International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Hybrid approach for Recurrence of Cardiovascular Disease Prediction Using Machine Learning Techniques

A Manjesh Gowda[1], Anish H G[2], Dr. Lakshmi K[3]

School of Computer Science and Applications, REVA University, Bengaluru, Karnataka, India

Abstract

Cardiovascular disease remains a global health concern, being a leading cause of mortality. In response to the critical nature of heart-related conditions, this research focuses on the development of smart systems leveraging machine learning algorithms for accurate and timely diagnosis. The study explores various machine learning approaches to predict recurrence of heart diseases based on patient data encompassing key health factors. With heart disease being a prevalent cause of death worldwide, real-time forecasting methods from medical data sources have become crucial.

The implementation of machine learning in healthcare demonstrates its potential for early and precise recurrence of disease detection. Despite the abundance of health information generated by medical institutions, there is an underutilization of this data, leading to a healthcare system that is "data rich" but "knowledge poor." This work aims to address this gap by presenting a reliable recurrence heart disease prediction system. The research highlights the necessity for effective analysis methods to uncover connections and patterns within healthcare data.

The proposed prediction system utilizes a diverse set of health factors, contributing to a comprehensive understanding of heart disease. By leveraging machine learning algorithms, the system aims to enhance the accuracy and efficiency of diagnosis. The findings of this research not only contribute to the advancement of predictive healthcare but also underscore the significance of unlocking valuable insights from the vast pool of medical data. Ultimately, the implementation of such intelligent systems holds promise for improving patient outcomes and reducing the burden of cardiovascular diseases on global healthcare systems.

The paper demonstrated Hybrid methods: Support Vector Machine (SVM), Random Forest (RF), and Naïve Bayes (NB), to build the prediction models. Data preprocessing and feature selection steps were done before building the models. The models were evaluated based on the accuracy, precision, recall, and F1-score. The Random Forest model performed best with 94.27% accuracy.

Keywords: Recurrence Heart Disease Prediction, Machine Learning, Hybrid Approach, Support Vector Machine, Naïve Bayes, Random Forest.

I. INTRODUCTION

Cardiovascular Disease (CVD), commonly referred to as heart disease, encompasses a wide range of conditions that affect the heart, with the two most common conditions being ischemic heart diseases and strokes. The World Health Organization lists the most significant behavioural risk factors for CVD as

maintaining an unhealthy diet, a sedentary lifestyle, tobacco use, and excessive consumption of alcohol. Prolonged exposure to these risk factors can present itself as an initial sign of CVD, which include elevated blood pressure, elevated blood glucose, raised blood lipids, and obesity. Warning signs listed by the American Heart Association include having one or more of the following: shortness of breath, persistent coughing or wheezing, swelling of the ankles and feet, constant fatigue, lack of appetite, and impaired thinking. Efficient early diagnosis can substantially reduce the risk and global burden of CVD by initiating treatment rapidly to prevent further health deterioration. Thus, there is an urgent need to develop machine learning models that can predict the probability of developing CVD depending on the risk factors present. Recently, machine learning models have successfully lent a hand in diverse cases in the medical field [5]. They have been effective in analyzing, evaluating, and predicting different medical conditions [6]. In this paper, we are proposing a machine learning approach to predict the presence of cardiovascular diseases in patients based on major health data.

This paper is organized as follow: Section II covers the related works where machine learning was used for heart disease prediction. Section III explains the methodology, where the dataset is described, pre-processed, and split. As well as the applied algorithms and the corresponding model design parameters, the evaluation metrics selected to evaluate the performance of the model are described. Section IV discusses the experimental results. Finally, in Section V, the remarks and conclusions about this work are presented.

II. RELATED WORK

Heart disease prediction was addressed in the literature using several methods. In [7], Naïve Bayes, SVM, and Functional Trees were used to predict the possibility of heart diseases with an accuracy of 84.5%, using measurements from wearable mobile technologies with the same inputs used in our work. Furthermore, Naïve Bayes was solely used in [8] with a slightly better accuracy of 86.4%, using the same dataset.

Another work [9] used several algorithms; Logistic Regression, KNN, NN, SVM, NB, Decision Tree, and RF, with three feature selection algorithms: Relief, mRMR, and LASSO to predict the existence of heart disease with the same dataset used in this work. The Logistic Regression algorithm had the best performance and yielded predictions with an accuracy as high as 89%.

Moreover, a work done in 2020 [10] applied 4 algorithms with a very high accuracy of 90.8% for the KNN model, and minimum accuracy of 80.3% for the other models.

In [11], a hybrid Random Forest and Naïve Bayes model achieved an accuracy of 84.16% using 10 features, which were selected using Recursive Feature Elimination and Gain Ratio algorithms.

In a recent work done in 2021 [12], Logistic Regression, Random Forest, and KNN were used for the prediction. The maximum accuracy was 87.5%.

All the previous is very promising for the future of heart diseases and failure prediction, especially with the current advances in portable electronic measurement devices.

III. METHODOLOGY

A. Data Collection

The dataset was collected from Kaggle [13]. The dataset contains a total of 2010 instances with 13 attributes as described in Table I.

TABLE I. CARDIOVASCULAR DISEASE DATASET DESCRIPTION

Data Element	Description	Type	Range	Remarks
Gender	Gender of the individual	Nom	Male/Female	-
Age	Age of the individual	Numa	29-77	Average is 52.78
Diabetes mellitus	Presence of diabetes mellitus	Bi	Yes/No	-
Current Smoking	Current smoking status	Bi	Yes/No	30.76% Female, 69.23% Male
BMI	Body Mass Index	Num	-	-
Systolic	Systolic blood pressure	Num	94-200	Average is 133.5
Diastolic	Diastolic blood pressure	Num	-	-
Cholesterol	Cholesterol level	Num	126-417	Average is 237.71
Atrial Fibrillation	Presence of atrial fibrillation	Bi	Yes/No	-
Number of Vascular Beds	Number of major vessels	Nom	0/1/2/3 1: Asymptotic 2: Non-anginal pain 3: Typical angina	-
Cardiovascular Event in Past Year	Previous cardiovascular events	Nom	Yes/No	-
Type of Cardiovascular Treatment	Treatment type for cardiovascular disease	Nom	Medicine/Surgery	-
Recurrence of Heart Disease	Recurrence of heart disease	Bi	Yes/No	-

[a] Numerical, [b] Binary, [c] Nominal.

B. Data Preprocessing

The performance of a machine learning model is greatly determined by the quality of the data used to build it, which makes data preprocessing very important. Data preprocessing includes cleaning the data by removing corrupted or missing data points and outliers, in addition to transforming the data, resampling it, and applying feature selection.

2) Checking for Imbalances

Imbalance in the output can distort the prediction accuracy. Therefore, the balance of the output "target" was verified as shown in Figure 2. After inspection, the data turned out to be balanced with a 10:11 ratio between the two categories.

Fig. 2. Histogram and boxplot of the output "target"

3) Data Transformation

Transformation is applied when the dataset includes data of different formats, or when different datasets are combined. In this case, the nominal features were transformed into factors, for them to be used in Google Colab.

4) Dimensionality Reduction

In machine learning, dimensionality reduction refers to the process of reducing the number of features to decrease the complexity and prevent overfitting, by either feature selection or extraction.

Feature selection is done by selecting a subset of features from the original set, and is done by methods such as CFS (Correlation-based Feature Selection), Chisquared test and ridge regression.

Weka software was used for feature selection as it has several options of attributes evaluator to test and use.

5) Data Splitting

In machine learning, the data is usually split into training and testing sets, where the training set is used to train the model, and the testing set is to test it and predict the output. In this work with 80% of the data used in training and 20% used for testing.

C. Applied Algorithms 1) Hybrid approach

A hybrid model in machine learning combines the strengths of multiple individual models to improve predictive performance by leveraging their unique capabilities. Combining Naive Bayes (NB), Support Vector Machine (SVM), and Random Forest (RF) models in a hybrid approach may enhance the overall accuracy and robustness of the predictive model. Each algorithm has its own strengths and weaknesses, and by combining them, the hybrid model may achieve better results than any single model alone.

2) Naïve Bayes (NB)

Naïve Bayes is a supervised learning algorithm, that is based on the Bayes Theorem, and assumes that all features are independent and have equal contribution to the target class. Bayes' theorem calculates the posterior probability of an event A, given some prior probability of event B, as in (3).

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (3)$$

3) Random Forest (RF)

RF is also a supervised machine learning algorithm, used for both classification and regression. It utilizes ensemble learning, which is a technique that combines several classifiers to make accurate predictions in complex situations. RF algorithms establish the prediction based on the results of multiple decision trees through bagging or bootstrap aggregation as shown in Figure 3.

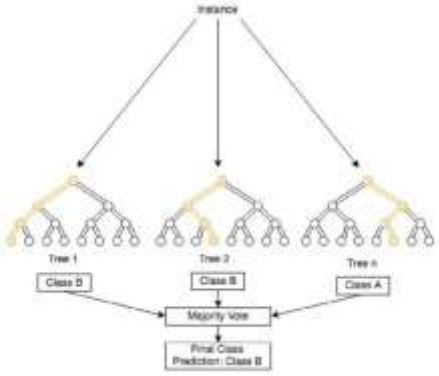

Fig. 3. Random Forest prediction method [14]

When training the model in RStudio, the following parameter were fixed: 500 decision trees, 3 variables try at each split, in classification method.

4) Support Vector Machine (SVM)

SVM is a supervised learning method used for classification, regression, and outlier detection. It seeks to establish a decision boundary between different classes, to label prediction using one or more feature vectors as shown in Figure 5. This decision boundary, known as the hyperplane, is oriented to be as far away as possible from the nearest data points. Those nearest locations are referred to as support vectors.

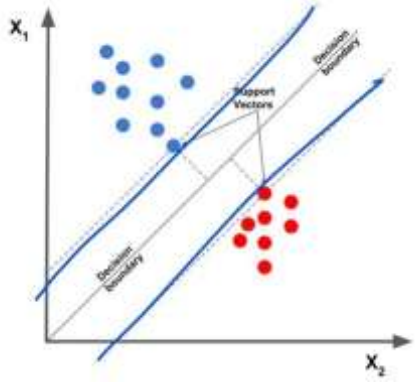

Fig. 5. SVM hyperplane separation of data points [16]

D. Evaluation Metrics

Evaluation metrics are used to test the quality and performance of the machine learning model. In this paper, the best model was chosen based on the following evaluation metrics:

Confusion Matrix: An N*N matrix, where N is the number of classes being predicted. For a prediction problem of two possible outputs, like in this work, the confusion matrix dimension is 2*2.

		Predicted class: P	Predicted class: N
Actual class	P	True Positive (TP)	False Negative (FN)
	N	False Positive (FP)	True Negative (TN)

The elements of the matrix are the counts of the correct and incorrect predictions, separated by class. For example, a True Positive is the number of the correctly classified Positive class (in this case, the number

of correctly diagnosed heart diseases). Similarly, a True Negative is the number of correctly classified Negative class (In this case, the count of correctly predicted absence of heart disease).

Accuracy: The percentage of the total number of predictions that were classified correctly, and is obtained from the confusion matrix by the following equation:

A = (TP + TN) / (TP + TN + FP + FN)

Precision: The percentage of the positive cases that were classified correctly, and is obtained from the confusion matrix by the following equation:

PR = TP / (TP + FP)

Sensitivity or Recall: The percentage of the actual positive cases that were classified correctly, and is obtained from the confusion matrix by the following equation:

RE = TP / (TP + FN)

F1 Score: If the target is to get the best precision and recall, F measure would be the best choice as it provides a harmonic mean of the recall and the precision values in classification problem, and is obtained from the confusion matrix by the following equation:

F1 = 2TP / (2*TP + FP + FN)

III. RESULTS AND DISCUSSION

In our study, we conducted experiments to combine Naive Bayes (NB), Support Vector Classifier (SVC), and Random Forest (RF) into a hybrid model for predicting the recurrence of cardiovascular diseases (CVD). The goal was to leverage the strengths of these algorithms and enhance predictive accuracy.

We conducted experiments to combine Naive Bayes (NB) and Random Forest (RF) into a hybrid model for predicting the recurrence of cardiovascular diseases (CVD). Our results showed that the hybrid model's accuracy of 0.9427

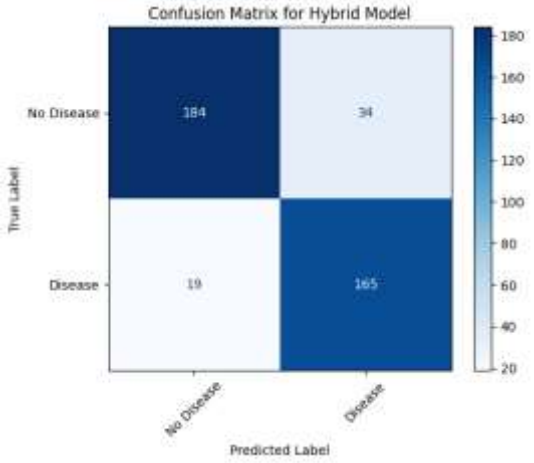

Fig. 6. Confusion matrix for NB and RF Hybrid model

Further, we combined Naive Bayes (NB) and Support Vector Machine (SVM) into a hybrid model for predicting the recurrence of cardiovascular diseases (CVD). The goal was to leverage the strengths of these algorithms and enhance predictive accuracy. However, our results showed that the hybrid model's accuracy of 0.8159 was too lower than the accuracy achieved by the RF and NB Hybrid model. This outcome suggests that the combination of NB and SVM did not significantly improve predictive performance for our dataset and problem domain.

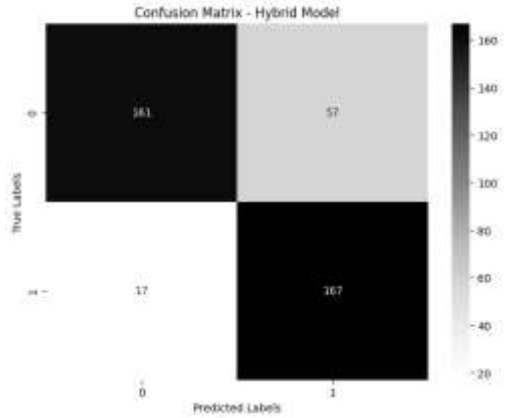

Fig. 7: Confusion Matrix for NB and SVM Hybrid Model.

Finally, we combined Random Forest (RF) and Support Vector Machine (SVM) into a hybrid model for predicting the recurrence of cardiovascular diseases (CVD). The aim was to leverage the strengths of these algorithms and enhance predictive accuracy. However, our results indicated that the hybrid model's accuracy of 0.9215.

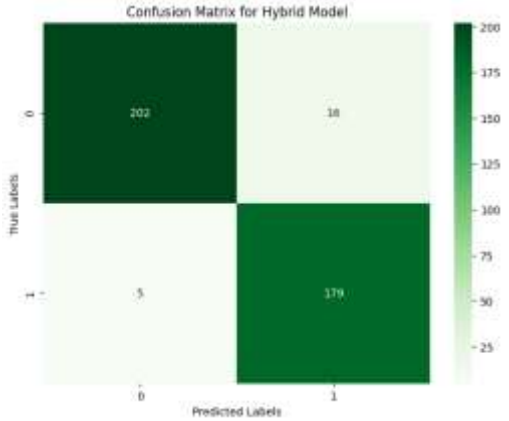

Fig. 8: Confusion Matrix for RF and SVM Hybrid Model.

V. CONCLUSION AND FUTURE WORK

This study establishes and validates a risk model to predict the recurrence of cardiovascular disease in patients according to specific health measurements.

The paper demonstrated 3 hybrid classification mechanism to build the prediction model. The data was collected and cleaned from any missing values and outliers. The model was trained and tested for each hybrid machine learning algorithms. RF and NB algorithms with had the best results with a 94.27% accuracy, 91.28% precision, 96.73% recall, and F1 Score of 93.93%. Each algorithm has its own strengths and weaknesses, and by combining them, the hybrid model may achieve better results than any single model alone.

Hybrid model in machine learning combines the strengths of multiple individual models to improve predictive performance by leveraging their unique capabilities. Combining Naive Bayes (NB), Support Vector Machine (SVM), and Random Forest (RF) models in a hybrid approach may enhance the overall accuracy and robustness of the predictive model.

Machine learning algorithms can also be applied to other types of diseases, especially with the generation of more accurate datasets in the medical field in the future.

REFERENCES

1. S. Rehman, E. Rehman, M. Ikram, and Z. Jianglin "Cardiovascular disease (CVD): assessment, prediction and policy implications," *BMC Public Health*, vol. 21, no. 1, p. 1299, 2021, doi: 10.1186/s12889-021-11334-2.
2. O. Atef, A. B. Nassif, M. A. Talib, and Q. Nassir, "Death/Recovery Prediction for Covid-19 Patients using Machine Learning," 2020.
3. A. B. Nassif, I. Shahin, M. Bader, A. Hassan, and N. Werghi, "COVID-19 Detection Systems Using Deep-Learning Algorithms Based on Speech and Image Data," *Mathematics*, 2022.
4. H. Hijazi, M. Abu Talib, A. Hasasneh, A. Bou Nassif, N. Ahmed, and Q. Nasir, "Wearable Devices, Smartphones, and Interpretable Artificial Intelligence in Combating COVID-19," *Sensors*, vol. 21, no. 24, 2021, doi: 10.3390/s21248424.
5. O. T. Ali, A. B. Nassif, and L. F. Capretz, "Business intelligence solutions in healthcare a case study: Transforming OLTP system to BI solution," in *2013 3rd International Conference on Communications and Information Technology, ICCIT 2013*, 2013, pp. 209–214, doi: 10.1109/ICCITechnology.2013.6579551.
6. A. Nassif, O. Mahdi, Q. Nasir, M. Abu Talib, and M. Azzeh, "Machine Learning Classifications of Coronary Artery Disease." Jan. 2018.
7. A. F. Otoom, E. E. Abdallah, Y. Kilani, A. Kefaye, and M. Ashour, "Effective diagnosis and monitoring of heart disease," *Int. J. Softw. Eng. its Appl.*, vol. 9, no. 1, pp. 143–156, 2015, doi: 10.14257/IJSEIA.2015.9.1.12.
8. K. Vembandasamyp, R. R. Sasipriyap, and E. Deepap, "Heart Diseases Detection Using Naive Bayes Algorithm," *IJISETInternational J. Innov. Sci. Eng. Technol.*, vol. 2, no. 9, 2015, Accessed: Dec. 11, 2021. [Online]. Available: www.ijiset.com.
9. A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, R. Sun, and I. GarciáMagarinõ, "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms," *Mob. Inf. Syst.*, vol. 2018, 2018, doi: 10.1155/2018/3860146.
10. D. Shah, S. Patel, · Santosh, and K. Bharti, "Heart Disease Prediction using Machine Learning Techniques," vol. 1, p. 345, 2020, doi: 10.1007/s42979-020-00365-y.
11. K. Pahwa and R. Kumar, "Prediction of heart disease using hybrid technique for selecting features," *2017 4th IEEE Uttar Pradesh Sect. Int. Conf. Electr. Comput. Electron. UPCON 2017*, vol. 2018-January, pp. 500–504, Jun. 2017, doi: 10.1109/UPCON.2017.8251100.
12. H. Jindal, S. Agrawal, R. Khera, R. Jain, and P. Nagrath, "Heart disease prediction using machine learning algorithms," doi: 10.1088/1757-899X/1022/1/012072.
13. "Heart Disease UCI | Kaggle." https://www.kaggle.com/ronitf/heart-disease-uci
14. D. Murphy, "Using Random Forest Machine Learning Methods to Identify Spatiotemporal Patterns of Cheatgrass Invasion through Landsat Land Cover Classification in the Great Basin from 1984 - 2011," 2019.
15. S. Liu, Z. Fang, and L. Zhang, "Research on Urban Short-term Traffic Flow Forecasting Model," *J. Phys. Conf. Ser.*, vol. 1237, no. 5, Jul. 2019, doi: 10.1088/1742-6596/1237/5/052026.
16. "Support Vector Machines (SVM) | LearnOpenCV #." https://learnopencv.com/support-vector-machines-svm/ (accessed Jan. 10, 2022).