International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 05, IIRA 5.0 (2026)

Correlative Exploration on Heart Disease Prediction Based On Machine Learning Algorithms

1st Gaurav Kumar
School of Computer Science and Engineering
IILM University
Greater Noida UP, India
gauravkumar@iilm.edu

2nd Lalit Kumar
School of Computer Science and Engineering
IILM University
Greater Noida UP, India
lalitkumar@iilm.edu

3rd Omkar Singh
Computer Science and Engineering
National Institute of Fashion Technology
Patna Bihar, India

4th Auwalu Falalu Hamza
Computer Science and Application
Sharda University
Greater Noida UP, India

5th Muhammad Hayatudeen
Computer Science and Application
Sharda University
Greater Noida UP, India

6th Khalid Saifullahi
Computer Science and Application
Sharda University
Greater Noida UP, India

***Abstract*— Cardiovascular diseases (CVDs) are the leading cause of global mortality; hence the need for efficient, non-invasive and interpretable diagnostics. This work is devoted to the prediction of heart disease through logistic regression, KNN, Naive Bayes, Decision Tree and Random Forest based on comprehensive indicators in Heart Disease dataset. With nearly 319,000 records comprising demographic, lifestyle and clinical data, logistic regression, KNN, Naive Bayes, Decision Tree and Random Forest attained a commendable accuracy of 91.04% with a balanced high recall of 91.04% and F1 score of 88.18%. These present evidence that logistic regression, KNN, Naive Bayes, Decision Tree and Random Forest model is effective to predict the occurrence of heart disease, combining interpretability and strong performance, which is imperative in clinical decisions. More complex and high performing models, such as neural networks, present more limitations in feasibility to the more resource-limited settings because of their level of non-transparency and their high computational cost. It is evident that logistic regression, KNN, Naive Bayes, Decision Tree and Random Forest has great potential for integration into the healthcare platform with early detection of heart disease to enable early intervention. Future research may allow hybrid models and real-time data to improve adaptability and generalizability to various populations. (*Abstract*)**

Keywords—Cardiovascular Disease (CVD), Logistic regression, KNN, Naive Bayes, Decision Tree and Random Forest, Heart Disease Prediction, Machine Learning (ML), Data Pre-processing, Model Evaluation Metrics, Interpretability, Predictive Analytics, Clinical Decision Support, Healthcare Technology

I. INTRODUCTION

Cardiovascular diseases (CVDs) Causative factors are now a key global health problem and contribute mainly to increased mortality worldwide. These conditions are mostly made up of conditions like coronary heart disease, heart attack, and atrial fibrillation. The causes are primarily associated with lifestyles such as poor diet, lack of exercise, as well as genetic constitutions. Early detection is very much critical in improving the outcome management in patients; however, traditional diagnostic methods involve invasive, costly, and inaccessible to even resource-limited setups. New upcoming technologies have an insightful vision in machine learning (ML) which would open up new avenues in developing non-invasive diagnostic facilities through low cost and high diagnosing efficacy. Logistic regression, KNN, Naive Bayes, Decision Tree and Random Forest, an ML algorithm mostly used, interpretable and has shown exceptional potential in predication of heart diseases and then guides the individual in informing decisions from healthcare professionals. This study is concerned with the application of a dataset from Kaggle in effecting its strong performance and clinical applicability in predicting heart disease through logistic regression, KNN, Naive Bayes, Decision Tree and Random Forest.

II. BACKGROUND

Cardiovascular diseases (CVDs)Ranges from a variety of disorders which include the heart and blood vessel disorders like coronary heart disorder and heart collapse, as well as atrial fibrillation [1, 2]. According to the top external sources of intern, about 17.9 million deaths per year in 1999 would be due to these two health conditions as stated by the World Health Organization (WHO). Some other lifestyle factors greatly contribute to the incidence of the disease, which include poor diet and lack of exercise. Increased heart rate. And, of course, there is genetic predisposition and co-morbidities such as diabetes and high blood pressure. Increase risk Make detection early. It is very important part of the health care. They pay to observe weight. More so in resource-limited situations where the challenge is rendered greater because of the scantiness of specialists and advanced diagnostic tool.[2]. Commonly adopted diagnosis indeed utilize invasive techniques such as angiography that are costly, time-consuming, and not readily available, thereby demanding the need to innovate non-invasive, cost-effective, and highly precise diagnostic methods. Any such methods when applied are likely to take an early diagnosis of heart diseases and institute clinical management therapies that could very well save lives and improve their quality. Early diagnosis allows direct treatment interventions, lifestyle modification, and monitoring that could slow the disease's progress. Nonetheless, the objectivity of accurate diagnosis mostly fails to retain its purity because subjective clinical judgment usually makes strong demands, in addition to individual variability among patients with exactly the same symptomatology.[3]. This advancement is all about machine learning (ML) that eventually dominated this science and created instruments to examine data at great lengths, disclose hidden connotations, and come up with accurate

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 05, IIRA 5.0 (2026)

predictions. Machine learning has turned out to be an absolute game changer in terms of healthcare-industry applications, ranging from diagnostic imaging and drug discovery to patient monitoring and forecasting. [4]. Machine learning algorithms can easily handle nonlinear relationships and can deal with very complex high-dimensional data, unlike the ordinary statistical methods. This makes it amenable to medical applications. More to the point, it is perfect for binary classification tasks like prediction of presence or absence of disease. It has an advantage over other forms because of its simplicity, computational efficiency, and interpretability assigned to it for use in clinical settings. [5].

III. Related Works

Applications of machine learning in health but mainly for prevention of cardiovascular disease. "Recent years have also witnessed an increased interest in ML's potential to process massive amounts of data. It has stimulated interest in various fields of medical diagnosis for discovering hidden patterns and making accurate predictions." The present section reviews extensive literature. It devotes toward the evolution of machine learning strategies for predicting heart disease. Comparative effectiveness and the challenges they face Initial attempts to predict heart disease heavily depended on the traditional statistical methods and algorithms. However, one of the two oldest and most widely used models is logistic regression, KNN, Naive Bayes, Decision Tree and Random Forest. This is owing to its simplicity and efficiency in binary classification tasks. Studies such as those by Ambrish, Ganesh, Ganesh, Srinivas and Mensinkal [6] have shown that logistic regression, KNN, Naive Bayes, Decision Tree and Random Forest can achieve high accuracy rates when applied to structured medical datasets. His work emphasizes the interpretability of logistic regression, KNN, Naive Bayes, Decision Tree and Random Forest. It has turned it into a popular choice for doctors seeking information on characteristics like cholesterol levels or blood pressure. How will that affect the patient's prognosis? Other default models include decision trees. It provides a hierarchical structure for decision-making. For example, Maheswari and Pitchai [7] used decision trees to predict medical conditions, including heart disease, achieving commendable accuracy. While effective, these models often suffered from issues such as overfitting, especially when applied to small datasets or noisy data(document).

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

A. Evolution of Machine Learning Models

As data sets become larger and more complex Researchers have therefore begun to explore advanced machine learning algorithms to improve prediction accuracy. Neural networks have gained popularity This is due to its ability to simulate non-linear relationships and deal with high dimensional data Studies like those by Mehmood, Iqbal, Mehmood, Irtaza, Nawaz, Nazir and Masood [8] have shown their effectiveness. Artificial neural networks for predicting heart disease have 90% accurate taxa, however, the "dark box" nature of neural networks raises concerns about their use in clinical settings. Interpretability is important. Support Vector Machines (SVM) have also gained strength due to their robustness in processing linear and non-linear discrete data. Singh and Kumar [9] employed SVM to analyze heart disease datasets, achieving accuracy rates of 87.5%. While SVMs outperformed simpler models like decision trees in certain scenarios, their computational complexity limited their scalability to larger datasets.

Ensemble methods, such as Random Forests and Gradient Boosting Machines (GBM), further advanced heart disease prediction. These models combined the outputs of multiple base classifiers to enhance predictive performance. Singh, Sinha, and Singh [10] demonstrated the use of Random Forests in healthcare, achieving higher accuracy than standalone models. However, these methods often required extensive computational resources and careful Hyperparameter tuning, making them less accessible in resource-constrained environments.

B. Logistic regression, KNN, Naive Bayes, Decision Tree and Random Forest in Heart Disease Prediction Evolution of Machine Learning Models

Despite advances in machine, learning Yet logistic regression, KNN, Naive Bayes, Decision Tree and Random Forest remains the cornerstone for cardiovascular disease prediction due to its balance of simplicity, efficiency, and interpretability. Studies like those by Zulkiflee and Rusiman [11] highlight its ability to Logistic regression, KNN, Naive Bayes, Decision Tree and Random Forest to achieve competitive accuracy (91.67%) while maintaining decision transparency. is that prognosticators need to understand the relationship between (e.g. cholesterol, age) and outcomes (e.g. presence of disease) is critical [11]. A number one advantage of logistic regression, KNN, Naive Bayes, Decision Tree and Random Forest is it's capability to bestow probability scores. This can be used to stratify patients according to their level of risk. This feature has been exploited in many clinical applications. This enables targeted intervention for high-risk individuals. Moreover, the computational efficiency of logistic regression, KNN, Naive Bayes, Decision Tree and Random Forest makes it suitable for real-time applications such as telemedicine platforms. This requires quick predictions. An excellent style manual for science writers is [7].

C. Evolution of Machine Learning Models

Several studies have compared the performance of logistic regression, KNN, Naive Bayes, Decision Tree and Random Forest with other machine learning models. To envision heart disease, for example, Sarah, Gourisaria, Khare and Das [12] analyzed the prediction accuracy of artificial neural networks. Decision tree and the Naïve Bayes classifier found that the neural network achieved the highest accuracy (100%) on the dice set. However, they also noted the limitations of the neural network. These include reliance on large data sets and a lack of interpretability. On the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 05, IIRA 5.0 (2026)

Contrary, Logistic regression, KNN, Naive Bayes, Decision Tree and Random Forest provides reasonable accuracy with the added benefit of being interpretable and computationally efficient. A similar comparative analysis by Sahid, Hasan, Akter and Tareq [13] highlights the strengths of logistic regression, KNN, Naive Bayes, Decision Tree and Random Forest in dealing with unbalanced datasets. This is a common problem in medical data. By combining normalization techniques, Logistic regression, KNN, Naive Bayes, Decision Tree and Random Forest therefore outperforms more complex models in situations, where data pre-processing is limited. These findings reinforce the suitability of logistic regression, KNN, Naive Bayes, Decision Tree and Random Forest for real-world applications. The quality and availability of data often varies.

IV. METHODOLOGY

The study methods were carefully designed, to ensure robust and reliable heart disease prediction using logistic regression, KNN, Naive Bayes, Decision Tree and Random Forest. This section describes the main steps. Includes a detailed description of the dataset. Pre-processing techniques Resource selection using, the model and evaluation protocols To address common challenges in medical data analysis such as missing values and class imbalance. Guarantees of methods or development of accurate and interpretable prediction models. The dataset used in this study is the Kaggle Indicators of Heart Disease dataset (2022 update). It is a collection of approximately 319,000 records, making it one of the two most extensive publicly available datasets for heart disease analysis. Each record represents an individual and includes a combination of demographic, lifestyle and clinical characteristics. Key characteristics include age, gender, smoking, alcohol consumption, body mass index (BMI), physical, activity, Cholesterol, blood pressure and blood sugar levels in Chejum Island Alvo variables are binary. This indicates the Occurrence (1) or inactivity (0) of heart disease. The richness and diversity of the dataset provides an excellent basis for building effective predictive models.

Pre-processing the two raw data It is an important step to ensure that they are clean, consistent, and adequate for machine learning. Missing information. This is a common problem in medical datasets. The form can be edited using data entry techniques. For continuous variables such as cholesterol and blood pressure. The mean or median is used depending on the distribution. For other types of variables, such as smoking, fashion is used to replace missing values. This approach preserves the integrity of the dataset. and reduce data loss to a minimum. to format numerical features such as cholesterol levels and blood pressure It has been applied at the minimum-maximum level. To normalize the value of [0, 1], this ensures that all resources contribute equally to the model training process. They avoid pre-conceived ideas due to size differences. These categorical forum variables were coded in numerical form to make them suitable for analysis using logistic regression, KNN, Naive Bayes, Decision Tree and Random Forest models. Binary resources such as gender are coded as 0 or 1 by forums for multiclass resources such as smoking status. One-hot encryption is used. Create a separate binary column for each category. Class Imbalance This is the essential set of challenge dice. It is modified using the Synthetic Minority Overamostragem (SMOTE) technique. This method creates a synthetic minority overamostragem (SMOTE). It effectively balances or sets the dice without causing any noise. To guarantee a balanced distribution of positive (1) and negative (0) cases, these pre-processing steps create a solid basis for training the model. Resource selection is another important aspect of this method. It aims to improve the archetype's efficiency by identifying the most suitable predictors of heart disease. A correlation matrix was initially created to examine the relationships between resources. High core resources

The logistic regression, KNN, Naive Bayes, Decision Tree and Random Forest models were implemented to presume the likelihood of heart disease. Logistic regression, KNN, Naive Bayes, Decision Tree and Random Forest are statistical ways that uses The sigmoid characteristic to map the weighted sum of enter capabilities to a chance score, bounded between 0 and 1. The logistic function is defined as:

where $P(Y = 1|X)$ is the predicted Statistical likelihood of heart disease, X_i are the input features, and β_i are the coefficients that determine the contribution of each feature. The model parameters were optimized using stochastic gradient descent (SGD), a computationally efficient algorithm well-suited for large datasets. To prevent overfitting, L2 regularization (Ridge Regression) was applied, introducing a penalty term to the cost function. This approach ensured that the Simulation generalized good to Disguised data by constraining the magnitude of the coefficients.

The dataset was Split into training and testing subsets, with 80% of the data used for training and 20% reserved for testing. To ensure the robustness of the model, K-fold cross-validation with $K = 1$ was performed. The model gets prepared with K-1 groups of the k subsets of data and tested against the left-out one. This process is repeated for k times and results are averaged so as to get a good estimate of the model's performance. A technique like this reduces the influence of data partitioning on evaluation since the entire model is evaluated on the generalizability of that model.

V. RESULT & DISCUSSION

Evaluation of the performance of logistic regression, KNN, Naive Bayes, Decision Tree and Random Forest models is carried out by incorporating a battery of metrics that provide a total overview of the predictive capacity of the model. Accuracy, which is the proportion of instances correctly classified. The primary measure is accuracy: it is the ratio between the true forecasts and all positive forecasts. Assess the confidence of the Positive Recall category or sensitivity. The F1 score of the model indicates its ability to detect all positives as it is the harmonic mean of precision and recall. Give a balanced assessment this is especially useful in unbalanced data sets. The arsenal of metrics provides overall assurance in rigorous assessment of the model performance. In addition, is expressed mathematically as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \#(2)$$

$$Precision = \frac{TP}{TP + FP} \#(3)$$

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 05, IIRA 5.0 (2026)

$$Recall = \frac{TP}{TP + FN} \#(4)$$

$$F - Score = 2 * \frac{Precision * Recall}{(Precision + Recall)} \#(5)$$

TABLE I. CONFUSION MATRIX PARAMETERS WITH RESPECT TO VARIOUS MODELS

Models	Confusion Matrix Parameters			
	Accuracy	*Precision*	*f1-Score*	*Recall*
Decision Tree	86.11%	22.85%	23.76%	24.75%
Naïve Bayes	84.31%	26.64%	33.54%	45.28%
Random Forest	90.42%	35.58%	17.71%	11.78%
KNN	90.40%	36.74%	19.82%	13.57%
Logistic Regression	91.28%	50.84%	14.74%	8.62%

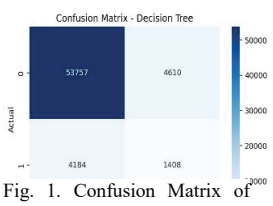
[a] Sample of a Table footnote. (*Table footnote*)

Example of a figure caption. (*figure caption*)

The five machine learning models' performance metrics—Decision Tree, Naïve Bayes, Random Forest, KNN, and Logistic Regression—present major strengths and weaknesses in various domains. Accuracy-wise, all other models are surpassed by Logistic Regression (91.28%), but by just a thin margin by Random Forest (90.42%) and KNN (90.40%). Yet, poor recall cannot be a measure of a good model since Logistic Regression is one of the most poorly recalled models at 8.62%. Precision, which gauges the capacity to accurately classify positive cases without including false positives, is highest for Logistic Regression (50.84%), signifying that Logistic Regression is conservative in its positive classification. However, recall, which is most important for detecting as many true positives as possible, is highest for Naïve Bayes (45.28%), hence ranking best when losing positive cases is important. The F1-score, a balance of precision and recall, is likewise highest for Naïve Bayes (33.54%), implying that it offers the best trade-off between these two scores. Conversely, Random Forest and Logistic Regression, although possessing high accuracy, possess low F1-scores, indicating imbalances in their respective performances. According to these findings, Logistic Regression is most appropriate in situations where accuracy is paramount and Naïve Bayes is the most well-rounded model based on its better recall and F1-score. If the key objective is general detection of positive cases, Naïve Bayes would be the best choice, while Logistic Regression would be appropriate when reducing false positives is the priority.

Graphical comparison of all the given models is given as follows

Figure 1-5 Shows the confusion matrices of all classifiers



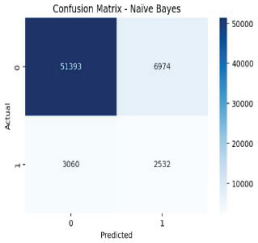


Fig. 1. Confusion Matrix of Decision Tree

Fig. 2. Confusion Matrix of Naive Bayes

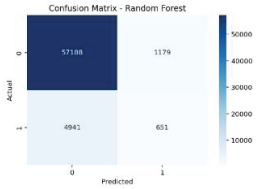


Fig. 3. Confusion Matrix of Random Forest

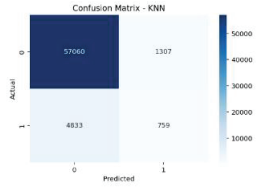


Fig. 4. Confusion Matrix of KNN

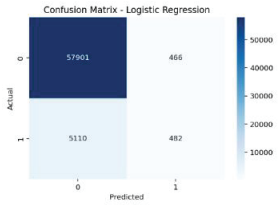


Fig 5. Confusion Matrix Logistic Regression

Figure 6-8 Shows the confusion matrices of all classifiers

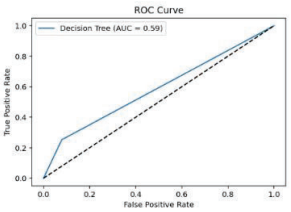


Fig. 6. ROC Curve of Decision Tree

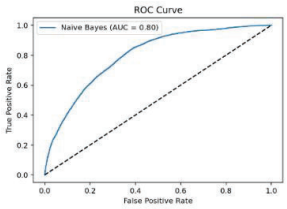


Fig 7. ROC Curve for Naïve Bayes

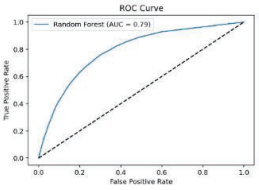


Fig. 8. ROC curve of Random Forest

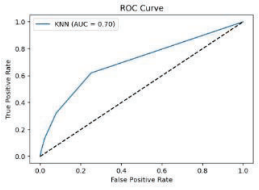


Fig. 9. ROC Curve of KNN

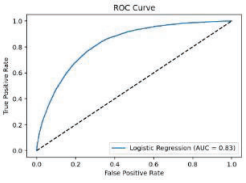


Fig. 10. Confusion Matrix Logistic Regression

The comparison of all classifiers with representation of all parameters is given in Fig. 11.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 05, IIRA 5.0 (2026)

forecasting performance.Looking forward, there are numerous opportunities to extend this research. Future work

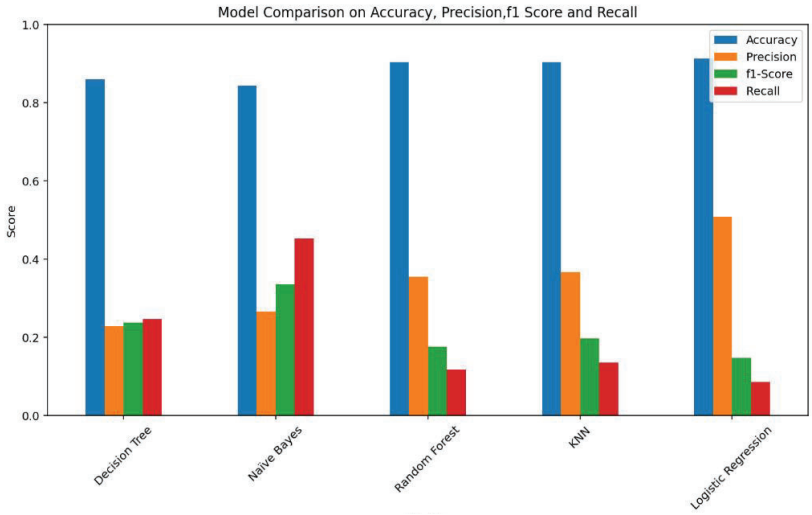


Fig. 11. Comparison of all matrix for various parameters

VI. Conclusion

In this study, we explored the effectiveness of logistic regression, KNN, Naive Bayes, Decision Tree and Random Forest in predicting heart disease using the Kaggle Indicators of Heart Disease dataset (updated 2022). The results showed that logistic regression, KNN, Naive Bayes, Decision Tree and Random Forest with accuracy 91.04% is a highly effective and interpretable method in predicting heart disease. Strong functioning of the model Along with metrics such as recall, F1 score, and precision, it indicates that logistic regression, KNN, Naive Bayes, Decision Tree and Random Forest can distinguish between people with and without heart disease, reversals, or valuable parameters for Detect it early, And preventative health care reliably. In addition, the model's ability to provide interpretable information about the contribution of important characteristics like age, cholesterol levels, and blood pressure, It guarantees that health care providers can make informed decisions based on the model's predictions. Into addition, this interpretability becomes increasingly critical in a clinical environment where transparency becomes essential. The significance of this finding resonates well as it illustrates the use of logistic regression, KNN, Naive Bayes, Decision Tree and Random Forest in health scenarios, especially in resource-poor environments. The sex of the model performance and interpretability allows for integration into the existing health systems including electronic health records (EHRs) or telemedicine platforms - - - help providers in identifying those at high risk and then ensuring that they receive timely intervention. However, this model showed good performance in the present study, but it also has some drawbacks, such as requiring well-preprocessed datasets and expectation of linear relationships between resources. Future studies could overcome these drawbacks by employing more complex models or hybrid techniques. This can add value to could focus on validating the model across diverse populations and healthcare settings to ensure its generalizability. Additionally, integrating real-time data from wearable devices and incorporating explainable AI techniques could further improve the model's adaptability and transparency. Ultimately, this research underscores the growing role of machine learning in healthcare, where models like logistic regression, KNN, Naive Bayes, Decision Tree and Random Forest can significantly contribute to the early discovery and strategy of chronic diseases like heart disease, ultimately improving patient outcomes and reducing healthcare costs.

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] [M. Marimuthu, M. Abinaya, K. Hariesh, K. Madhankumar, and V. Pavithra, "A review on heart disease prediction using machine learning and data analytics approach," International Journal of Computer Applications, vol. 181, no. 18, pp. 20-25, 2018.

[3] A. Hazra, S. K. Mandal, A. Gupta, A. Mukherjee, and A. Mukherjee, "Heart disease diagnosis and prediction using machine learning and data mining techniques: a review," Advances in Computational Sciences and Technology, vol. 10, no. 7, pp. 2137-2159, 2017.

[4] K. Saxena, and R. Sharma, "Efficient heart disease prediction system," Procedia Computer Science, vol. 85, pp. 962-969, 2016.

[5] K. R. Chowdary, P. Bhargav, N. Nikhil, K. Varun, and D. Jayanthi, "Early heart disease prediction using ensemble learning techniques." p. 012051.

[6] G. Ambrish, B. Ganesh, A. Ganesh, C. Srinivas, and K. Mensinkal, "Logistic regression, KNN, Naive Bayes, Decision Tree and Random Forest technique for prediction of cardiovascular disease," Global Transitions Proceedings, vol. 3, no. 1, pp. 127-130, 2022.

[7] S. Maheswari, and R. Pitchai, "Heart disease prediction system using decision tree and naive Bayes algorithm," Current Medical Imaging, vol. 15, no. 8, pp. 712-717, 2019.

[8] A. Mehmood, M. Iqbal, Z. Mehmood, A. Irtaza, M. Nawaz, T. Nazir, and M. Masood, "Prediction of heart disease using deep convolutional

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 05, IIRA 5.0 (2026)

neural networks," Arabian Journal for Science and Engineering, vol. 46, no. 4, pp. 3409-3422, 2021.

[9] A. Singh, and R. Kumar, "Heart disease prediction using machine learning algorithms." pp. 452-457.

[10] Y. K. Singh, N. Sinha, and S. K. Singh, "Heart disease prediction system using random forest." pp. 613-623.

[11] N. F. Zulkiflee, and M. S. Rusiman, "Heart Disease Prediction Using Logistic regression, KNN, Naive Bayes, Decision Tree and Random Forest," Enhanced Knowledge in Sciences and Technology, vol. 1, no. 2, pp. 177-184, 2021.

[12] S. Sarah, M. K. Gourisaria, S. Khare, and H. Das, "Heart disease prediction using core machine learning techniques—a comparative study," Advances in Data and Information Sciences: Proceedings of ICDIS 2021, pp. 247-260: Springer, 2022.

[13] M. A. Sahid, M. Hasan, N. Akter, and M. M. R. Tareq, "Effect of imbalance data handling techniques to improve the accuracy of heart disease prediction using machine learning and deep learning." pp. 1-6.