Prediction of Cardiovascular Disease Using Machine Learning Model

1st I R Oviya
Department of Artificial Intelligence Engineering,
Amrita School of Computing,
Amrita Vishwa Vidyapeetham
Chennai,India

2nd Jerome Santiago J
Department of Artificial Intelligence Engineering,
Amrita School of Computing,
Amrita Vishwa Vidyapeetham
Chennai,India

3rd Shashidhar R
Department of Artificial Intelligence Engineering,
Amrita School of Computing,
Amrita Vishwa Vidyapeetham
Chennai,India

4th Sri Bhuvana Sankar T
Department of Artificial Intelligence Engineering,
Amrita School of Computing,
Amrita Vishwa Vidyapeetham
Chennai,India

5rd Vignesh S
Department of Artificial Intelligence Engineering,
Amrita School of Computing,
Amrita Vishwa Vidyapeetham
Chennai,India

6th Adhithiya MS
Department of Artificial Intelligence Engineering,
Amrita School of Computing,
Amrita Vishwa Vidyapeetham
Chennai,India

***Abstract*—According to a variety of diagnostic factors, this research examines how well machine learning algorithms predict cardiovascular disease (CVD). A collection of heart and blood vessel conditions called cardiovascular disease (CVD) is brought on by a person's genetic makeup, environment, and way of life. Chest discomfort, high blood pressure, high cholesterol, fasting blood sugar, resting ECG findings, maximal heart rate during exercise, and exercise-induced angina are typical warning signs. When paired with additional variables like age, gender, and family history, these characteristics are utilized in predictive models to evaluate a person's CVD risk. This allows healthcare professionals to determine the threat and suggest suitable measures to lower the risk of CVD. Every algorithm is trained and tested on various patient datasets to provide prediction models. Techniques for choosing features are applied to determine the most important characteristics for CVD prediction. Metrics including accuracy, precision, recall, and F1 score are used to evaluate the models' efficiency. This illustrates how differently the algorithms execute, with each exhibiting qualities in particular areas of prediction. KNN well captures local patterns, while SVM best handles complicated data and strong discriminatory power. Both Logistic Regression and Naive Bayes offer interpretable predictions and insights into the relevance of particular features. The research highlights the significance of assessing the efficiency of models across an extensive range of measures, such as recall, accuracy, and F1 score, to obtain a more sophisticated comprehension of predictive skills. These metrics render it possible to evaluate the algorithm's predicted accuracy as well as its accuracy in identifying positive CVD cases.**

***Index Terms*—Cardiovascular Disease, K-Nearest Neighbors method, Support Vector Machine, Naive Bayes, Logistic Regression.**

I. Introduction

Cardiovascular diseases (CVDs) remain to be one of the world's main causes of death and morbidity, making them significant threats to global public health. The prevalence of CVDs is rising despite advancements in medical research and healthcare infrastructure, placing a strain on economies and healthcare systems alike. Accurate risk assessment, timely intervention, and early detection are still essential for controlling and reducing CVD-related consequences. Traditional approaches to assessing the risk of CVD have primarily depended on well-established prediction models such as the Framingham Risk Score, which incorporate clinical and demographic variables to calculate an individual's risk of cardiovascular events over a specified period. Although these models have shown to be useful in identifying high-risk individuals within the general population, they frequently falter in providing personalized risk assessments because they do not account for the complex interactions between genetic predispositions, environmental factors, and lifestyle factors that raise the risk of CVD. Using innovative computational methods, especially machine learning (ML), to create more accurate and customized predictive models for CVD risk assessment has gained popularity in recent years. Machine learning algorithms can examine large and diverse datasets, interpret complex patterns, and create prediction models that can capture the complex aspects of cardiovascular disease risk. The goal of incorporating machine learning (ML) into cardiovascular disease (CVD) risk prediction comes from its ability to combine a wide range of patient-specific information, including clinical biomarkers, medical histories, genetic predispositions, lifestyle decisions, and environmental exposures, to produce customized risk assessments. By utilizing machine learning (ML), scientists and medical professionals hope to improve the precision, depth, and clinical applicability of CVD risk prediction models. This will allow for more specialized preventative care and individualized treatment plans. This study's main goal is to investigate the use of machine learning (ML) techniques in the prediction of cardiac disease, with a focus on developing reliable and practically useful predictive models. Using large-scale datasets drawn from various patient cohorts, medical environments, and geographic locations, we aim to train and evaluate machine learning algorithms to predict the beginning, course, and severity of cardiovascular diseases. Our goal is to uncover important predictors, improve model performance, and provide new viewpoints on the complex etiology and pathophysiology of heart disease through the rigorous assessment and development of predictive models. This research is important because it has the potential to completely change how CVD risk assessment and management are currently done. Predictive models powered by machine learning (ML) have the potential to provide physicians with more accurate and customized risk assessments, allowing for more focused interventions and customized preventive measures. The work carried out is significant because it has the potential to completely change how CVD risk assessment and management are currently done. Predictive models based on machine learning (ML) have the potential to provide physicians with more accurate and customized risk assessments, allowing for more focused interventions and customized preventive measures. Healthcare practitioners can reduce the burden of heart disease and improve patient outcomes by identifying individuals at heightened risk of CVDs early on and implementing preventive measures such as medication adjustments, lifestyle modifications, and patient education. Predictive models based on machine learning also can guide resource allocation plans and population-level health interventions. Public health authorities can reduce the overall burden of heart disease in communities by implementing targeted screening and intervention programs, allocating resources optimally, and identifying high-risk populations and geographic locations with elevated risk of CVD. In conclusion, the integration of machine learning methods into the risk assessment of cardiovascular disease signals a positive step forward in the direction of customized therapy and improved cardiovascular health outcomes. ML-driven predictive models have the potential to revolutionize current practices, provide clinicians with practical insights, and ultimately aid in the prevention and management of CVDs worldwide by leveraging the abundance of data available within modern healthcare systems.

II. Literature Review

There is a rising interest in using machine learning (ML) algorithms for early detection and prediction as cardiovascular diseases (CVDs) become more common. Jindal et al. (2021) concentrate on creating a heart disease prediction system (HDPS) by machine learning approaches, stressing the significance of precise and effective disease prediction because of the difficulties in detecting cardiac disorders. They highlight other research that looked at the effectiveness of several machine learning (ML) algorithms to forecast cardiac illnesses using clinical variables, such as logistic regression, K-nearest neighbors (KNN), and Random Forest Classifier. The authors emphasize the importance of data choice and data pretreatment in model prediction optimization. The refinement of input data for machine learning algorithms, which guarantees accuracy and generalization, is largely dependent on techniques like data cleaning, normalization, and feature extraction. Efficient training and validation of prediction models require access to high-quality datasets with complete medical records. The significance of early identification and treatment in reducing the negative consequences of CVDs is highlighted by ML-based heart disease detection. The accuracy and effectiveness of conventional diagnostic techniques, such as electrocardiography (ECG), are limited. By utilizing intricate datasets to detect trends and forecast the course of diseases, machine learning algorithms present a viable substitute. The promise of Naive Bayes with a balanced method, support vector machine (SVM) with XGBoost, and an upgraded SVM using a dual optimization methods to increase the accuracy of heart disease identification is highlighted by Nagavelli et al. (2022). Recent research has made great progress in combining machine learning and deep learning methods to increase the effectiveness of prediction models. Although Bharti et al. (2021) showed outstanding accuracy rates, their research had some noteworthy faults, including a lack of openness about the dataset utilized and a lack of a full description of the reasoning underlying their decisions and how they affected the efficacy of the model. The significance of AI-based methods, particularly ML integration with deep learning, was emphasized by Subramani et al. (2023) about CVD prediction. Their suggested models proved to be more accurate than those using conventional methods. To further improve the efficacy of prediction models, future research in this subject should concentrate on using bigger and more varied data from numerous medical institutions. The forecasting of heart disease through the use of deep learning (DL) and machine learning (ML) approaches has been thoroughly investigated. While Golande and T (2019) discovered random forest to perform best with 92.5 Methods for feature selection and optimization help to enhance the performance of models. The goal of this research is to create an enhanced CNN model for the precise categorization of heart disease. The study by Diaa Salama et al emphasizes how crucial early identification is to lessen the effects of cardiac disease. The authors show how ML approaches may be used to increase prediction accuracy, which might lead to better patient outcomes and more immediate intervention. The use of several datasets and ML algorithms, such as k-Nearest demonstrates a comprehensive investigation of predictive techniques Neighbor, Gradient Boosting, Random Forest, Na¨ıve Bayes, Decision Tree, and Logistic Regression. Numerous research works have looked at machine learning methods for predicting cardiac disease. The Hybrid Random Forest with Linear Model (HRFLM) was first presented by Mohan et al. (2019), which achieved an accuracy of 88.7 The ability of machine learning and deep learning approaches to foresee and identify cardiovascular disease or CVD, a major global cause of death has been demonstrated by a recent study. Across a range of machine learning (ML) and deep learning (DL) classifiers, Avvaru et al. (2022) achieved the greatest accuracy of 95.7 Multiple studies have used datasets such as the Cleveland heart disease dataset and the Pima Indian diabetes dataset to investigate machine learning algorithms for heart disease prediction.

Hybrid Random Forest with Linear Model (HRFLM), developed by Ramesh Ponnala et al. (2021), combines random forest with a linear model to achieve 100

III. THE FRAME WORK

A. Dataset

There are 1025 rows and 14 columns in this cardiac data. It includes information on age, sex, blood pressure, cholesterol, kind of chest discomfort, fasting blood sugar, and resting ECG readings. The target factor, which indicates the existence of cardiac disease, is in the final section. Under the other characteristics and predictors, it will be utilized to develop patterns of classification that will forecast heart disease.

B. Comparison of Machine Learning Models for Heart Disease Prediction

Machine learning models such as Support Vector Machine (SVM), Naive Bayes, Logistic Regression, and K-Nearest Neighbors (KNN) were assessed in comparative research for their ability to predict cardiac disease. SVM received 78% on training data and 76% on test data, to maximize the margin between classes. Assuming substantial feature independence, Naive Bayes performed the worst, achieving scores of 76% and 74% on the training and test sets, respectively. On training and test sets, the accuracy of logistic regression was 78% and 76%, respectively. KNN achieved 76% and 74% accuracy on training and test sets, respectively, by depending on nearest neighbors. Because of Random Forest’s strong generality and accuracy, it was the best performer overall.

Machine learning models, such as K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Logistic Regression, and Naive Bayes, were compared with the dataset. Scikit-learn was used for implementing these models. The simplest algorithm is KNN, which predicts new data using the k nearest training instances. Although it takes a lot of memory, it is relatively simple to learn and use. Using a high-dimensional feature space as a map, SVM determines the best separation hyperplane between groups. Although logistic regression implies linear correlations among variables, it fits data to a logistic function to determine class likelihood. Depending on the Bayes theorem, the Naive Bayes classifier is a straightforward probabilistic algorithm that presumes variable independence. When compared to other models, KNN produced the greatest accuracy of 82% on test data for the heart disease dataset. KNN’s non-parametric character, which enables it to represent interactions that are nonlinear better than logistic regression, is one of the main factors contributing to its success.

Because parameters like age, sex, and other medical considerations rely on one another, Naive Bayes performs badly in these situations. Logistic regression and SVM make assumptions about linear connections that might not be entirely true. As is often the case with medical data, KNN makes no assumptions and functions effectively even in situations where noise and sampling are problems. Finally, without adjusting any hyperparameters, the straightforward non-parametric K-Nearest Neighbors method produced the greatest results, demonstrating that fundamental learning strategies function best when their underlying presumptions are in line with the task at hand.

C. KNN Model Implementation

The algorithm explains the KNN model, the heart disease dataset is processed, its performance is assessed, and forecasts are made depending on user input. The first step is to load the required libraries, such as sklearn for machine learning features, pandas for data processing, and numpy for numerical calculations. A Pandas DataFrame (heart data) containing the heart disease data is imported and divided into features (X) and target variables (Y). Each of the columns makes up the characteristics, except the ’target’ column, which has a binary classification label indicating whether cardiovascular disease is present or absent.

D. KNN Model Training and Evaluation

The K-nearest neighbors (KNN) classifier divides data at random to anticipate cardiovascular disease. The KNN algorithm is a simple and effective method for applications involving classification. Using the training data, the model is taught to recognize patterns and provide forecasts. The efficiency score method is used to evaluate the models post-training to calculate the accuracy of the predictions. The user is prompted by the algorithm to provide values for heart-related variables such as age, sex, and cholesterol level. Resize (1, -1) is used to assemble the input provided by the user into a list and then convert it into a NumPy array. The person’s cardiovascular status is then ascertained by loading the previously saved model from the file. The goal variable’s distribution of classes in each set is guaranteed by this approach.

E. Figures and Table

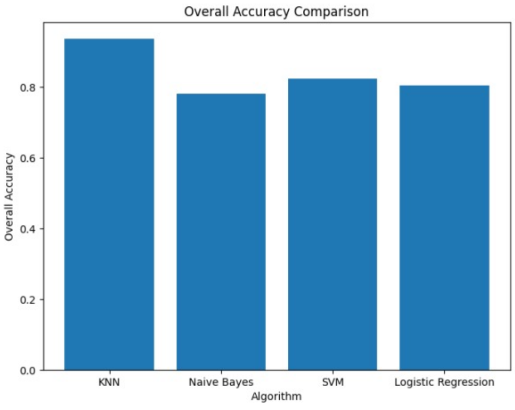

Fig. 1. Overall Accuracy

TABLE I
COMPARISON TABLE

S.no	Algorithm	Accuracy on Training Data	F1 Score	Precision	Recall	Overall Accuracy
1	KNN	0.9927	0.9378	0.9423	0.9333	0.9366
2	Naive Bayes	0.8390	0.7907	0.7727	0.8095	0.7805
3	SVM	0.8695	0.8435	0.7760	0.9333	0.8244
4	Logistic Regression	0.8524	0.8230	0.7686	0.8857	0.8049

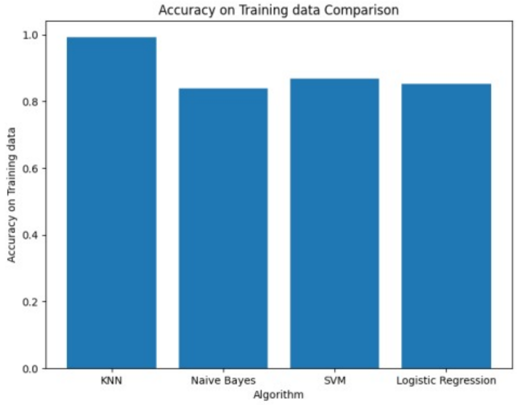

Fig. 2. Accuracy on Training Data Comparsion

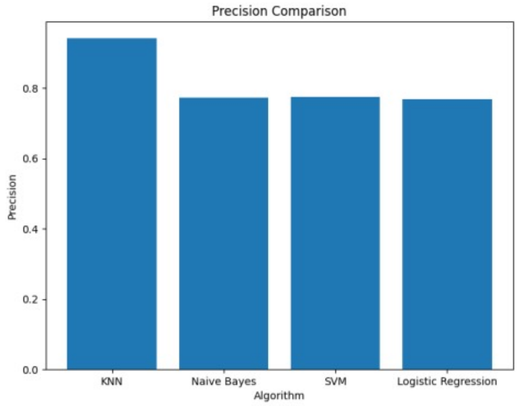

Fig. 3. Precision Comparsion

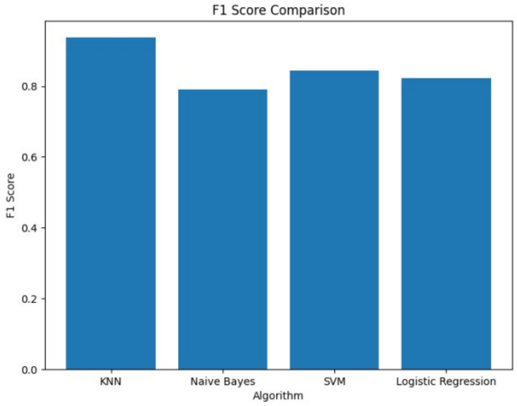

Fig. 4. F1 Score Comparsion

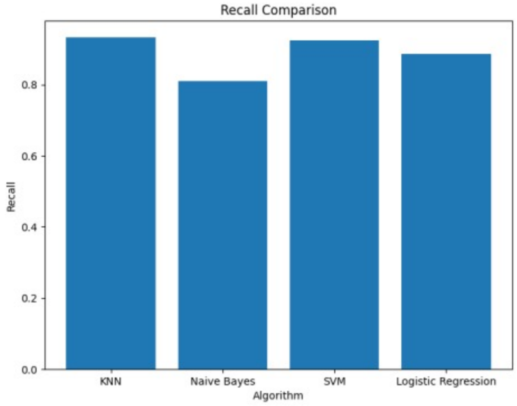

Fig. 5. Recall Comparsion

IV. RESULTS AND CONCLUSION

With the highest F1 score (0.9378), precision (0.9423), recall (0.9333), and total accuracy (0.9366), KNN exhibits exceptional performance. It is especially good at capturing the intricate correlations found in the dataset when the underlying patterns are not clear-cut or linear. Even with an acceptable overall accuracy (0.7805), Naive Bayes is not as good as KNN. It displays a lower recall (0.8095), precision (0.7727), and F1 score (0.7907). Because Naive Bayes relies on the independence of characteristics, it may perform poorly in situations where there are complex dependencies. With a balanced F1 score of 0.8435, SVM performs well in terms of precision (0.7760) and recall (0.9238). It does, however, lag somewhat below KNN in terms of total accuracy (0.8244). SVM works well with non-linear interactions and is renowned for its ability to capture complex decision limits. With a balanced F1 score (0.8230), precision (0.7686), recall (0.8857), and total accuracy (0.8049), logistic regression produces competitive results. Its performance may differ depending on how linear the underlying patterns are, and it is predicated on a linear relationship between features and the target variable. In this comparison, KNN is the best algorithm for predicting Cardiovascular Disease, showing higher performance on all evaluation metrics.

References

[1] Umarani Nagavelli, Debabrata Samanta, Partha Chakraborty, ”Machine Learning Technology-Based Heart Disease Detection

[2] Models”, Journal of Healthcare Engineering, vol. 2022, Article ID 7351061, 9 pages, 2022.

[3] Nazrul Anuar, N., Hafifah, A. H., Mohd Zubir, S., Noraidatulakma, A., Rosmina, J., Nurul Ain, M. Y., . . . Rahman, A. J. (2020). Cardiovascular Disease Prediction from Electrocardiogram by Using Machine Learning. International Journal of Online and Biomedical Engineering (iJOE), 16(07), pp. 34–48.

[4] Rohit Bharti, Aditya Khamparia, Mohammad Shabaz, Gaurav Dhiman, Sagar Pande, Parneet Singh, ”Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning”, Computational Intelligence and Neuroscience, vol. 2021, Article.

[5] Rohit Bharti, Aditya Khamparia, Mohammad Shabaz, Gaurav Dhiman, Sagar Pande, Parneet Singh, ”Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning”, Computational Intelligence and Neuroscience, vol. 2021, Article

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Trans. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] Harshit Jindal et al 2021 IOP Conf. Ser.: Mater. Sci. Eng. 1022 012072

[8] Journal of Computing and Communication Vol.2, No.1, PP. 50-65, 2023 Heart Disease Prediction Using Machine Learning Diaa Salama AbdElminaam, Nada Mohamed, Hady Wael, Abdelrahman Khaled, Adham Moataz

[9] Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques SENTHILKUMAR MOHAN, CHANDRASEGAR THIRUMALAI AND GAUTAM SRIVASTAVA VOLUME 7, 2019 4

[10] Suneetha, A. R. V. N. ., Mahalngam, T. (2022). Cardiovascular Disease Prediction Using ML and DL Approaches. International Journal on Recent and Innovation Trends in Computing and Communication, 10(10), 161–167.

[11] Ramesh Ponnala, K. Sai Sowjanya (2021) Heart Disease Prediction using Machine Learning Techniques. Volume 8, Issue 4 Page Number: 42-47

[12] Nandan Kishor A SMART HEALTHCARE ALGORITHM FOR ANALYSIS AND PREDICTION OF HEART DISEASE USING ML Vol. 03, Issue 06, June 2023, pp: 437-447.