training and testing, with an

80:20 ratio. The training set was utilized to teach the machine learning algorithms, while the testing set was utilized to evaluate their efficacy. The dataset was preprocessed before training the algorithms. To ensure data integrity and consistency, any incomplete values in the dataset were replaced with the average value of the corresponding feature. Additionally, the categorical features were transformed into binary dummy variables.

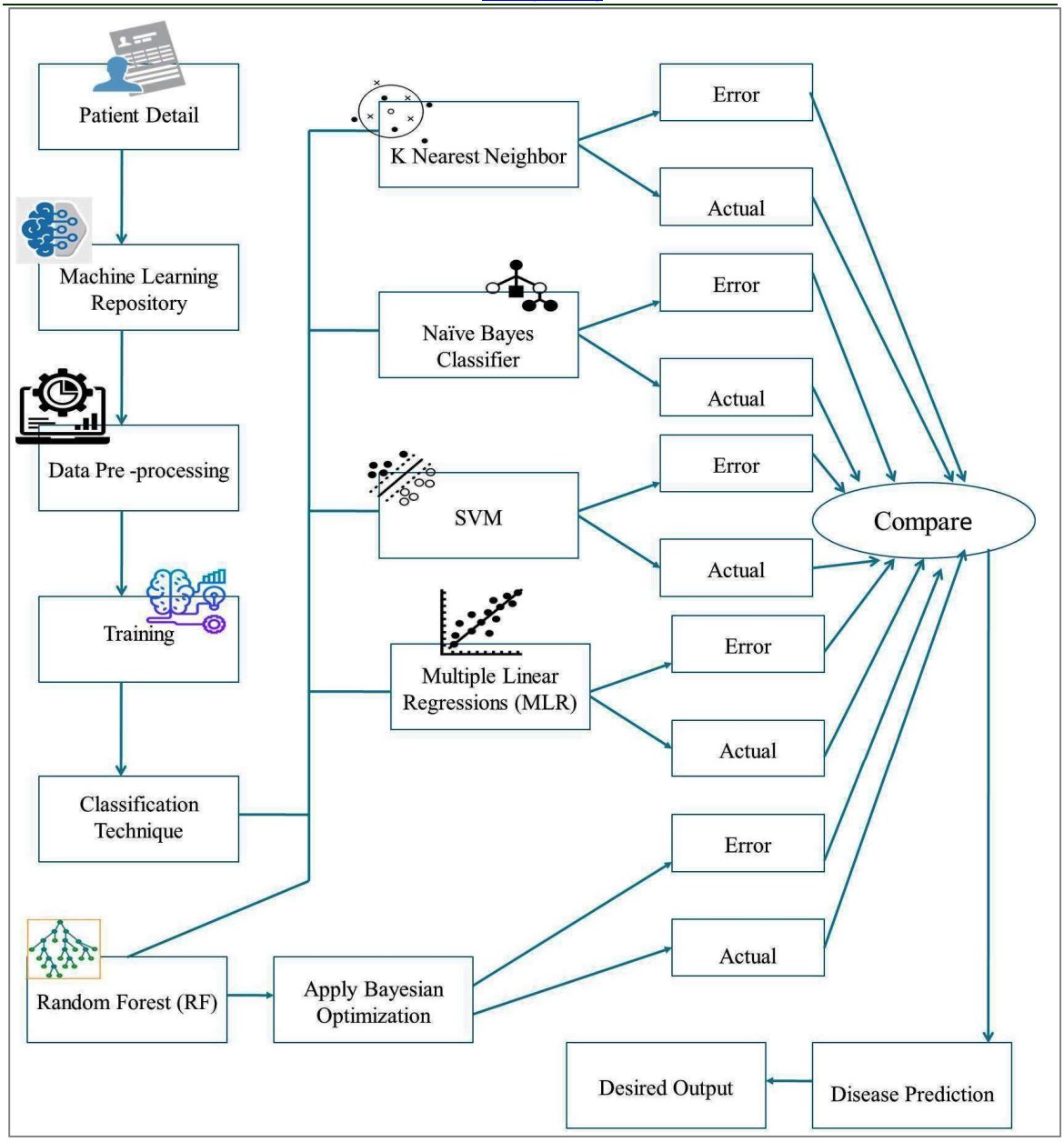

Figure-2 Architecture model of the heart disease diagnosis technique

The k-nearest neighbor (KNN) algorithm is a non-parametric machine learning technique utilized for both classification and regression tasks. The KNN algorithm functions by identifying k number of nearest neighbors to a new data point and classifying it based on the majority of the neighbors. The KNN algorithm is easy to implement and can handle both binary and multi-class classification problems. In this study, the KNN algorithm was utilized to predict the presence or absence of heart disease by leveraging the Cleveland heart disease dataset.

The Naive Bayes Classifier is a frequently used machine learning classification algorithm that applies Bayes' theorem to approximate the probability of a hypothesis (class) based on observable evidence (features). The term "naive" stems from its assumption that the features are independent, even though in actuality, they might

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

be interrelated. The classifier computes the likelihood of each feature given the class, and the prior probability of the class, subsequently leveraging them to deduce the most probable class for a given set of features. This assumption simplifies the computation needed to approximate the probabilities, resulting in a highly efficient algorithm.

A. Data Preparation and Preprocessing

The data for the heart disease prediction project can be made available in a variety of file formats. Once the data is obtained, it needs to be loaded into the R environment for further analysis. One common way of storing data in R is to use a data frame, which can hold different types of data (e.g., numeric, categorical, textual) and can be manipulated using R's built-in functions. In order to use the Random Forest classification algorithm, it is necessary to install and load the Random Forest package in the R environment. Data pre-processing is a crucial component of the data analysis process, which encompasses various techniques for refining, reshaping, and standardizing raw data into a suitable format that can be further processed for analysis. Common problems in raw data include noise, missing values, and uncertainty, which can be addressed using various statistical techniques. Dealing with missing values is a crucial aspect of data preparation and can be addressed by utilizing imputation methods like mean imputation, mode imputation, or regression imputation. In this project, the data was preprocessed to handle missing values, normalize the data, and prepare it for use in the classification algorithms.

B. Multivariate Regression for Heart Disease Prediction

Regression analysis is a statistical approach to investigate the relationship between one or more independent variables and a dependent variable. Its application is widespread in forecasting future outcomes based on historical data. Simple regression and multiple regression are the two main types of regression analysis used.

Multiple linear regression, a type of regression analysis, assumes that the independent variables are not strongly correlated with each other, the observations are independent, and the dependent and independent variables have a linear relationship. Moreover, the variance of the residuals should be constant for accurate prediction.

C. Using Random Forest Algorithm for Prediction

The Random Forest algorithm is a versatile method that can be used for both regression and classification problems. Developed by Leo Breiman and Adele Cutler in 2001, the algorithm generates a large number of decision trees. As shown in Figure-3, each tree in the Random Forest is constructed independently using a random sample of the data.

During the prediction phase, a new data point is passed through each tree in the forest, producing multiple outcomes. For classification problems, the Random Forest algorithm selects the most common class predicted by the trees, while for regression problems, the algorithm takes the mean of the outputs generated by each tree as the predicted value. This approach helps to reduce over fitting and improve the accuracy of the model.

Machine learning algorithms require setting certain parameters and hyper parameters to optimally fit the model to the data. Parameters are learned during training of the model and reflect the relationship between the input features and output variable. On the other hand, hyper parameters are pre-defined settings that govern how the model learns the parameters from the data. In some algorithms, such as linear and logistic regression, the parameters are the coefficients and biases, while the hyper parameters include regularization terms and learning rates.

The process of finding the optimal hyper parameters for a model is known as hyper parameter tuning, and it involves experimenting with different combinations of hyper parameters to find the best settings that yield the highest performance on a validation set. This process is crucial to ensure that the model is not over fitting or

under fitting the data, and to achieve the highest accuracy possible.

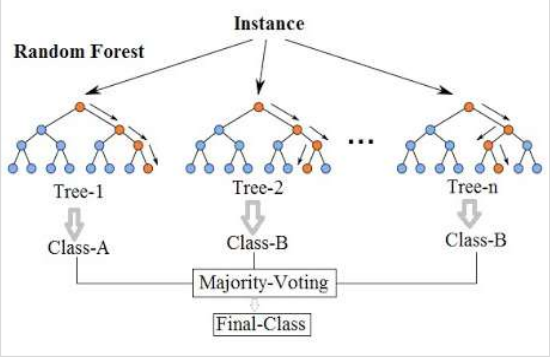

Figure-3 Tree formation by Random Forest

To achieve the best possible accuracy, it is crucial to fine-tune the most significant hyper parameters of the decision tree model illustrated in Figure 3.

Table 1: Performance Evaluation Metrics of the Tuned Decision Tree Model Using Optimized Hyper parameters

Max depth	Min samples split	Min samples leaf	Max features	Criterion	Accuracy
10	18	15	Auto	Entropy	81%
15	25	12	Auto	Gini	82%
18	11	10	Sqrt	Gini	72%
20	15	20	Sqrt	Gini	83%
25	10	50	Auto	Entropy	71%
30	12	30	Auto	Entropy	74%
35	22	25	Sqrt	Entropy	75%
40	8	14	Sqrt	Gini	78%
45	5	16	Auto	Entropy	73%
50	14	0	Auto	Gini	78%
70	17	18	Auto	Entropy	75%
80	13	0	Auto	Entropy	84%
100	20	0	Sqrt	Entropy	84%

The hyper parameters are optimized by systematically varying their values and testing the model's performance on the test dataset. Careful evaluation is necessary to prevent over fitting, which occurs when the model performs well on the training data but poorly on the test data. After tuning the hyper parameters, the model's performance is assessed on the test data to obtain the final performance metrics. Table 3 shows the results of the decision tree model after hyper parameter tuning. Although different combinations of hyper parameters may lead to varying outcomes, we only report the combinations that yield the highest accuracy.

Random Forest with Bayesian Optimization:

Bayesian Optimization is a statistical technique that aims to minimize a given objective function. Its primary objective is to find the input value(s) that generate the optimal output value. Compared to other optimization methods, it can produce better results, reduce optimization time, and improve performance during testing. The Hyperopt package is a Python library that provides an implementation of Bayesian Optimization, and it requires three primary parameters to the fmin function:

1. Bayesian Optimization for Function Minimization
   - Probability-based approach to finding the minimum of a function
   - Objective is to identify input values that produce the lowest output value
   - Outperforms other methods in terms of testing performance and optimization time
2. Implementation of Bayesian Optimization using Hyperopt Library in Python
   - Function fmin has three main parameters
   - Bayesian Optimization is a method for minimizing a given function by probabilistically determining the input value that produces the lowest output value. It outperforms other methods, resulting in enhanced testing performance and decreased optimization time.
   - The Hyperopt package in Python implements Bayesian Optimization, which requires the specification of three main parameters: the objective function that defines the loss to be minimized, the domain space that specifies the range of input values to test, and the optimization algorithm that determines the best input value through search.
3. Components of the Domain Space Parameter
   - Range of input values used for optimization
   - Creates a probability distribution for each hyper parameter used in Bayesian Optimization
4. Optimization Algorithm Options
   - Random Search: Simple search algorithm that randomly selects input values
   - TPE (Tree-structured Parzen Estimator): Sequential search algorithm that uses Bayesian probability model to select input values
   - Adaptive TPE: Adaptive version of TPE algorithm that dynamically adjusts the probability model based on previous evaluations.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

```
SMBO(f, M_0, T, S)
1   H ← ∅,
2   For t ← 1 to T,
3       x* ← argmin_x S(x, M_{t−1}),
4       Evaluate f(x*),        ▷ Expensive step
5       H ← H ∪ (x*, f(x*)),
6       Fit a new model M_t to H.
7   return H
```

Figure 4: SMBO of pseudo-code

Sequential Model-Based Optimization is also known as Bayesian Optimization (Figure-4), which is a technique used to minimize the objective function by approximating it with a surrogate model. This method is sequential in nature because hyper parameters are added one by one to update the surrogate model. The surrogate function is cheaper to evaluate as compared to the true objective function.

The following terminologies are used in SMBO:

- Observation History (H): A record of (hyper parameter, score) pairs observed so far.
- Max Number of Iterations (T): The maximum number of iterations allowed to find the best hyper parameters.
- True Objective Function (f): A function to minimize (e.g., RMSE function).
- Surrogate Function (M): An approximation of the true objective function, updated with each new observation.
- Acquisition Function (S): A function that guides the search for the next set of hyper parameters to evaluate.
- Next Chosen Hyper parameter to Evaluate (x*): The set of hyper parameters selected for evaluation in the next iteration.

1. Start by initializing a surrogate model and an acquisition function.
2. During each iteration, determine the hyper parameter x* that maximizes the acquisition function. The acquisition function is built using the surrogate model rather than the true objective function, which is explained further later on. Note that in the provided pseudo-code, x* is obtained when the acquisition function is minimized, but this depends on the specific definition of the acquisition function being used. For the commonly used Expected Improvement function, the goal is to maximize it.
3. Get the objective function score for x* to evaluate its performance.
4. Add the (hyper parameter x*, objective function score) to the history of previous samples.
5. Update the surrogate model using the most recent set of evaluated hyper parameters and their corresponding objective function scores.
6. Repeat this process until the maximum number of iterations is reached. Finally, return the history of (hyper parameter, true objective function score) pairs. It should be noted that the last recorded pair may not necessarily correspond to the best achieved score. Therefore, the pairs should be sorted to determine the optimal hyper parameters.

4. EXPERIMENTAL RESULTS & INVESTIGATION.

The hyper parameters of the random forest model were exhaustively searched and fine-tuned, including but not limited to N estimators, max depth, min sample split, min sample leaf, and max features.

The optimized random forest model was evaluated and its performance is presented in Table 2. The table displays the various combinations of hyper parameters that were explored, including criterion, max depth, max features, N estimators, and min sample leaf. From the results, it is observed that the optimal accuracy of 87% was achieved by setting the hyper parameters as criterion=Gini, max depth=50, max features=Auto, and N estimators=100. It should be noted that different combinations of hyper parameters can result in

varying performance metrics, and the reported values are based on the best performing configuration. The model's true positive rate and true negative rate are 87% and 84%, respectively, and the model achieves an accuracy of 86%. The precision of the model is also 86%, and the misclassification rate is 13%. Furthermore, the model achieves an AUROC score of 86%, indicating its high performance in distinguishing between positive and negative cases. The results of the optimized random forest model's experiments illustrate how crucial it is to tune the hyper parameters for achieving optimal performance.

Table 2 - An Empirical Evaluation of the Optimized Random Forest Model: Impact of Hyper Parameter Tuning on Model Performance.

Criterion	Max depth	Max features	N-Estimators	Min samples leaf	Accuracy
Gini	70	0	0	0	85%
Entropy	60	Auto	0	0	86%
Gini	50	Auto	100	0	87%
Entropy	80	Auto	100	100	73%
Gini	100	Auto	100	50	76%
Entropy	30	0	80	60	80%
Gini	40	0	90	40	78%
Gini	25	Auto	70	30	75%
Entropy	20	Auto	40	25	82%
Entropy	35	Auto	30	20	81%
Gini	45	0	60	35	80%

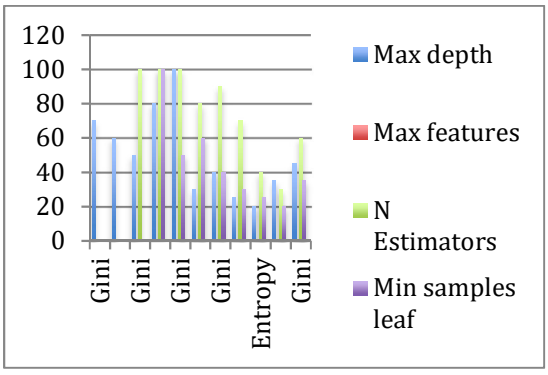

Figure-4 "Visualization of Performance Metrics for Optimized Random Forest Model"

Figure-4 graphical representation of the experimental results for the optimized random forest model is shown in Figure 1. It can be observed that the highest accuracy of 87% was achieved with the hyper parameter combination of criterion = Gini, max depth = 50, max features = auto, and N estimators = 100. The performance metrics for this combination were also the most favorable, with a true positive rate of 87%, a true negative rate of 84%, and a precision and accuracy of 86%. The experimental results indicate that the accuracy of the model is highly dependent on the choice of hyper parameters, as there is a significant variation in the performance across different combinations. In fact, some of the combinations resulted in accuracy as low as 73%, highlighting the importance of carefully selecting the appropriate hyper parameter values to achieve optimal performance. Overall, the results demonstrate the importance of hyper parameter tuning in optimizing the performance of a random forest model.

Table 3 - *Shows the experimental results of the MSE , RMSE and Accuracy*

Algorithm	MSE	RMSE	Accuracy
k-Nearest Neighbor	0.180	0.424	0.803
Naive Bayes Classifier	0.230	0.480	0.787
Support Vector Machine	0.150	0.387	0.820
Multiple Linear Regression	0.127	0.357	0.847
Random Forest	0.120	0.346	0.860

Table 3 presents the comparative evaluation of five distinct machine learning algorithms on a specific dataset, measured by their performance metrics such as mean squared error (MSE), root mean squared error (RMSE), and accuracy. Among the algorithms, Random Forest shows the best performance with the lowest MSE of 0.120, the lowest RMSE of 0.346, and the highest accuracy of 0.860. The next best algorithm is Multiple Linear Regression, with an MSE of 0.127, an RMSE of 0.357, and an accuracy of 0.847. Support Vector Machine also performs well, with an MSE of 0.150, an RMSE of 0.387, and an accuracy of 0.820. Naive Bayes Classifier and k-Nearest Neighbor show relatively lower performance, with an accuracy of 0.787 and 0.803, respectively. Figure-5, These results suggest that Random Forest and Multiple Linear Regression are promising algorithms for this dataset, while further optimization and tuning may be required for Naive

Bayes Classifier and k-Nearest Neighbor to achieve better performance.

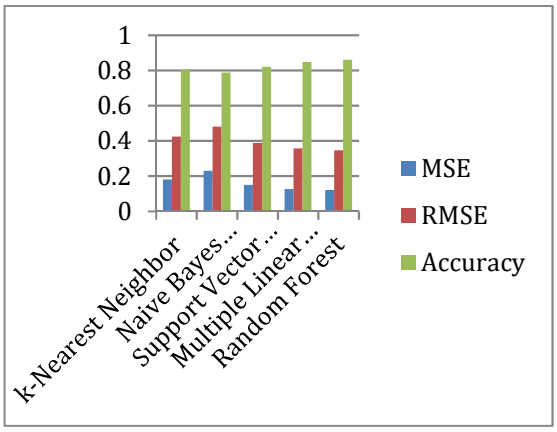

Figure-5 Experimental results of the experimental results of the MSE, RMSE and Accuracy

Table 4: Performance Evaluation Metrics of the Developed Optimized Models for Heart Disease Classification

	Performance measures of the models					
Models	TPR	TNR	Accuracy	Precision	Error rate	AUROC
Naive Bayes	83%	80%	82%	82%	15%	82%
K-NN	87%	81%	84%	83%	15%	85%
SVM	80%	82%	82%	86%	18%	82%
Multiple Linear recursion	84%	82%	82%	84%	14%	85%
Random forest	87%	84%	87%	86%	13%	87%

Table 4 displays the evaluation metrics of the optimized heart disease models, which were assessed using several performance measures, including accuracy, precision, true positive rate (TPR), true negative rate (TNR), error rate, and area under the receiver operating characteristic curve (AUROC). Among the models, the random forest model performed the best with TPR of 87%, TNR of 84%, accuracy of 87%, precision of 86%, error rate of 13%, and AUROC of 87%. The k-Nearest Neighbor (K-NN) model also performed well with TPR of 87%, TNR of 81%, accuracy of 84%, precision of 83%, error rate of 15%, and AUROC of 85%. The Multiple Linear Regression model had a TPR of 84%, TNR of 82%, accuracy of 82%, precision of 84%, error rate of 14%, and AUROC of 85%. The Naive Bayes Classifier and Support Vector Machine (SVM) models had comparable performance with TPR of 83% and 80%, TNR of 80% and 82%, accuracy of 82% and 82%, precision of 82% and 86%, error rate of 5% and 18%, and AUROC of 82% and 82%, respectively. In general, the random forest algorithm demonstrated superior accuracy and resilience compared to the other models evaluated.

The evaluation of various machine learning models in predicting heart disease has shown promising outcomes. The Random Forest, k-Nearest Neighbor, and Multiple Linear Regression models displayed superior performance in comparison to the Naive Bayes Classifier and Support Vector Machine models, in terms of precision, accuracy, and AUROC score. Among these models, the Random Forest algorithm showed the highest precision of 86%, an accuracy of 87%, and an AUROC score of 87%.

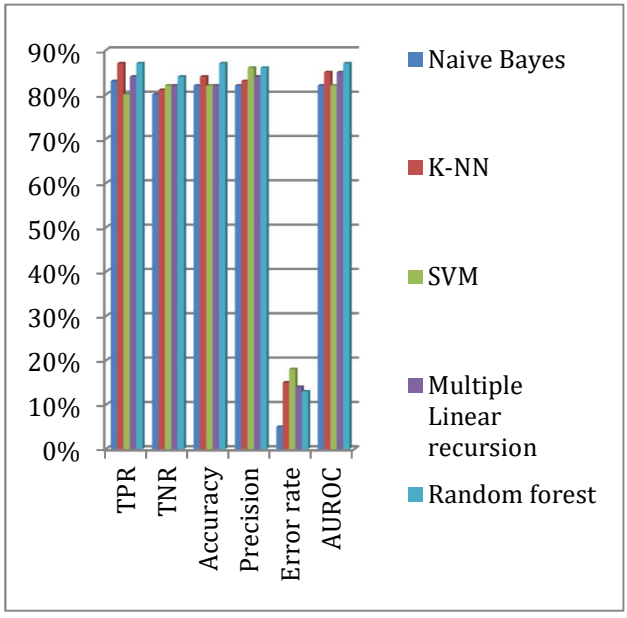

Figure-6: Performance Measures of the Developed Optimized

Heart Disease Models Evaluated with High-Level Metrics.

However, it's important to note that performance measures like TPR, TNR, and error rate are also important indicators of the models' performance. Overall, these results provide valuable insights into the potential use of machine learning models in predicting heart disease and can guide future research in this area.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

5. CONCLUSION AND FUTURE ENHANCEMENT

In this study, Machine Learning techniques were employed to predict the likelihood of heart disease development in patients. Through hyper parameter tuning, the Random Forest and Multiple Linear Regression algorithms were compared for their efficacy. The outcomes indicated that the Random Forest model achieved the highest accuracy score of 87%, while the Multiple Linear Regression model demonstrated an accuracy of 85%. The Random Forest algorithm has the potential to assist physicians in predicting cardiac illness and improving medical care. Future research can enhance the findings by integrating additional machine learning algorithms and deep learning techniques. Furthermore, the dataset will be validated using deep learning methods, and the algorithm's accuracy will be evaluated. This approach has the potential to identify the appropriate algorithm for diagnosing heart disease in patients.

REFERENCES

[1] Alperen Erdogan, Selda Guney, “Heart Disease Prediction by Using Machine Learning Algorithms", IEEE Signal Processing and Communications Applications Conference, 2020.

[2] Jian Ping Li, Amin Ul Haq, Salah Ud Din, Jalaluddin Khan,” Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare”, IEEE Access, Volume: 8, 2020, pp: 107562 – 107582

[3] Seyedamin Pouriyeh, “A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease ", IEEE International Conference on Computers and Communications, 2017

[4] An Automated Strategy for Early Risk Identification of Sudden Cardiac Death by Using Machine Learning Approach on Measurable Arrhythmic Risk Markers”, Dakun Lai; Yifei Zhang; Xinshu Zhang; Ye Su; Md Belal Bin Heyat, IEEE Access, Year: 2019, Vol. 7.

[5] Samrat Kumar Dey, Ashraf Hossain, "Implementation of a Web Application to Predict Diabetes Disease: An Approach Using Machine Learning Algorithm", International Conference of Computer and Information Technology, Dec 2018.

[6] Zameer Khan, et.al., "Empirical Study of Various Classification Techniques for Heart Disease Prediction", IEEE International Conference on Computing Communication and Automation, 2020.

[7] S. Nayak, M. K. Gourisaria, M. Pandey and S. S. Rautaray, "Prediction of Heart Disease by Mining Frequent Items and Classification Techniques," 2019 International Conference on Intelligent Computing and Control Systems (ICCS), 2019.

[8] M. Gudadhe, K. Wankhade, and S. Dongre, “Decision support system for heart disease based on support vector machine and artificial neural network,” IEEE International Conference on Computer and Communication Technology, 2010, pp. 741–745.

[9] Yiwen Meng, William Speier, “A Machine Learning Approach to Classifying Self-Reported Health Status in a Cohort of Patients With Heart Disease Using Activity Tracker Data”, IEEE Journal of Biomedical and Health Informatics, Volume: 24, Issue: 3, March 2020.

[10] G. N. Geweid and M. A. Abdallah, ``A new automatic identication method of heart failure using improved support vector machine based on duality optimization technique," IEEE Access, vol. 7, pp. 149595149611, 2019.

[11] U. Haq, J. Li, M. H. Memon, J. Khan, and S. U. Din, A novel integrated diagnosis method for breast cancer detection, J. Intell. Fuzzy Syst., vol. 38, no. 2, pp. 2383-2398, 2020.

[12] S. Nazir, S. Shahzad, S. Mahfooz, and M. Nazir, ``Fuzzy logic based decision support system for component security evaluation," Int. Arab J. Inf. Technol., vol. 15, no. 2, pp. 224231, 2018.

[13] Balakrishnan, S., Syed Muzamil Basha, & Ravi Kumar Poluru., 2019. Heart Disease Prediction Using Machine Learning Algorithm. International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue- 10.

[14] Rajathi, N., Kanagaraj, S., Brahmanambika, R. and Manjubarkavi, K., 2018. Early detection of

dengue using machine learning algorithms. International Journal of Pure and Applied Mathematics, 118(18), pp.3881-3887.

[15] Raju, C., Philipsy, E., Chacko, S., Suresh, L.P. and Rajan, S.D., 2018, March. A Survey on Predicting Heart Disease using Data Mining Techniques. In 2018 Conference on Emerging Devices and Smart Systems (ICEDSS) (pp. 253-255). IEEE.

[16] Thomas, J. and Princy, R.T., 2016, March. Human heart disease prediction system using data mining techniques. In 2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT) (pp. 1-5). IEEE.

[17] Tikotikar, A., & Kodabagi, M., 2017. A survey on technique for prediction of disease in medical data. In 2017 International Conference on Smart Technologies for Smart Nation (Smart Tech Con) (pp. 550-555). IEEE

[18] Jamgade, A.C. and Zade, S.D., 2019. Disease prediction using machine learning. International Research Journal of Engineering and Technology, 6(5), pp.6937-6938.

[19] Prasad, R., Anjali, P., Adil, S. and Deepa, N., 2019. Heart disease prediction using logistic regression algorithm using machine learning. International journal of Engineering and Advanced Technology, 8, pp.659-662.

[20] A. U. Haq, J. Li, J. Khan, M. H. Memon, S. Parveen, M. F. Raji, W. Akbar, T. Ahmad, S. Ullah, L. Shoista, and H. N. Monday, ‘‘Identifying the predictive capability of machine learning classifiers for designing heartdisease detection system,’’ in Proc. 16th Int. Comput. Conf. Wavelet Act.Media Technol. Inf. Process., Dec. 2019, pp. 130–138..

[21] A. Ul Haq, J. Li, Z. Ali, J. Khan, M. H. Memon, M. Abbas, and S. Nazir, ‘‘Recognition of the Parkinson’s disease using a hybrid featureselection approach,’’ J. Intell. Fuzzy Syst., vol. 39, pp. 1–21, May 2020, doi: 10.3233/JIFS-200075.

[22] A. U. Haq, J. P. Li, J. Khan, M. H. Memon, S. Nazir, S. Ahmad, G. A. Khan, and A. Ali, ‘‘Intelligent machine learning approach for effective recognition of diabetes in E-healthcare using clinical data,’’ Sensors, vol. 20, no. 9, p. 2649, May 2020.

[23] A. U. Haq, J. Li, M. H. Memon, J. Khan, S. U. Din, I. Ahad, R. Sun, and Z. Lai, ‘‘Comparative analysis of the classification performance of machine learning classifiers and deep neural network classifier for prediction of parkinson disease,’’ in Proc. 15th Int. Comput. Conf. Wavelet Act.Media Technol. Inf. Process. (ICCWAMTIP), Dec. 2018, pp. 101–106.

[24] A. U. Haq, J. P. Li, M. H. Memon, J. Khan, A. Malik, T. Ahmad, A. Ali, S. Nazir, I. Ahad, and M. Shahid, ‘‘Feature selection based on L1-norm support vector machine and effective recognition system for Parkinson’sdisease using voice recordings,’’ IEEE Access, vol. 7, pp. 37718–37734, 2019.

[25] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, ‘‘A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms,’’ Mobile Inf. Syst., vol. 2018, pp. 1–21, Dec. 2018

[26] Sun J, Reddy CK (2013) Big data analytics for healthcare. Tutorial presentation at the SIAM International Conference on Data Mining, Austin, TX.

[27] Ghadge P, Girme V, Kokane K, Deshmukh P (2015) Intelligent heart attack prediction system using big data. International Journal of Recent Research in Mathematics Computer Science and Information Technology 2: 73-77. Ghadge P, Girmev V, Deshmukh P, Kokane K (2016)

[28] Intelligent Heart attack prediction system using big data. International Journal of Advanced Research in Computer and Communication Engineering 5: 723-725.

[29] Center for Information Policy Leadership, Hunton and Williams LLP (2013) Big Data and Analytics: Seeking Foundations for Effective Privacy Guidance. A Discussion Document, pp: 1-16.