

Predicting Heart Disease with Machine Learning: A Comparative Study

Swetha R

Department of computer science,
CHRIST University
Bangaluru ,India,560029
swetha.r@msam.christuniversity.in

Helen. K.Joy

Department of computer science,
CHRIST University
Bangaluru ,India ,560029
helenjoy88@gmail.com
<https://orcid.org/0000-0001-7776-3856>

Warusia Yassin,

Universiti teknikal malaysia melaka
Serdang, Selangor, Darul Ehsan,
Malaysia
s.m.warusia@utem.edu.my

Sridevi.R

Department of computer science.
CHRIST university
Bangalore ,India,560029
sridevi.r@christuniversity.in

Drakshayini M N

Department of ECE
T. John Institute of Technology
Bengaluru , India
mndrakshayini@gmail.com

Fabiola Hazel Pohrmen

Department of Computer Science
CHRIST university
Bangalore ,India,560029
fabiola.hazel@christuniversity.in

Abstract— Heart disease is considered as the leading causes of death worldwide. Detection in early stage helps to prevent the disease. Nearly 17 million heart disease deaths occur every year. If heart disease is detected early, it could save many lives. Using machine learning algorithms to detect heart disease early will benefit many people. The present study explains the utilization of machine learning algorithms and artificial intelligence in the detection of cardiac illness. Data science plays a crucial role in healthcare by facilitating the analysis and processing of large volumes of data, therefore enabling the application of Artificial Intelligence (AI) and Machine Learning (ML). This study aims to identify the optimal model for predicting heart disease by employing different classification methods of machine learning, including Logistic Regression (LR) and Support Vector Machine. keywords Artificial intelligence, machine learning, prediction of heart disease, selection of features, decision trees, logistic regression, and accuracy.

Keywords— Machine Learning, Logistic Regression, Support Vector Machine, Accuracy, Precision.

I. INTRODUCTION

Comprising about 17 million heart disease cases every year. The heart is a fundamental and essential organ of the human body. Since it is vital to protect our hearts. current heart disease diagnosis approaches are inefficient in early detection. Thus creating an efficient method to detect cardiovascular disease early. The integration of Artificial Intelligence (AI) and Machine Learning (ML) has become a highly promising technique in healthcare, providing a revolutionary method for evaluating risks and detecting cardiovascular problems at an early stage. Using proper machine learning algorithms can predict heart disease at an early stage. Artificial intelligence and supervised machine learning algorithms like support vector machine (SVM), and logistic regression can be used in datasets to predict accuracy for detecting heart disease. Heart disease varies from patient to patient; based on a person's health, it is essential to understand their health conditions and build algorithms. Training algorithms on giant data sets to get proper results is essential. Doing proper feature selection is very critical to get accurate results. The development of AI models will help doctors to detect diseases very precisely. These ML

algorithms take patient historical data and detect the occurrence of heart disease..

II. LITERATURE SURVEY:

As well as Sujata Chakravarthy, etc. Within the scope of this research study, a robust system for predicting ischemic cardiovascular disease through the application of machine learning applications is proposed. The study, which critically evaluates the performance of six alternative algorithms concurrently, demonstrates that the K Nearest Neighbour (KNN) method yields the most precise predictions. The research methodology entails gathering data from a number of different datasets, preparing and analysing the data, as well as training and testing machine learning algorithms. The purpose of this project is to improve the accuracy of prognosis for ischemic cardiovascular disease by utilising the KNN algorithm. Additionally, the performance of the algorithm would be evaluated using a variety of assessment metrics. After the hyperparameters have been tuned, the results demonstrate that the KNN model achieves the highest performance attainable. In addition, the study contains references that pertain to the application of machine learning and data mining techniques in the medical field. These approaches include research that investigate the prediction of cardiac disease through the utilisation of random forest and gradient-boosting classification techniques. The use of deep convolutional neural networks for the classification of breast lesions in mammographic images is another topic that is covered in this study. [1]

Obstructive sleep apnea (OSA) is the concentration of Daniele Padovano and colleagues. It is a condition that is associated with heart disease in which a person stops breathing while they are sleeping. Obesity or pressure in the chest cavity are two possible causes of this condition. Predicting the electrocardiogram (ECG) can be accomplished with the application of machine learning methods. Using public datasets from sources such as Kaggle and the University of California, Irvine does not

give adequate accuracy. The author came to the conclusion that in order to achieve complete accuracy, additional datasets should be trained on. This study is the first of its kind to analyse OSA in terms of epochs.2 [2]

The research conducted by Abdallah Abdel-Latif and colleagues focuses on developing a comprehensive method for identifying cardiac illness and classifying its severity through the application of machine learning and hyperparameter optimisation techniques. In order to deal with imbalanced data, the model uses the Synthetic Minority Oversampling Technique (SMOTE), and it also assesses six different machine learning classifiers. For the purpose of determining which hyperparameters are optimal for the classifiers, hyperparameter optimisation is utilised. The findings establish that the suggested approach is capable of effectively identifying cardiac disease and categorising the severity levels of the condition. In addition to this, the paper analyses the performance of several machine learning models and emphasises the efficacy of hyperparameter optimisation in terms of enhancing model performance.[3]:

To anticipate the severity of cardiac illness, Rina S. Patil and colleagues propose an architecture that combines machine learning methods with the Internet of Things (IoT). For machine learning algorithms to achieve the highest possible level of accuracy, the study highlights the significance of feature selection. Based on the findings of experiments, it has been demonstrated that integrating several classifiers, such as support vector machines, results in improved forecasting skills. The fourth paragraph by Senthil Kumari delineates the application of machine learning algorithms with the explicit aim of predicting heart attacks. The author conducts an analysis of the effectiveness of several categorization algorithms and authenticates the system based on a collection of characteristics. Hence, the study suggests that implementing machine learning methods, particularly the random forest algorithm, has the capacity to enhance the accuracy of heart attack prediction and improve healthcare outcomes. There are a variety of themes that are covered by the references in this text that are associated with artificial intelligence and data analytics in the healthcare industry. Some of these topics include personalised precision medicine, machine learning in healthcare communications, and big data analytics in healthcare. Using the random forest method, the researchers were able to achieve a 95 percent accuracy rate in their prediction of heart attacks, which was the finest result of the study. Because this accuracy rate is higher than the work that has been done previously, it demonstrates that the approach that has been proposed is effective [5].

It is discussed by Mert Ozcan and colleagues that a classification and regression tree (CART) method can be utilised for the purpose of predicting cardiac disease. For the purpose of this study, a comprehensive dataset was utilised, and the CART algorithm was utilised to extract decision rules and rank the characteristics that influence

high blood pressure. The algorithm was able to accurately predict heart disease with an accuracy rate of 87 percent, with the ST slope and the type of chest discomfort being the most relevant factors. According to the model, these algorithms have the potential to be utilised for early disease prediction. On the other hand, the study has several shortcomings, and additional research is required to validate the findings and make the model better. In addition, the article makes reference to additional research publications and articles that focus on the prediction of cardiac disease through the utilisation of data mining and machine learning approaches. [6]

Through the utilisation of machine learning techniques, Yasar Amin and colleagues have proposed a reliable system for the prediction of ischemic cardiovascular disease. The study, which evaluates the performance of six different algorithms, reveals that the K Nearest Neighbour (KNN) method achieves the best degree of performance, with an accuracy of 91.8 percent. The limits of previous research are brought to light in this work, and the authors intend to overcome such constraints by employing a comprehensive dataset and include essential characteristics. Collecting and combining data from a number of different datasets, preparing and analysing the data, and training and testing machine learning algorithms are all components of the study technique. The purpose of this project is to improve the accuracy of prognosis for ischemic cardiovascular disease by utilising the KNN algorithm. Additionally, the performance of the algorithm would be evaluated using a variety of assessment metrics. After the hyperparameters have been tuned, the results demonstrate that the KNN model achieves the highest performance attainable. Using the OBL-GWO technique for feature selection, Sivasubramaniam et al. address the construction of an optimal artificial neural network (ANN) based heart disease classification model. [7] This model was developed by Sivasubramaniam and colleagues. For the purpose of forecasting heart disease, the model that has been proposed reaches an accuracy of 92.54 percent. In comparison to previous hybrid models, the OBLGWO algorithm is superior in terms of choosing the best possible subset of features. In addition to this, the article discusses the significance of feature selection methods in the diagnosis of heart disease and presents the OBL-GWO model as a method for the detection of heart illness. The proposed approach for the identification of heart disease incorporates fourteen significant characteristics and is capable of making reliable predictions. Feature selection approaches contribute to the optimisation of the prediction model by lowering the costs of processing and increasing the accuracy of the model. Within the framework of the GWO algorithm, the social order of grey wolves serves as a source of inspiration. An improvement in accuracy, recall, and precision in illness classification has been demonstrated by the utilisation of ANN models in conjunction with feature selection methods such as OBL-GOA and IMGWO. [8]:

Jian Ping Li and colleagues have proposed a method that utilises machine learning techniques to develop a system

```
[ ] data=pd.read_csv("/content/heart.csv")
```

```
data.head()
```

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
1	49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
2	37	M	ATA	130	283	0	ST	98	N	0.0	Up	0
3	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
4	54	M	NAP	150	195	0	Normal	122	N	0.0	Up	0

Fig. 1. Preview of Heart Disease Dataset Using Pandas in Python

that is both effective and accurate in diagnosing cardiac disease. Support Vector Machine, Logistic Regression, Artificial Neural Network, K-Nearest Neighbour, Naive Bayes, and Decision Tree are some of the classification techniques that are utilised by the system. Standard feature selection algorithms are utilised in order to facilitate the selection of pertinent features and enhance the accuracy of categorization. After being evaluated on the Cleveland Heart Disease dataset, the above mentioned approach exhibits an enabling level of accuracy in relation to the systems that are already implemented on the market today. In addition, the study presents a novel feature selection technique that is referred to as Fast Conditional Mutual

Information (FCMIM), and it evaluates the performance of this algorithm in comparison to the performance of previous feature selection algorithms. [9] Abdul Kareem and others etc. The proposed system in the study article has claimed 92.37% of accuracy by using the FCMIM in combination with the SVM [7] classifier for the feature selection process. Other classification algorithms, used by the system, include; Logistic Regression, Artificial Neural Network, K- Nearest Neighbour, Naive Bayes, Decision Tree [8-10] and so on. Some of the traditional feature selection methods that were used in this study includes Relief, Minimal Redundancy Maximal Relevance (mRMR), [11] Least Absolute Shrinkage Selection Operator (LASSO) [12-14], and Local Learning. It was determined that the performance of the system was superior to that of the approaches that were already in use, particularly in terms of precision. [15]

III. A. DATASET DETAILS: A COMPREHENSIVE HEART DISEASE DATASET

This dataset marks a significant advancement in heart disease research by consolidating five previously independent datasets into a single, comprehensive resource. By integrating these diverse datasets based on eleven common features, it has turned into one of the largest-scale databases of heart diseases that can be used in analyses at the moment, offering the potential for more robust analyses and generalizable findings in heart disease prediction studies.

The five source datasets and their respective sample sizes are as follows: 303 observation was in Cleveland, 294

observation in Hungarian 123 in Switzerland, 200 in Long Beach VA and 270 in Stalog (Heart) Data Set.

The dataset includes 14 features, each providing critical information relevant to heart disease prediction:

Age: Heart disease risk rises with age, therefore patient age is important.

Important feature: Encoded 1 for male and 0 for female since heart disease risk and symptoms vary by gender.

cp chest pain type: Typical angina chest pain due to decreased cardiac blood flow, atypical angina, non-anginal pain (esophageal spasms), and asymptomatic chest pain without disease.

- Trestbps: Hospital admission resting blood pressure in mm Hg, a heart disease risk factor.

- High serum cholesterol (chol) levels (mg/dl) increase artery plaque and heart disease risk.

- Fasting blood sugar (fbs) above 120 mg/dl may indicate diabetes and heart disease risk.

- Restecg shows normal, ST-T wave anomalies, and probable or definite left ventricular hypertrophy. These findings may indicate heart disease.

Patient's highest heart rate reflects cardiovascular fitness and heart health.

Exercise-induced Angina (exang) tests whether the patient has chest pain during activity (1 = yes; 0 = no), which may suggest coronary artery disease.

- ST Depression (oldpeak): Compares exercise-related cardiac stress to rest on an electrocardiogram.

Maximum exercise slope ST Segment: Upslope, flat, or downslope depending on heart disease severity and kind.

The number of main arteries (0-3) during fluoroscopic screening shows coronary artery disease progression.

Blood disease thalassaemia can compromise cardiac function and increase risk. Normal, fixed, or reversible.

Heart Disease: The categorical variable indicating heart disease. The large dataset will help researchers and data scientists build and validate heart disease machine learning models.

IV. ALGORITHMS AND TECHNIQUES

1. Logistic Regression:

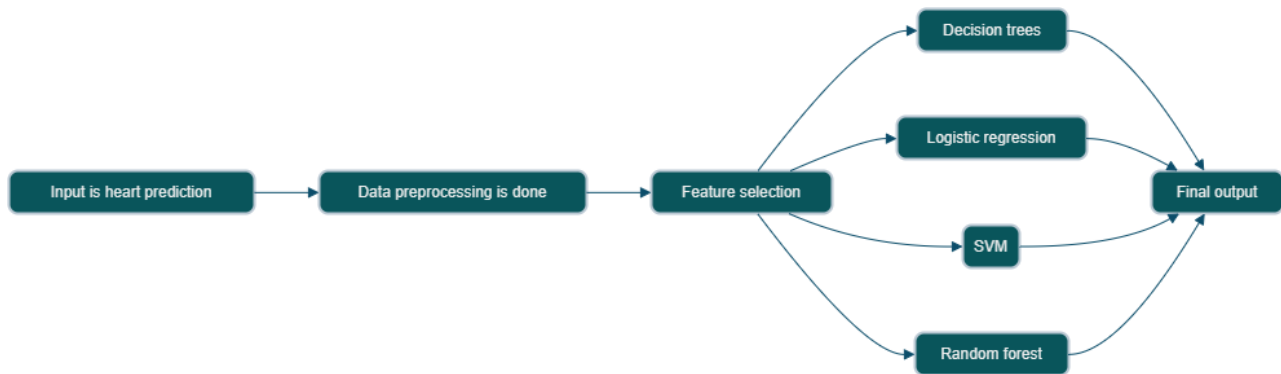


Fig. 2. Overview of proposed work

Logistic regression plays a crucial role in predictive analytics, especially in solving classification problems with binary outcomes. Its widespread application spans various fields such as social sciences, healthcare, marketing, and finance. This statistical method is highly regarded for its ability to model complex interactions between input variables and binary outcomes, not only offering predictions but also providing probabilistic insights. In healthcare, logistic regression has become a valuable tool for disease prediction, striking a balance between technical accuracy and interpretability.

In the context of heart disease prediction, logistic regression is particularly powerful due to its ability to accurately predict outcomes while remaining interpretable. It models the probability of a patient having heart disease based on a set of input features, providing both a binary classification and insights into the relative significance of each feature in the diagnostic process. This interpretability is crucial in medical settings, where understanding the reasoning behind a prediction is important. Logistic regression allows for the quantification of the impact of various risk factors on the likelihood of developing heart disease, making it a valuable asset in developing evidence-based diagnostic tools. Its straightforward yet effective approach makes logistic regression a key player in efforts to enhance early detection and risk assessment in cardiovascular health.

$$f(x) = 1 / (1 + e^{(-x)}) \quad (1)$$

B. Support Vector Machines (SVM):

Support Vector Machines (SVM) represent a sophisticated class of machine learning algorithms that have revolutionized pattern recognition and classification. In heart disease prediction, SVM stands out as a highly effective method, particularly when dealing with the complexities of medical data. SVM operates on the principle of finding an optimal hyperplane that best separates different classes in a high-dimensional feature space. In heart disease prediction, these classes typically represent the presence or absence of cardiovascular disease. The "support vectors" are the data points closest to the decision boundary, and the algorithm

utilizes them to maximize the margin of separation between classes. This approach not only facilitates accurate classification but also enhances the model's generalization

ability, making it robust when applied to new, unseen data.

One of SVM's most notable strengths is its ability to handle high-dimensional data effectively. Medical datasets, such as those used in heart disease prediction, often comprise a wide range of variables, from basic patient demographics to complex physiological parameters. While many traditional algorithms struggle with high-dimensional data, SVM excels by navigating these dimensions to identify the most informative features, thereby overcoming the "curse of dimensionality" that hampers other machine learning algorithms. Moreover, SVM is adept at capturing complex, nonlinear relationships between variables, a skill particularly valuable in medical diagnostics where interactions between health indicators are often intricate and non-linear. SVM achieves this through the use of kernel functions, mathematical tools that allow the algorithm to operate in high-dimensional spaces without explicitly computing the coordinates. Common kernel functions, such as polynomial and radial basis function (RBF) kernels, enable SVM to generate nonlinear decision boundaries, leading to more nuanced and accurate classifications in complex datasets.

In heart disease prediction, SVM can effectively distinguish between patients with and without the disease by analyzing multiple risk factors simultaneously, such as age, blood pressure, cholesterol levels, and ECG readings. This multi-factor approach aligns with the complex nature of cardiovascular disease, where numerous interrelated factors contribute to overall risk.

Combining SVM with other algorithms, such as logistic regression, in a hybrid model offers a strategic approach to heart disease prediction. While logistic regression provides clear interpretability and insights into each feature's contribution, SVM complements this by handling nonlinear relationships and excelling in high-dimensional spaces. This combination leverages the strengths of both algorithms, resulting in a more comprehensive and effective diagnostic tool for assessing cardiovascular health.

As research in this field progresses, integrating SVM with advanced techniques, such as deep learning and genetic

algorithms, presents new opportunities for enhancing the accuracy and utility of heart disease prediction models. This ongoing evolution of machine learning methods holds the potential to significantly impact early detection and management strategies in cardiovascular health, ultimately saving lives through more timely and precise diagnoses. Support Vector Machines (SVM) represent a sophisticated class of machine learning algorithms that have revolutionized pattern recognition and classification. In heart disease prediction, SVM stands out as a highly effective method, particularly when dealing with the complexities of medical data.

SVM operates on the "principle of finding an optimal hyperplane that best separates different classes in a high-dimensional feature space. In heart disease prediction, these classes typically represent the presence or absence of cardiovascular disease. The "support vectors" are the data points closest to the decision boundary, and the algorithm utilizes them to maximize the margin of separation between classes. This approach not only facilitates accurate classification but also enhances the model's generalization ability, making it robust when applied to new, unseen data." [7]

One of SVM's most notable strengths is its ability to handle high-dimensional data effectively. Medical datasets, such as those used in heart disease prediction, often comprise a wide range of variables, from basic patient demographics to complex physiological parameters. While many traditional algorithms struggle with high-dimensional data, SVM excels by navigating these dimensions to identify the most informative features, thereby overcoming the "curse of dimensionality" that hampers other machine learning algorithms.

Moreover, SVM is adept at capturing complex, nonlinear relationships between variables, a skill particularly valuable in medical diagnostics where interactions between health indicators are often intricate and non-linear. SVM achieves this through the use of kernel functions, mathematical tools that allow the algorithm to operate in high-dimensional spaces without explicitly computing the coordinates. Common kernel functions example, radial basis function (RBF) kernels and polynomial, enable SVM to generate nonlinear decision boundaries, leading to more nuanced and accurate classifications in complex datasets.

In heart disease prediction, SVM can effectively distinguish between patients with and without the disease by analyzing multiple risk factors simultaneously, such as age, blood pressure, cholesterol levels, and ECG readings. This multi-factor approach aligns with the complex nature of cardiovascular disease, where numerous interrelated factors contribute to overall risk.

Combining SVM with other algorithms, such as logistic regression, in a hybrid model offers a strategic approach to heart disease prediction. While logistic regression provides clear interpretability and insights into each feature's contribution, SVM complements this by handling nonlinear relationships and excelling in high-dimensional spaces. This

combination leverages the strengths of both algorithms, resulting in a more comprehensive and effective diagnostic tool for assessing cardiovascular health.

As research in this field progresses, integrating SVM with advanced techniques, such as deep learning and genetic algorithms, presents new opportunities for enhancing the accuracy and utility of heart disease prediction models. This ongoing evolution of machine learning methods holds the potential to significantly impact early detection and management strategies in cardiovascular health, ultimately saving lives through more timely and precise diagnoses..

C. Random Forest:

Random Forest is a powerful ensemble learning technique that has gained significant traction in various fields, including medical diagnostics and, more specifically, heart disease prediction. This method builds on the concept of decision trees, combining their simplicity and interpretability with the robustness of ensemble methods to create a highly accurate and versatile prediction model.

The underlying principle of Random Forest is the "wisdom of crowds," which posits that a collective decision made by a group of uncorrelated models often outperforms that of any individual model. During training, RF generates a large number of decision trees and aggregates their predictions to determine the final output. For classification tasks like heart disease prediction, the model predicts the class that receives the majority vote from the individual trees.

Key components of Random Forest include:

Decision Trees: These binary tree structures make decisions based on different dataset attributes, such as age, blood pressure, and cholesterol levels, in the context of heart disease prediction.

Bagging (Bootstrap Aggregation): "This technique creates multiple subsets of the original data through random sampling with replacement. Each decision tree in the forest is trained on a different bootstrap sample, introducing diversity among the trees." [6]

Random Feature Selection: At each node of a tree, only a random subset of features is considered for splitting, further enhancing the diversity of the trees.

Voting/Averaging: For heart disease classification, each tree "votes" for a class, and the majority vote determines the model's prediction.

The Random Forest process for heart disease prediction involves several steps:

Data Preparation: The heart disease dataset, containing various patient attributes and their corresponding diagnoses (presence or absence of heart disease), is prepared for analysis.

Bootstrap Sampling: Multiple subsets of the original data are generated using random sampling with replacement, with each subset containing approximately 63.2% of the original data points.

Tree Creation: A decision tree is built for each bootstrap sample, where a random subset of features is chosen for splitting at each node, and the tree is grown to its full depth or until a stopping criterion is met.

Prediction: To classify a new patient, their data is passed through all trees in the forest, with each tree voting for a class.

The class with the most votes becomes the model's prediction.

The underlying principle of Random Forest is the "wisdom of crowds," which posits that a collective decision made by a

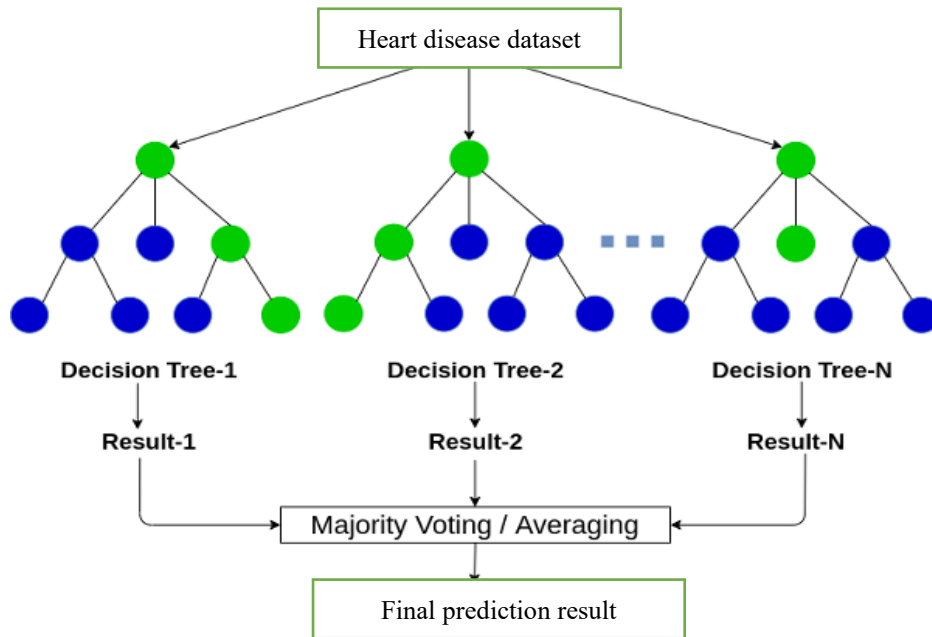


Figure3: Ensemble Learning Using Decision Trees for Heart Disease Prediction

Random Forest offers several advantages for heart disease prediction:

Handling Non-linearity: Random Forest can capture complex, nonlinear relationships between features, which is crucial in heart disease prediction, where risk factors often interact in intricate ways.

Feature Importance: The algorithm provides insights into the importance of different features, helping to identify the most critical factors, such as blood pressure or cholesterol, in predicting heart disease.

Robustness to Outliers and Noise: Due to its ensemble nature, Random Forest is less sensitive to outliers and noise in the data.

Handling Missing Data: Random Forest can effectively manage missing values, a common occurrence in medical datasets.

Reduced Overfitting: The ensemble approach, combined with random feature selection, helps mitigate overfitting, a common challenge in medical data analysis.

Despite its strengths, Random Forest does have some challenges and considerations:

Interoperability & Computational Cost: The algorithm can be computationally intensive and inter operable, especially with large datasets or a high number of trees.

Tuning: RF generally performs well with default parameters, fine-tuning hyperparameters can enhance performance in specific applications, such as heart disease prediction.

It is a powerful ensemble learning technique that has gained significant traction in various fields, including medical diagnostics and, more specifically, heart disease prediction. This method builds on the concept of decision trees, combining their simplicity and interpretability with the robustness of ensemble methods to create a highly accurate and versatile prediction model.

group of uncorrelated models often outperforms that of any individual model. During training, Random Forest generates a large number of decision trees and aggregates their predictions to determine the final output. For classification tasks like heart disease prediction, the model predicts the class that receives the majority vote from the individual trees. Key components of Random Forest include:

Decision Trees: These binary tree structures make decisions based on various dataset attributes, such as age, blood pressure, and cholesterol levels, in the context of heart disease prediction.

Bagging (Bootstrap Aggregation): "This technique creates multiple subsets of the original data through random sampling with replacement. Each decision tree in the forest is trained on a different bootstrap sample, introducing diversity among the trees." [6]

Random Feature Selection: At each node of a tree, only a random subset of features is considered for splitting, further enhancing the diversity of the trees.

Voting/Averaging: For heart disease classification, each tree "votes" for a class, and the majority vote determines the model's prediction.

The Random Forest process for heart disease prediction involves several steps:

Data Preparation: The heart disease dataset, containing various patient attributes and their corresponding diagnoses (presence or absence of heart disease), is prepared for analysis.

Bootstrap Sampling: Multiple subsets of the original data are generated using random sampling with replacement, with each subset containing approximately 63.2% of the original data points.

Tree Creation: A decision tree is built for each bootstrap sample, where a random subset of features is chosen for

splitting at each node, and the tree is grown to its full depth or until a stopping criterion is met.

Prediction: To classify a new patient, their data is passed through all trees in the forest, with each tree voting for a class. The class with the most votes becomes the model's prediction.

Random Forest offers several advantages for heart disease prediction:

Handling Non-linearity: "Random Forest can capture complex, nonlinear relationships between features, which is crucial in heart disease prediction, where risk factors often interact in intricate ways."

Feature Importance: This algorithm provides an understanding about the importance of different features, helping to identify the most critical factors, such as blood pressure or cholesterol, in predicting heart disease."

Random Forest Accuracy: 0.900709219858156
Random Forest Classification Report:

	precision	recall	f1-score	support
0	0.91	0.88	0.89	67
1	0.89	0.92	0.91	74
accuracy			0.90	141
macro avg	0.90	0.90	0.90	141
weighted avg	0.90	0.90	0.90	141

Random Forest Confusion Matrix:
[[59 8]
[6 68]]

Fig. 4. Random forest Model Performance Metrics for Heart Disease Classification.

Robustness to Outliers and Noise: Due to its ensemble nature, Random Forest is less sensitive to outliers and noise in the data.

Handling Missing Data: Random Forest can effectively manage missing values, a common occurrence in medical datasets.

Reduced Overfitting: The ensemble approach, combined with random feature selection, helps mitigate overfitting, a common challenge in medical data analysis.

Despite its strengths, Random Forest does have some challenges and considerations:

Interpretability: While Random Forest provides feature importance, the model itself is less interpretable than a single decision tree or logistic regression.

Computational Cost: The algorithm can be computationally intensive, especially with large datasets or a high number of trees.

Tuning: Although Random Forest generally performs well with default parameters, fine-tuning hyperparameters can enhance performance in specific applications, such as heart disease prediction.

V. RESULTS AND DISCUSSIONS:

For the purpose of evaluation the following factors are considered .

$$precision = TP / (TP + FP) \quad (1)$$

$$recall = TP / (TP + FN) \quad (2)$$

$$F1 = 2 * (precision * recall) / (precision + recall) \quad (3)$$

Accuracy: 0.8794326241134752
Classification Report:

	precision	recall	f1-score	support
0	0.88	0.87	0.87	67
1	0.88	0.89	0.89	74
accuracy			0.88	141
macro avg	0.88	0.88	0.88	141
weighted avg	0.88	0.88	0.88	141

Confusion Matrix:
[[58 9]
[8 66]]

Fig. 5. LR Model Performance Metrics for Heart Disease Classification"

$$accuracy = (TP + TN) / (TP + FN + TN + FP) \quad (4)$$

$$specificity = TN / (TN + FP) \quad (5)$$

1. Logistic Regression (LR):

The logistic regression model achieved an accuracy of 87%, with a precision of 88%, indicating its strong capability in correctly identifying true positive cases among the predicted positive results. The model's recall rate was also 87%, reflecting its effectiveness in accurately identifying patients with heart disease. Additionally, the F1-score, which represents the harmonic mean of precision and recall, reached 89%.

2. Support Vector Machine (SVM)

The Support Vector Machine (SVM) model exhibits robust performance across several key metrics. It achieved an accuracy of 88%, which is slightly better than that of the logistic regression model. The precision of the SVM model is 90%, indicating a high rate of correctly identifying positive cases. With a recall rate of 92%, the model demonstrates its capability to effectively identify true positives. Additionally, the F1 score for the SVM is 92%, reflecting a balanced performance with precision and recall rates being nearly

SVM Accuracy: 0.8794326241134752
SVM Classification Report:

	precision	recall	f1-score	support
0	0.90	0.84	0.87	67
1	0.86	0.92	0.89	74
accuracy			0.88	141
macro avg	0.88	0.88	0.88	141
weighted avg	0.88	0.88	0.88	141

SVM Confusion Matrix:
[[56 11]
[6 68]]

Fig. 6. SVM Model Performance Metrics for Heart Disease Classification.

equal.

3. Random Forest:

The model demonstrated a high level of accuracy, achieving a correctness rate of 90%. It also exhibited impressive precision with a true-positive rate of 91%, indicating robust performance in prediction tasks. The recall of 92% highlights the model's strong capability in identifying

patients with heart disease. As a result, the F1 score reached 92%, reflecting an overall exceptional performance in predictive accuracy.

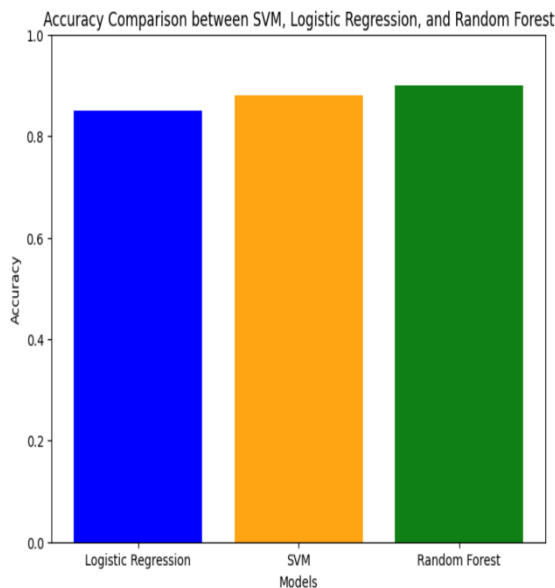


Fig. 7. Performance Comparison

Comparison of Models:

Among all the models, Random Forest yielded the highest accuracy, precision, recall and F1-score, hence, was deemed most accurate in heart disease prediction. SVM also did very well, especially in terms of accuracy and recall though worse than the accuracy of the Random Forest classification method. Among the three models, Logistic Regression was the least accurate, however, it was still rather satisfactory with fairly good precision against recall. From these outcomes, it can be pointed out that all three algorithms work effectively in terms of prediction the heart disease but the Random Forest algorithm can be presented in terms of high accuracy and ability to detect the true cases of heart diseases. This means that by employing the Random Forest as an ensemble method can greatly improve the accuracy of the predictions of heart disease hence may improve patients' outcome due to early detection in Clinical Medicine practice.

VI. CONCLUSION:

To summarize, heart disease continues to be the top cause of death worldwide, accounting for over 17 million deaths each year. Early identification is important for halting the progression of this disease and saving lives. Our findings show that machine learning (ML) and artificial intelligence (AI) can help improve early identification and risk assessment for cardiac disease. The classification models we created, including Logistic Regression (LR) and Support Vector Machines (SVM), can diagnose cardiac disease at an early stage with high accuracy. The datasets used produced exact results, highlighting the need of ongoing research and interdisciplinary collaboration among data scientists and

healthcare practitioners. Future developments should focus on incorporating larger and more diverse datasets, as well as investigating advanced ML algorithms to improve model performance even more. Furthermore, extensive collaboration with clinical professionals, as well as thorough validation in real-world clinical settings, will be required to ensure the predictive models' dependability and practical application. By continuing to progress this science, we will be able to dramatically improve heart disease prediction and save many lives.

REFERENCES:

- [1] R. T. Selvi and I. Muthulakshmi, An optimal artificial neural network based big data application for heart disease diagnosis and classification model, *J. Ambient Intell. Humanized Comput.*, vol. 12, no. 6, pp. 61296139, 2021.
- [2] Senthil Prediction in Kumari, [2022], Machine Learning Heart Attack Environment, <https://www.researchgate.net/publication/362630808>.
- [3] Mert Ozcan, [2023], Healthcare Analytics <https://doi.org/10.1016/j.health.2022.100130>.
- [4] Rina S. Patil, [2021], Prediction of Heart Disease Severity Measurement Using Deep Learning Techniques, <https://doi.org/10.1177/2048004016645467>.
- [5] Joy, Helen K; Kounte, Manjunath R; ,Deep learning-based switchable network for in-loop filtering in high efficiency video coding, *International Journal of Electrical and Computer Engineering (IJECE)*, 13, 4, 4640-4648, 2023, IJECE
- [6] Chao, C.; Peiliang, Z.; Min, Z.; Qu, Y.; Bo, J. Constrained transformer network for ecg signal processing and arrhythmia classification. *BMC Med. Inf. Decis. Mak.* 2021, 21, 184.
- [7] Joy, Helen Kottarathil; Kounte, Manjunath Ramachandra; ",A Comprehensive Review of Traditional Video Processing," *Advances in Science, Technology and Engineering Systems Journal*", 5, 6, 272-279, 2020,
- [8] Shubhanshi et al., *American International Journal of Research in Science, Technology, Engineering & Mathematics*, 23(1), June-August, 2018, pp. 257-261
- [9] S. Mohan, C. Thirumalai, and G. Srivastava, Effective heart disease prediction using hybrid machine learning techniques, *IEEE Access*, vol. 7, pp. 8154281554, 2019.
- [10] Joy, Helen K; Kounte, Manjunath R; Joy, Ajin K; ,Deep learning approach in intra-prediction of high efficiency video coding, "2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE)", , 134-138, 2020, IEEE
- [11] Joy, Helen K; Kounte, Manjunath R; ,Deep CNN depth decision in intra prediction, *Proceedings of International Conference on Power Electronics and Renewable Energy Systems: ICPERES 2021*, , 1-10, 2022, Springer Singapore
- [12] A. Gupta, R. Kumar, H. S. Arora, and B. Raman, MIFH: A machine intelligence framework .
- [13] K. Feng, X. Pi, H. Liu, and K. Sun, Myocardial infarction classification based on convolutional neural network and recurrent neural network, *Appl. Sci.*, vol. 9, no. 9, p. 1879, May 2019.
- [14] [9] A. Singh and A. Payal, "CAD diagnosis by predicting stenosis in arteries using data mining process," *Intell. Decis. Technol.*, vol. 15, no. 1, pp. 59-68, Mar. 2021.
- [15] Vahini B, Sanjeev S, Narenthiran CK, Chandrasekar K (2021) A review on rheumatic heart disease. *Curr Asp Pharm Res Dev* 5:33-42.