Heart Disease Risk Prediction System using Machine Learning

Mariyam Tariq
Deparment of Computer Science & Engineering
Amity University Uttar Pradesh, India

Syed Wajahat Abbas Rizvi
Deparment of Computer Science & Engineering
Amity University Uttar Pradesh, India

ABSTRACT
In today's world, accurately estimating the risk of heart disease in patients is a very critical problem. It's also crucial to detect the risk of heart disease as it reduces death tolls. If the heart disease risk is predicted in its early stage, then with proper medication, the death toll caused by it can be reduced. Machine learning (ML) here plays a key role in helping the doctor throughout the prediction and detection phase of heart disease(HD) by catering to a stronger bias for decision making and prediction based on the patient's dataset provided by hospitals. This paper's primary goal is to develop a suitable, precise machine-learning model that can anticipate HD at an early stage. A diversity of machine learning techniques has been proposed for finding the most accurate method and feature subset. Different classification algorithms have been utilised, including logistic regression, random forest algorithm, gradient boosting, support vector machines (SVM), KNN, and decision trees (DT). An accurate model for predicting HD is obtained by employing several cross-validation approaches and using both public and private datasets. Applying the random forest classification approach to the combined dataset yields the best outcomes, with an accuracy of 85.25%.

Keywords
Machine learning, Heart disease, Support vector machine, K-Near est neighbor, Decision trees.

1. INTRODUCTION
Heart disease is a primary contributor to heart related disorders, which remain the leading cause of mortality across the globe. The World Health Organization indicates that heart disease and stroke cause 18.5 million deaths a year, or roughly 80% of all fatalities due to CVD. Hypertension, diabetes mellitus, hypercholesterolemia, psychological stress, and cigarette smoking, obesity, inactivity, neck and shoulder pain, poor appetite, and back pain are some of the many risk factors for cardiovascular illnesses [23]. The primary symptoms of heart disease are weakness, shortness of breath, arm and shoulder discomfort, and chest pain. There are certain traditional ways in which heart disease can be diagnosed, such as X-rays, MRI, and angiography. By accurately predicting cardiac disease by assessing complicated patterns, such as the correlations between many elements like age, cholesterol level, etc., machine learning models are better. This is because it is very difficult to find patterns and risk factors in more complex patient data that traditional methods can overlook. Several variables, including data noise, incompleteness, and abnormalities, make it challenging to anticipate reliable findings when using standard techniques for diagnosing cardiac disease. Preprocessing approaches are used on the dataset to remove incorrect data for a better accuracy rate during the evaluation of the model. Numerous techniques, comprising logistic regression, random forest classifier, decision trees, and support vector machines, can be utilized to diagnose the early stage of heart failure. Clustering (unsupervised machine learning), classification or regression (supervised machine learning), and reinforcement learning (error-driven) are the three fundamental methods for machine learning estimates.

Heart disease risk factors can be reduced by changing our lifestyle through exercising, reducing stress, controlling weight, quitting smoking, and taking proper medica tion. Through the vast amount of patient data that is accessible due to an enormous number of modern healthcare systems, it is now viable to make prediction models. [1,2,3] Machine learning, sometimes referred to as data storage models, employs an array of machine learning strategies and algorithms to transform data into knowledge-based mining. High classification accuracy is anticipated in this machine learning model with the use of logistic regression, KNN algorithms, and the random forest classifier. A comparative assessment of different machine learning approaches is attempted for accurately forecast the probability of developing heart disease. The results demonstrate that, with an accuracy of 85%, random forest delivers the best outputs.

2. LITERATURE REVIEW
The heart disease prognosis has attracted considerable focus lately because of the rising death rate linked to cardiac diseases. Timely and reliable identification of heart disease is crucial for decreasing fatal results and optimizing patient treatment. As healthcare data has quickly expanded, machine learning methods have been extensively investigated to help clinicians diagnose heart disease more effectively [21][22]. Alternative studies concentrate on assessing various machine learning algorithms, such as KNN, SVM, Decision Tree, and Random Forest, to determine the most dependable methods for predicting heart disease. These investigations highlight the significance of performance indicators like precision, recall, and accuracy to guarantee dependable predictions [5].

Numerous researchers have utilized supervised machine learning algorithms like Decision Tree, Random Forest, XGBoost, and Multilayer Perceptron to predict heart disease. In one investigation, the integration of clustering-based preprocessing with classifiers enhanced prediction accuracy, with the Multilayer Perceptron delivering the highest performance among evaluated models utilizing a large real-world dataset [3]. This emphasizes the advantage of integrating data preprocessing methods with sophis ticated classifiers Furthermore, sophisticated methods combine preprocessing strategies such as outlier identification, data normalization, and feature selection with robust classifiers like XGBoost and SVM. These combined models enhance prediction accuracy and aid clin ical decision support systems for diagnosing heart disease [1].

Comparative research utilizing datasets from the UCI Machine Learning Repository shows that ensemble models, especially Random Forest, surpass conventional algorithms like Naive Bayes and Logistic Regression in forecasting heart disease risk, reaching accuracy over 90% [4]. These findings validate the

strength of ensemble learning techniques in medical forecasting activities. In general, the literature shows that machine learning significantly contributes to predicting heart disease by enhancing diagnostic precision and minimizing human mistakes. Nonetheless, the effectiveness of prediction models significantly relies on suitable data preprocessing, feature selection, and optimization of the model. These results encourage additional studies focused on creating more precise and understandable heart disease prediction systems through sophisticated machine learning methods.

3. METHODOLOGY

For predicting heart disease, diverse classification algorithms based on machine learning are used on the patient dataset. Prior to beginning the feature selection process, the dataset must first be preprocessed; in this case, the dataset was selected from the Kaggle dataset. The dataset is separated into training and testing parts, with 80% of the dataset being utilized for training and the remaining 20% being employed for testing. A ma chine-learning model that forecasts cardiac illness is constructed using the training dataset, and the classification strategy is evaluated using the testing datasets.

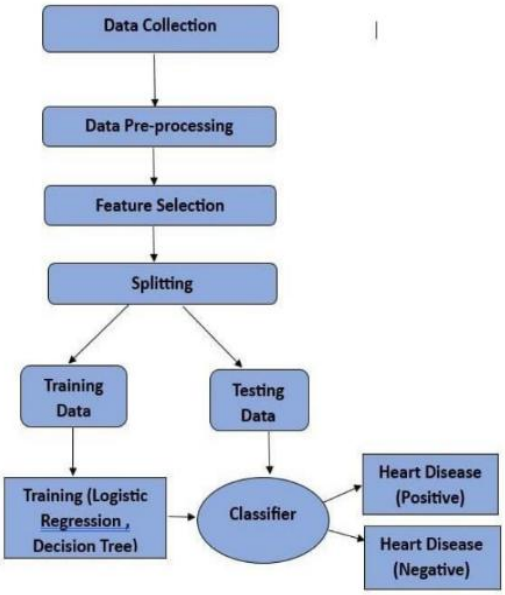

Fig. 1. Methodology of Heart Disease Risk Prediction System

Data collection, data preprocessing on the given dataset, exploratory data analysis, feature selection, model evaluation, splitting the data into training and testing data, optimization, and, lastly, model prediction and deployment are the steps involved in developing a machine learning prediction model. [11,12,13] Let's understand each step:

1) Data collection

Here, the dataset of over 1100 patients is used from Kaggle, which is a .csv file. To prevent biases, we should ensure that the dataset comprises both positive and negative people at risk of heart disease.

2) Data preprocessing

Data preparation deals with the identification of outliers, missing values, and inconsistencies since raw data contains noise and abnormalities. [14,15,16] Data preprocessing steps are:

a) *Cleaning:* deleting unnecessary entries and resolving missing data. Errors, outliers, and inconsistencies are removed in order to clean the data.

b) *Transformation:* employing a variety of methods, including smoothing, normalization, and aggregation, to transform the data into an appropriate format.

c) *Data encoding:* This stage involves separating numerical and categorical variables, encoding categorical data, and identifying unique values.

d) *Reduction:* This involves getting rid of unnecessary features. To avoid biases, duplicate entries are removed.

3) Exploratory Data Analysis

It can reveal hidden patterns, connections, and trends within a dataset. As a result, this method involves correlation analysis and visualization.

4) Feature selection

The most relevant feature from the dataset has been selected. Following data reduction, processing, and cleaning, it is completed. [17,18,19] This model generates predictions based on BP, cholesterol, and fasting sugar levels. A correlation matrix is used.

5) Data splitting

There is a particular testing-to-training (model training) ratio in this dataset. Eighty percent of the dataset comprises training data, while twenty per cent consists of testing data.

6) Evaluation (model assessment)

Performance indicators, like sensitivity, accuracy, precision, and F1 score, are utilized to examine the efficiency of the machine learning model. Accuracy represents the general performance of the model, and precision indicates the exactness of its positive predictions, and sensitivity is the pro portion of accurate positive predictions out of all positives (recall). The F1 metric is calculated as the harmonic mean of precision and recall.

4. TOOLS AND TECHNOLOGIES

Python programming is used. The sklearn and pandas libraries are used in data cleaning. Pandas are used for analyzing and transforming data. The NumPy library is used for numerical operations. Machine learning methods also make use of Sklearn Learn. Mat plotlib and Seaborn are Python libraries used to visualize (histograms, bar charts, etc). Coding is done in the Jupyter Notebook.

A. Dataset Description

The experimental analysis in this study utilizes the heart disease dataset which was collected from the UCI Machine Learning Repository. 1,100 patient records and 14 variables, make up this dataset. Age, sex, chest pain type (cp), resting blood pressure (trestbps), serum cholesterol (chol), fasting blood sugar (fbs), resting electrocardiogram results (restecg), maximum heart rate attained (thalach), exercise-induced angina (ex ang), exercise-induced ST depression (oldpeak), number of major vessels coloured by fluoroscopy (ca), and thalassaemia type (thal). The target variable demonstrates whether a patient has heart disease (1) or not (0). Because of its various clinical features and balanced makeup, this dataset is frequently utilized for predictive modelling and heart disease risk assessment research.

TABLE I. Dataset Attributes with their Description

Attribute	Description
Age	Age of the patient (years)
Sex	Gender (1 = male, 0 = female)
cp	Chest pain type (0–3)
trestps	Resting blood pressure (mm Hg)

chol	Serum cholesterol (mg/dl)
fbs	Fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
restecg	Resting ECG results (0–2)
thalach	Maximum heart rate achieved
exang	Exercise-induced angina (1 = yes; 0 = no)
oldpeak	ST depression induced by exercise
ca	Number of major vessels colored by fluoroscopy
thal	Thalassemia type (0–3)
target	Presence of heart disease (1 = yes; 0 = no)

B. Machine learning algorithms used

- SVM: A supervised classification algorithm that identifies the best hyperplane distinguishing data points from various classes.
- Random forest: An ensemble of multiple decision trees that improves prediction accuracy through majority voting among trees. [6,7,8]
- Decision tree: A model with a tree structure that divides data into branches according to decision rules to forecast results.
- Logistic regression: A logistic function based binary classification algorithm that forecasts the likelihood of an event.
- KNN: An algorithm that does not assume a fixed parameter and classifies data points are classified based on the dominant class of their k closest neighbors.
- Gradient Boosting: A method of ensemble learning that creates models in succession, with each successive model addressing the mistakes of earlier ones to enhance overall prediction precision.

5. IMPLEMENTATION

The heart disease risk prediction model is developed using Python as a programming language. Anaconda's command prompt has been utilized to complete the required code in a Jupyter notebook. Jupyter Notebook is better compared to PyCharm and Microsoft Visual. [9, 10, 20] With the help of Jupyter Notebook, we can code and simultaneously create graphs (scatter diagrams), and it also makes data analyses easier. Implementation of this machine learning risk prediction model is done in several steps:

- The Heart Disease related dataset in Comma Separated Values file format was imported into Jupyter Notebook using the Pandas read_csv() function to efficiently load and manage the data for analysis.
- To handle data, including feature engineering and data purification using the IsNull() function, a variety of libraries and resources are imported, such as pandas, numpy, matplotlib, seaborn, and sklearn. The pandas package is used to convert category data into numerical numbers.
- In order to ensure that every feature in the dataset contributes equally to the model, feature scaling is done to convert features to a common, comparable scale.
- Model selection: The model is chosen so that the dependent variables are represented as y and the features as x values. Training and testing data are separated using the sklearn program.
- The model exhibiting the best performance metrics and accuracy is utilized to predict the disease.

6. RESULTS AND DISCUSSION

For evaluating model effectiveness, the dataset is partitioned into training and test sets; where 80% of the dataset is utilized for model training, and 20% is set aside for evaluation. The dependability of the machine learning model is further examined using four performance metrics: accuracy, precision, sensitivity (recall), and F1 score. A confusion matrix is one table employed to evaluate the model's effectiveness. To assess the model's effectiveness, it presents both correct and incorrect predictions. Model predictions fall into four groups, including accurate outcomes (true positives and true negatives) and erroneous outcomes (false positives and false negatives).

- Accuracy: measures the model's correct prediction. It reflects the proportion of correctly classified instances among all predictions.
- Precision: It tells the model's prediction that a positive is actually positive. Out of all positive predictions, the actual correct prediction.
- Recall: it tells the occurrence of actual positive outcomes that were accurate positive outcomes. Out of all the positives, it tells us how many the model correctly identified.
- F1 score: This represents the fusion of accuracy and recall. It is the mean of recall and precision

The study examined the performance of Decision Trees, SVM, Random Forest, logistic regression, K-Nearest Neighbours, and Gradient Boosting in heart disease risk prediction tasks. The findings demonstrated that the inferior performance metrics of these models reflected their limited capability to understand the complexity involved in predicting heart disease. The Random Forest models were capable of differentiating between individuals with and without heart disease with high precision, recall, and ac curacy values, as well as prediction robustness. The healthcare industry is significantly impacted by the utilization of machine learning approaches in heart disease prediction. Physicians may identify heart disease issues early and take preventative action by utilizing machine learning techniques like random forest.

Accuracy, precision, recall, and F1 score were among the model performance indicators used to assess their performance. Random Forest did the best, with 85.25% ac curacy, 0.81 precision, 0.89 recall, and 0.85 F1-score, according to the statistics shown in the table below. Gradient Boosting and Logistic Regression with accuracies of 80.33% each but having somewhat lower F1 score than Random Forest. SVM and KNN scored worse, with accuracy of 59.01% and 63.93%, respectively; nevertheless, both had very high recall values, indicating that they were able to identify most positive cases, albeit with reduced accuracy.

TABLE II. Metrics Evaluation score of ML Algorithms

Algorithm	Accuracy	Precision	Recall	F1 Score
LR	80.327869	0.742857	0.757576	0.812500
SVM	59.016393	0.545455	0.827586	0.657534
KNN	63.934426	0.581395	0.862069	0.694444
DT	77.049180	0.741935	0.793103	0.766667
RF	85.245902	0.812500	0.896552	0.852459
GB	80.327869	0.757576	0.862069	0.806452

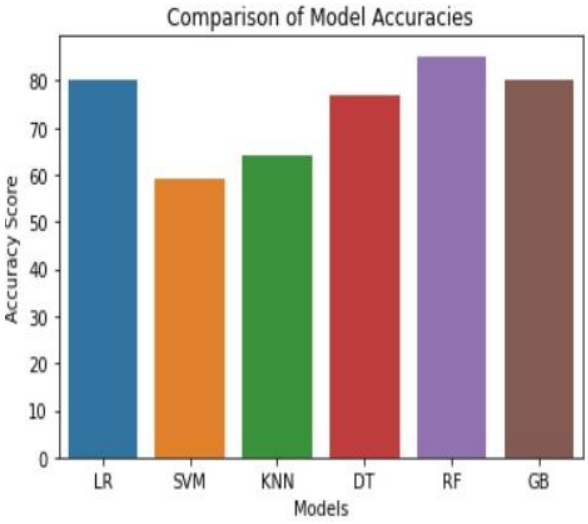

Fig. 2. Results of the ML Algorithms

Random Forest outperforms the other algorithms regarding accuracy and general balance between precision and recall, as the comparison study makes evident. Its ensemble technique, which combines several decision trees, explains why it performs better than Decision Tree, SVM, LR, KNN and GB. It can manage intricate feature intersections and minimize overfitting. Interestingly, SVM and KNN had lower accuracy and lower F1-scores even though they had high recall values (0.82 and 0.86, respectively). According to this, these models generated many false positives, which can result in needless medical alerts, even if they were able to detect the majority of actual positive instances.

Conversely, with accuracies of about 80%, Logistic Regression and Gradient Boosting provided a reasonable compromise between recall and precision. Even while these models performed consistently, the Random Forest still outperformed them in terms of overall predictive ability. Random Forest's excellent recall (0.89) is particularly important in a healthcare scenario where lowering false negatives (missed sickness cases) is critical. Because they could keep a patient from receiving timely medical at tention, False negatives pose a greater risk than false positives. Overall, the findings demonstrate that Random Forest is the most reliable model for risk estimation of heart disease in this study since it consistently and fairly performs across all evaluation criteria.

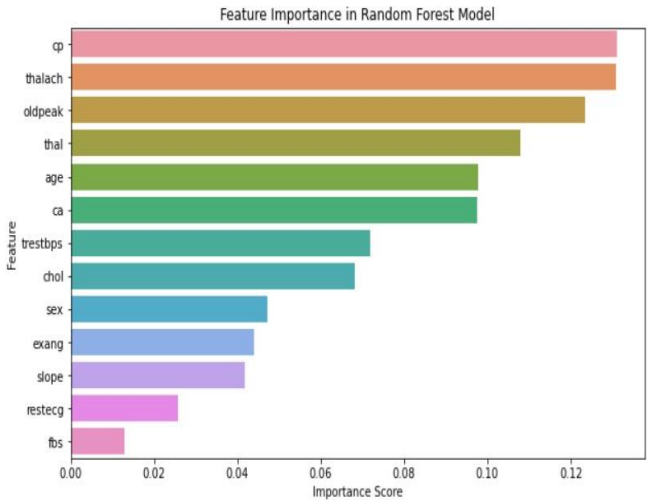

Fig. 3. Feature Importance in Random Forest Model

Clinical characteristics, including the type of chest pain (CP), the peak heart rate (thalach), and the ST depression induced by exercise (oldpeak), were identified as having the greatest impact on predicting heart disease through feature significance analysis. Age, the number of big vessels (ca), and thal were other important characteristics. Since aberrant exercise tolerance, chest discomfort, and ECG indications are important factors in determining cardiovascular risk, these findings are consistent with medical knowledge. On the other hand, characteristics like resting electrocardiogram (restecg) and fasting blood sugar (fbs) had comparatively lesser relevance, indicating limited predictive potential in this dataset.

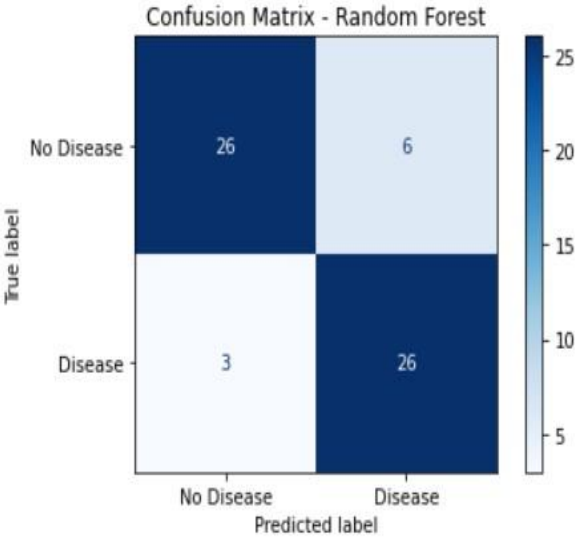

Fig. 4. Confusion Matrix of Random Forest

The Random Forest classifier acquired a balanced prediction capacity, as seen by the confusion matrix. 26 individuals without disease and 26 patients with disease were properly categorized by the model from the test samples. Three false negatives (patients with disease projected as healthy) and six false positives (patients without disease identified as diseased) were among the misclassifications. In medical diagnosis cases, when failure to discover a real case (false negative) is more crucial than an unneeded follow up test, this suggests a little larger inclination towards false positive predictions, which is acceptable.

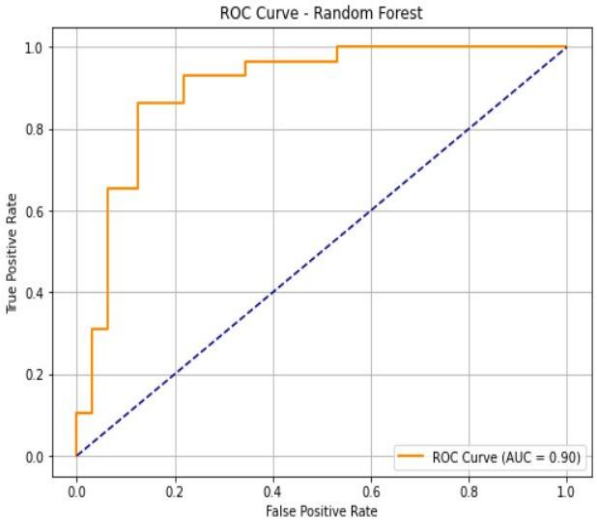

Fig. 5. ROC Curve of Random Forest

Strong discriminating capacity was demonstrated by the Random Forest model's Receiver Operating Characteristic (ROC) curve, which had an AUC value of 0.90. This implies that the model may be used to accurately identify between heart disease affected and non affected individuals. Additionally, the curve remained far above the diagonal baseline, showing that the classifier performed better than random guessing. Taking everything into account, the Random Forest model excelled in forecasting heart disease. The feature importance analysis emphasized the key risk factors, offering interpretability. The ROC-AUC score confirmed the precision of the predictions, whereas the confusion matrix validated the ability to classify data evenly. Although there are a few minor misclassifications, the model's tendency to minimize false negatives renders it particularly beneficial for medical purposes. These results indicate that Random Forest serves as an effective model for clinical decision support in forecasting heart disease risk, and its interpretability renders it a practical choice relative to more complex black box models.

7. FUTURE SCOPE

In the future, I will be working on a huge dataset to increase the dependability of the model. We will explore how to enhance the parameters of various machine learning classifiers by employing neural networks, such as CNN and different technologies like Federated learning and Ensemble learning. I'll be constructing a website with more complex capabilities, including the ability to upload the information straight to the website and then fill out the essential sections for the illness prediction. I'll also include a feature called Explainability that tells certain dataset variables (such as blood pressure, cholesterol, etc.) suggest the susceptibility to heart disease.

8. CONCLUSION

Heart disease is a deadly illness that takes many lives each year. Our objective is to use fewer characteristics and tests to improve prediction accuracy and effectiveness. In this work, investigation and assessment are undertaken utilizing a range of pre-processing techniques and machine learning algorithms. The prediction model is trained and assessed on three datasets. Following extensive testing and careful consideration of several crucial elements, the final application has an accuracy of 88%. This random forest accuracy was only attained for a limited set of 1100 patient records. A number of fac tors, including patient demographics, medical history, and lifestyle traits, may be included to help create powerful prediction models. Certain techniques, including logistic regression using decision trees, have differing degrees of success in accurately identifying those who are at risk. Issues include the necessity of comprehensive and varied datasets, as well as the interpretability of sophisticated models, and ethical issues around patient privacy continue to influence future research. Model validation and continuous improvement are essential. The application's accuracy and precision issues may be altered if it has a larger variety of data.

9. REFERENCES

[1] Nagavelli U, Samanta D, Chakraborty P. Machine learning technology-based heart disease detectionmodels. J Healthc Eng. 2022.

[2] Krittanawong C, Virk HU, Bangalore S, Wang Z, Johnson KW, Pinotti R, et al. Machine Learning prediction in cardiovascular diseases: a meta-analysis. Sci Rep. 2020.

[3] Bhatt CM, Patel P, Ghetia T, Mazzeo PL. Effective heart disease prediction using machine learning techniques. Algorithms. 2023.

[4] Rajdhan A, Agarwal A, Sai M, Ravi D, Ghuli P. Heart disease prediction using machine learning. Int J Eng Res Technol. 2020.

[5] Arunpradeep, N. & Niranjana, G. Different machine learning models based heart disease prediction. Int. J. Recent Technol. Eng. 8(6), 544–548 (2020).

[6] Alom, Z. et al. Early Stage Detection of Heart Failure Using Machine Learning Techniques. In Proceedings of the International Conference on Big Data, IoT, and Machine Learning, Cox's Bazar, Bangladesh, 23–25 September (2021).

[7] Rizvi, S.W.A., Singh, V.K. and Khan, R.A., 2017. Early stage software reliability modeling using requirements and object-oriented design metrics: fuzzy logic perspective. *International journal of computer applications*, *162*(2), pp.44-59.

[8] Zakria, N., Raza, A., Liaquat, F. & Khawaja, S. G. Machine learning based analysis of car diovascular disease prediction. J. Med. Syst. 41 (12), 207 (2017).

[9] Ngufor, C., Hossain, A., Ali, S. & Alqudah, A. Machine learning algorithms for heart disease prediction: a survey. Int. J. Comput. Sci. Inform. Secur. 14 (2), 7–29 (2016).

[10] Moon, S., Lee, W. & Hwang, J. Applying machine learning to Predict Cardiovascular dis eases. Healthc.Inf.Res. 25 (2),79–86. (2019).

[11] Wongkoblap, A., Vadillo, M. A. & Curcin, V. Machine learning classifiers for early detec tion of Cardiovascular Disease. J. Biomed. Inform. 88, 44–51. (2018).

[12] S.Nandhini, Monojit Debnath, Anurag Sharma and Pushkar, "Heart Disease Prediction us ing Machine Learning," International Journal of Recent Engineering Research and Devel opment (IJRERD), ISSN: 2455-8761, vol.3, pp.39-46, 2018.

[13] Abhijeet Jagtap, Priya Malewadkar, Omkar Baswat, Harshali Rambade "Heart Disease Pre diction using Machine Learning", Volume-2, Issue-2, February-2019.

[14] Galla Siva Sai Bindhika, Munaga Meghana, Manchuri Sathvika Reddy "Heart Disease Pre diction Using Machine Learning Techniques", (IRJET)April 2020.

[15] Rizvi, S.W.A., Singh, V. K. and Khan, R. A., "Revisiting software reliability engineering with fuzzy techniques," *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, New Delhi, India, 2016, pp. 1037-1042

[16] A. A. Stonier, R. K. Gorantla, and K. Manoj, "Cardiac disease risk prediction using machine learning algorithms," Healthcare Technol. Lett, vol. 11, pp. 213–217, November 2023.

[17] I. D. Mienye, and N. Jere, "Optimized ensemble learning approach with explainable AI for improved heart disease prediction," Information, vol. 15, p. 394, July 2024.

[18] H. F. El-Sofany, "Predicting heart diseases using machine learning and different data clas sification techniques," IEEE Access, vol. 12, pp. 106146–106160, August 2024.

[19] Garg, Apurv, Bhartendu Sharma, and Rijwan Khan. "Heart disease prediction using machine learning techniques." In IOP Conference Series: Materials Science and Engineering, vol. 1022, no. 1, p. 012046. IOP Publishing, 2021.

[20] Parveen, H, Syed Wajahat Abbas Rizvi, Raja Sarath Kumar

Boddu, “Fuzzy-Ontology based knowledge driven disease risk level prediction with optimization assisted ensemble classifier”, Data & Knowledge Engineering, Volume 151, https://doi.org/10.1016/j.datak.2024.102278, 2024.

[21] Muhammad Usman Aslam, Songhua Xu, Sajid Hussain, Muhammad Waqas, Nafiu Lukman Abiodun, “Machine learning-based classification of valvular heart disease using cardiovas cular risk factors‖ Sci Rep 14, 24396 (2024).

[22] A. Hammoud, A. Karaki, R. Tafreshi, S. Abdulla, and M. F. Wahid, "Coronary heart disease prediction: a comparative study of machine learning algorithms," J. Adv. Inf. Technol., vol. 15, pp. 27-32, 2024.

[23] Logabiraman G, Ganesh D, Kumar MS, Kumar AV, Bhardwaj N. Heart disease prediction using machine learning algorithms. MATEC Web Conf. 2024:392