Exploring the Path: Machine Learning Approaches to Cardiovascular Risk Assessment

NagaCharitavya Madala
Department of CSE,
SRM University-AP,
nagacharitavya_m@srmap.edu.in

Sai Durga Sardhi Pranu Deepak Tallapudi
Department of CSE,
SRM University-AP,
deepak_tallapudi@srmap.edu.in

Mahitha Chimata
Department of CSE,
SRM University-AP,
mahitha_chimata@srmap.edu.in

Venkata Srikari Malladi
Department of CSE,
SRM University-AP,
venkatasrikari_m@srmap.edu.in

Srilatha Tokala
ACT Lab Department of CSE,
SRM University-AP,
srilatha_tokala@srmap.edu.in

Murali Krishna Enduri
ACT Lab, Department of CSE,
SRM University-AP,
muralikrishna.e@srmap.edu.in

***Abstract*—Cardiovascular disease, refers to a variety of circumstances affecting the cardiovascular and blood vessels, such as cardiovascular failure and coronary artery illness. Cardiovascular disease, a major global health concern, is frequently caused by atherosclerosis, a condition in which plaque builds up and obstructs blood flow. This study introduces a predictive modeling methodology utilizing various machine learning (ML) algorithms. Additionally, hybrid models including Random Forest-Gradient Boosting, Genetic Algorithm-Support Vector Machine (GA-SVM), AdaBoost-Support Vector Machine (AdaBoost-SVM), Logistic Regression-Principal Component Analysis (LR-PCA), and Gradient Boosting Machines-Decision Tree (GBM-DT) have been integrated into the analysis. Using two distinct datasets, our study focuses on proactive heart disease management, addressing a significant health challenge. Notably, the Random Forest-Gradient Boosting Machines (RF-GBM) hybrid model exhibited exceptional performance, achieving an impressive 93.5% accuracy for both datasets in predicting heart disease. These results highlight the effectiveness of our integrated approach in advancing predictive modeling for improved cardiovascular health management.**

Keywords-cardiovascular disease, predictive modeling, machine learning algorithms, atherosclerosis.

I. INTRODUCTION

A variety of circumstances impacting the cardiovascular and blood vessels collectively known as cardiovascular illness pose a significant global health risk. The increasing prevalence of coronary artery illness and cardiovascular failure contributes to a concerning rise in global cases of cardiovascular disease (CVD). The urgency to address heart disease is underscored by compelling statistics: CVD-related deaths improved from 12.1 million in 1990 to 20.5 million in 2021, firmly establishing it as the leading cause of mortality worldwide [1]. The 2023 report from the World Cardiovascular Federation, presented at the World Cardiovascular Summit, brings attention to escalating risk and mortality rates associated with cardiovascular disease (CVD) on a global scale. Despite a worldwide decline in CVD-related deaths, Central Europe, Eastern Europe, and Central Asia continue to experience the highest rates. This concerning pattern, influenced by factors such as an aging population, necessitates a thorough and targeted approach to mitigate the impact of cardiovascular disease, particularly in low- and middle-income countries. In the United States, heart disease is still the leading cause of death. Affecting individuals of diverse genders and ethnic backgrounds. Alarmingly, approximately 695,000 individuals succumbed to CVD in 2021, accounting for one-fifth of all deaths [2]. This frightening impact translates into a major economic burden, costing the United States an estimated $239.9 billion per year from 2018 to 2019, including healthcare bills, medications, and lost productivity due to premature death. Considering these challenges, it is evident that accurate forecasting and efficient handling of cardiovascular conditions are paramount. Professor Fausto Pinto, a contributor to the WHF study, asserts that adopting proactive strategies holds the potential to avert as much as 80% of untimely heart attacks and strokes [3]. This research aims to elevate predictive modeling for heart disease by using a varied range of ML techniques. While exploring traditional models such as LR, NB, KNN, DT, RF, GBM, XGBoost, AdaBoost, SVM, PCA, GA, the research mainly focuses on hybrid models. The hybrid ML technique involves combining the features of many models to build strong combinations. Our analysis focused on models such as GBM-DT, LR-PCA, GA-SVM, AdaBoost-SVM, and RF-GBM, to improve predictive modeling for heart

disease [4]. This approach explores the connection among many models, pushing the limits of cardiovascular health prediction. The research focuses on discovering the most successful hybrid models, rather than independent ML algorithms, to provide an effective and fresh perspective regarding cardiovascular health care.

II. RELATED WORK

Senthilkumar Mohan et al. applied a range of machine learning methodologies in their study, introducing neural networks based on heart rate time series to enhance the precision of cardiac disease [5]. Training and testing were conducted using a Radial Basis Function Network (RBFN), with 30% reserved for testing purposes. The mixed algorithm Random Forest-Linear Model (RFLM) demonstrated the highest accuracy at 88.7%, underscoring the efficacy of their methodology. The inclusion of various metrics ensured a comprehensive evaluation of the model's effectiveness. This study adds valuable insights to the expanding knowledge base in the field by delivering insightful predictions related to cardiac disease. Bhanu Prakash et al. applied diverse machine learning techniques, including Naive Bayes, Decision Tree, Logistic Regression, SVM, Random Forest, and KNN in their investigation [6]. The study introduced a Hybrid model, merging Radial Basis Function (GA-RBF) and Genetic Algorithm. Among the 76 features in the dataset, 297 were utilized in the analysis, focusing on 14 key indicators. The performance assessment encompassed crucial metrics. Notably, the GA-RBF Hybrid model exhibited exceptional performance, achieving a maximum accuracy of 94.20%. This result underscores the effectiveness of methodology in accurately predicting heart disease, making a substantial contribution to the advancement of cardiovascular health prediction modeling. Ahmed et al. used a wide range of machine learning methods in their investigation, including Naive Bayes, KNN, Decision Trees, Random Forest, and Logistic Regression [7]. Notably, the study introduced an innovative Hybrid model that merged a Convolutional Neural Network with Long Short-Term Memory (CNN with LSTM) to enhance the prediction of heart disease Evaluation metrics such as F1 Score, Accuracy, Recall, and Precision were employed during the assessment phase. Notably, the suggested Stacking SVM demonstrated effectiveness with an astounding accuracy of 98.41%. of the model in improving heart disease prediction accuracy. This novel approach offers a strong methodology for precise and trustworthy cardiovascular health diagnostics, which makes a substantial contribution to the area. Ramalingam et al. investigation delved deeply into the exploration of machine learning methodologies for predicting heart disease, encompassing SVM, KNN, Naive Bayes, Decision Tree, and Random Forest [8]. The study emphasized the significance of dimensionality reduction by employing feature extraction and selection strategies. Notably, chi-square was used for feature selection. Following extensive testing of various machine learning techniques, Support Vector Machine (SVM) emerged as the most accurate, achieving an impressive accuracy rate of 92.1%. The authors' focus on dimensionality reduction reflects their commitment to enhancing the predictive model, thereby improving the overall effectiveness and accuracy of heart disease prediction. This research significantly contributes to the field by highlighting the pivotal role of SVM in optimal feature selection for enhancing forecast precision. Abhijeet et al. carried out a study utilizing diverse machine learning algorithms, including SVM, Logistic Regression, and Naive Bayes, utilizing a dataset from the University of California, Irvine [9]. The authors meticulously partitioned the dataset, allocating 25% for algorithmic accuracy evaluation and reserving 75% for training. With a specific emphasis on support vector machines (SVM), the study covered fundamental operations like feature scaling, factorization, data preprocessing, and cleaning. The cumulative efforts resulted in a noteworthy accuracy of 64.4%, marking the highest accuracy ever achieved by SVM. Despite its poor accuracy, the study offers light on the complexities of data preparation and algorithmic performance evaluation, offering useful insights into the use of machine learning approaches for cardiovascular disease prediction.

. III. MATERIALS AND METHODOLOGY

An overview of the steps taken in the suggested attempt is given in the sections that follow.

Proposed System:

The model's implementation sequence is shown in Fig. 1.

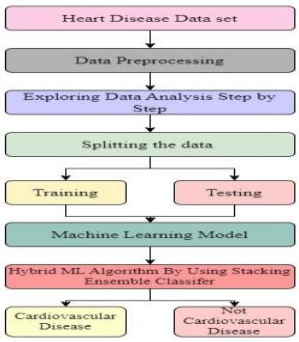

Fig.1.Algorithm for proposed Machine Learning Model

IV. DATASETS

For this study, we looked at two different sets of numbers. The initial dataset, called "heart.csv," is specifically designed to forecast whether heart disease will manifest or not. The dataset, which consists of 303 instances with 14 attributes—13 of which are numerical—includes physiological, clinical, and demographic data. Interestingly, the class attribute "target" indicates whether heart disease is present (1) or not (0). There are no missing values or null values present in the dataset. To obtain a better comprehension of the dataset we generated a distribution plot for each variable.

The goal of the second dataset, "framingham.csv," is to project the 10-year risk of cardiac disease development. This dataset, which contains 4240 instances and 16 attributes—15 of which are numerical—introduces the binary classification "TenYearCHD," which denotes the risk for the following ten years. Six binary-encoded attributes that reflect gender, smoking habits, medication usage, and health conditions are part of the essential feature set. Although the dataset 'heart.csv' shows no missing values, the dataset 'framingham.csv' initially had 645 null values or roughly 13% of the total data. Once these are subtracted, 3658 instances remain in the dataset. There are no missing values or null values present in the dataset. To obtain a better comprehension of the dataset we generated a distribution plot for each variable.

V. MODEL METHODOLOGY

The machine learning framework incorporates a range of models, discussed in this section. Moreover, the framework features hybrid algorithms that amalgamate the strengths of individual models. The GBM-DT hybrid combines Decision Trees' interpretability with gradient-boosting predictive ability. To improve classification, PCA-LR uses Principal Component Analysis to reduce dimensionality, followed by Logistic Regression. GA-SVM uses Genetic Algorithms to optimize Support Vector Machine parameters, which improves performance. AdaBoost-SVM uses AdaBoost ensemble learning to enhance SVM's predictive capabilities [10]. Finally, RF-GBM combines the advantages of Random Forest and Gradient Boosting into an ensemble model. Preprocessing steps specific to each model performance evaluation using relevant metrics are all emphasized in the implementation guidelines.

A. Decision Tree: It works by recursively splitting data into smaller groups by using input feature values, facilitating decision-making or predictions [11].

B. Random Forest: An ensemble learning algorithm called Random Forest produces a large number of Trees to determine the classification's mode or mean prediction [12].

C. Principal Component Analysis: Using Principal Component Analysis (PCA), dimensionality can be decreased by converting high-dimensional data into a lower-dimensional format while maintaining as much of the original variability as possible [13].

D. Support vector Machine: By working in a high-dimensional space, it finds the best hyperplane to split the data into different categories [14].

E. Genetic Algorithm: They apply evolutionary concepts to discover approximate solutions for optimization and search problems [15].

F. Logistic Regression: The statistical technique known as logistic regression is used to solve double and multiclass classifying issues [15].

G. Naive Bayes: The "naive" assumption that characteristics are independent is the foundation of the Bayes theorem-based probabilistic classification technique known as Naive Bayes [16].

H. K-Nearest Neighbor: K-Nearest Neighbours is a straightforward and adaptable method that can be used for both regression and classification (KNN) [16].

I. Gradient Boosting Machines: Gradient Boosting Machine (GBM) is an ensemble learning approach that generates a predictive model from a collection of weak learners, usually decision trees.

J. XGBoost: A scalable and optimized variant of gradient boosting is called XGBoost [17].

K. ADABoost: Using an ensemble learning algorithm called AdaBoost, a strong learner is produced by combining the predictions of weak learners. It gives varying weights to examples according to how accurately they were classified, giving more attention to examples that earlier learners misclassified [18].

L. Gradient Boosting Machine - Decision Tree (GBMDT): GBM-DT is a hybrid model that combines the potent predictive powers of Gradient Boosting with the interpretability of Decision Trees.

M. Principal Component Analysis - Logistic Regression (PCA LR): PCA-LR is a hybrid model that uses Principal Component Analysis (PCA) to minimize dimensionality and then Logistic Regression to improve classification [19].

N. Genetic Algorithm - Support Vector Machine (GASVM): GA-SVM is a hybrid model that uses Genetic Algorithms to optimize Support Vector Machine (SVM) parameters to improve SVM's performance [20].

O. AdaBoost - Support Vector Machine (AdaBOST SVM): AdaBoost-SVM enhances the Support Vector Machine's (SVM) predictive power by employing the ensemble learning technique AdaBoost [21].

P. Random Forest - Gradient Boosting Machines (RF-GBM): The Random Forest - Gradient Boosting Machine ensemble model, or RF-GBM for short, combines the best aspects of Gradient Boosting and Random Forest [19].

VI. RESULTS

The investigation thoroughly examined many categorization schemes to decide the well-organized way to diagnose cardiovascular disease. Strict adherence was maintained to validation procedures.

Accuracy: By dividing the number of cases that the model correctly predicts by the total number of instances, accuracy calculates the model's overall performance as shown in Eq.1.

$$\text{Accuracy} = \frac{TP+TN}{P+N} \quad (1)$$

Precision: It specifies how well a model predicts favorable outcomes. Out of all the anticipated positives, it displays the number of genuine positives as shown in Eq.2.

$$\text{Precision} = \frac{TP}{Tp+FP} \quad (2)$$

Recall: The complete number of true positives divided by the complete number of actual positives yields the recollect value. This shows how well the model finds every pertinent occurrence as shown in Eq.3.

$$\text{Recall} = \frac{TP}{P} \quad (3)$$

F1 Score: A balanced statistic that accounts for both false positives and false negatives is the F1 score. The harmonic mean of accuracy and recall is used to calculate it as shown in Eq.4.

$$\text{F1 score} = \frac{2*Precision*Recall}{Precision+Recall} \quad (4)$$

Fig.3 shows the performance measures of different models performed for algorithms in 'heart.csv' dataset. It is observed in the figure that for all the performance measures ADA Boost algorithm performs well compared to the other algorithms. The least performance is given by the KNN algorithm. Fig.4 demonstrates the performance of hybrid algorithms used in our analysis. It is observed in the analysis that ADA Boost combined with SVM method gives better performance that all the remaining algorithms. The least performance is given by the genetic algorithm and SVM methods. The results are tabulated for heart dataset in Table 1 and Table 2.

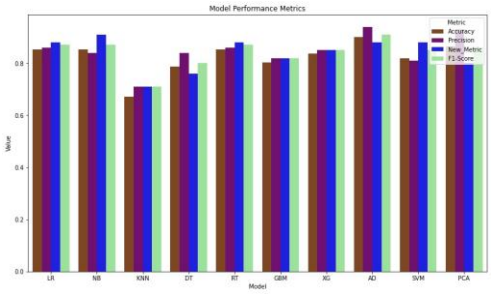

Fig-3: Performance measures of different models performed for algorithms in 'heart.csv' dataset

Algorithm	Accuracy	Precision	Recall	F1-Score
LR	0.85	0.86	0.88	0.87
NB	0.85	0.84	0.91	0.87
KNN	0.67	0.71	0.71	0.71
DT	0.78	0.84	0.76	0.80
RF	0.85	0.86	0.88	0.87
GBM	0.80	0.82	0.82	0.82
XGB	0.83	0.85	0.85	0.85
ADB	0.90	0.94	0.88	0.91
SVM	0.81	0.81	0.88	0.85
PCA	0.86	0.93	0.80	0.86

Table-1: 'heart.csv'results for ML algorithms

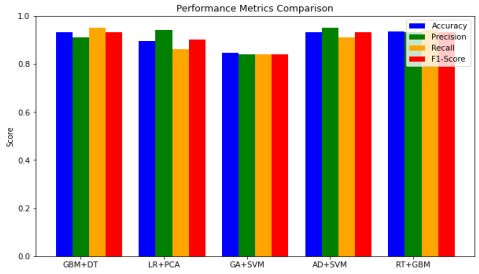

Fig-4: Performance measures of different models performed for hybrid algorithms in 'heart.csv' dataset

Algorithm	Accuracy	Precision	Recall	F1-Score
GBM+DT	0.93	0.91	0.95	0.93
LR+PCA	0.89	0.94	0.86	0.90
GA+SVM	0.84	0.84	0.84	0.84
AD+SVM	0.93	0.95	0.91	0.93
RT+GBM	0.93	0.93	0.94	0.93

Table-2: heart.csv'results for hybrid algorithms

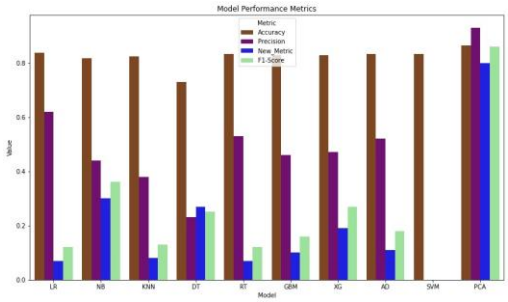

Fig-5: Performance measures of different models performed for algorithms in 'framingham.csv' dataset

Algorithm	Accuracy	Precision	Recall	F1-Score
LR	0.83	0.62	0.07	0.12
NB	0.81	0.44	0.30	0.36
KNN	0.82	0.38	0.08	0.13
DT	0.72	0.23	0.27	0.25
RF	0.83	0.53	0.07	0.12
GBM	0.82	0.46	0.10	0.16
XGB	0.82	0.47	0.19	0.27
ADB	0.83	0.52	0.11	0.18
SVM	0.83	0.00	0.00	0.00
PCA	0.86	0.93	0.80	0.86

Table-3: 'framingham.csv'results for ML algorithms

Fig. 5 shows the performance measures of different models performed for algorithms in 'framingham.csv' dataset. It is observed in the dataset that PCA method gives better performance than all the remaining methods. The least performance for the accuracy and precision measures is given by the decision tree method, and the random forest method gives the less score for the recall and F1-score method. Fig.6 demonstrates the performance measures of different models performed for hybrid algorithms in the 'framingham.csv' dataset. It is observed in the dataset that the hybrid method for decision tree and gradient boosting method gives better performance than all the methods. The least performance is shown by the hybrid method genetic algorithm and SVM. The results for framingham dataset are tabulated in Table 3 and Table 4.

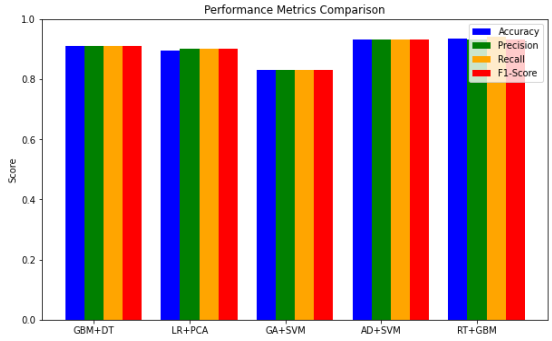

Fig-6: Performance measures of different models performed for hybrid algorithms in 'framingham.csv' dataset

Algorithm	Accuracy	Precision	Recall	F1-Score
GBM+DT	0.91	0.91	0.91	0.91
LR+PCA	0.89	0.90	0.90	0.90
GA+SVM	0.83	0.83	0.83	0.83
AD+SVM	0.93	0.93	0.93	0.93
RT+GBM	0.93	0.93	0.94	0.93

Table-4: 'framingham.csv'results for hybrid algorithms

VII. CONCLUSION

Our study found that Random Forest (RF) and Gradient Boosting Machine (GBM) performed better than other models and hybrid models on both datasets, with an accuracy rate of 93.5%. In navigating the complexity of our data, their iterative and ensemble optimization approaches proved to be very successful. With regard to our particular datasets, the RF and GBM's strong performance confirms their reputation as potent tools. A major factor in their success in producing better predictive results is their ability to recognize complex data relationships and adjust to a variety of patterns. Thus, for comparable predictive modeling tasks falling under our research purview, our results emphasize RF

and GBM as the best options for maximum accuracy. Furthermore, the predictability of our results was strengthened by our methodology, which takes into account every variable that was available for analysis. Our comprehensive approach also improves the forecasts' accuracy and reliability.

VIII. REFERENCES

[1]. A. S. Abdullah and R. R. Rajalaxmi, ‘‘A data mining model for predicting the coronary heart disease using random forest classifier,’’ in Proc. Int. Conf. Recent Trends Compute Methods, Commun. Controls, Apr. 2012, pp. 22–25.

[2]. N. Al-milli, ‘‘Back Propagation neural network for prediction of heart disease,’’ J. Theor. Appl.Inf. Technol., vol. 56, no. 1, pp. 131–135, 2013.

[3]. C. A. Devi, S. P. Rajamohanan, K. Umamaheswari, R. Kiruba, K. Karunya, and R. Deepika, ‘‘Analysis of neural networks based heart disease prediction system,’’ in Proc. 11th Int. Conf. Hum. Syst. Interact. (HSI), Gdansk, Poland, Jul. 2018, pp. 233–239.

[4]. Mourão-Miranda, J., Bokde, A.L.W., Born, C., Hampel, H., Stetter, M.: Classifying brain states and determining the discriminating activation patterns: support vector machine on functional MRI data. Neuroimage 28(4), 980–99

[5]. S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[6]. Doppala, Bhanu Prakash, et al. "A hybrid machine learning approach to identify coronary diseases using a feature selection mechanism on a heart disease dataset." Distributed and Parallel Databases (2021): 1-20.

[7]. Almulihi, Ahmed, et al. "Ensemble Learning Based on Hybrid Deep Learning Model for Heart Disease Early Prediction." Diagnostics 12.12 (2022): 3215.

[8]. Ramalingam, V. V., Ayantan Dandapath, and M. Karthik Raja. "Heart disease prediction using machine learning techniques: a survey." International Journal of Engineering & Technology 7.2.8 (2018): 684-687.

[9]. Jagtap, Abhijeet, et al. "Heart disease prediction using machine learning." International Journal of Research in Engineering, Science and Management 2.2 (2019): 352-355.

[10]. A. S. Abdullah and R. R. Rajalaxmi, ‘‘A data mining model for predicting the coronary heart disease using random forest classifier,’’ in Proc. Int. Conf. Recent Trends Compute Methods, Commun. Controls, Apr. 2012, pp. 22–25.

[11]. Vanisree, K., Singaraju, J.: Decision support system for congenital heart disease diagnosis based on signs and symptoms using neural networks. Int. J. Comput. Appl. 19(6), 6–12 (2011)

[12]. Nazir, S., Shahzad, S., Septem Riza, L.: Birthmark-based software classification using rough sets. Arab. J. Sci. Eng. 42(2), 859–871 (2017)

[13]. Hall, J.E.; Hall, M.E. Guyton and Hall Textbook of Medical Physiology e-Book; Elsevier Health Sciences: Amsterdam, The Netherlands, 2020

[14]. Bhowmick, A.; Mahato, K.D.; Azad, C.; Kumar, U. Heart Disease Prediction Using Different Machine Learning Algorithms. In Proceedings of the 2022 IEEE World Conference on Applied Intelligence and Computing (AIC), Sonbhadra, India, 17–19 June 2022; pp. 60–65.

[15]. Aluffi, B.; Alharbi, A.; Sahal, R.; Saleh, H. An Optimized Hybrid Deep Learning Model to Detect COVID-19 Misleading Information. Compute. Intell. Neurosci. 2021, 2021, 9615034.

[16]. Weissler, E.H.; Naumann, T.; Andersson, T.; Ranganath, R.; Elemento, O.; Luo, Y.; Freitag, D.F.; Benoit, J.; Hughes, M.C.; Khan, F.; et al. The role of machine learning in clinical research: Transforming the future of evidence generation. Trials 2021, 22, 1–15.

[17]. Ramadoss and Shah B et al.“A. Responding to the threat of chronic diseases in India”. Lancet. 2005; 366:1744–1749. doi: 10.1016/S0140-6736(05)67343-6.

[18]. Global Atlas on Cardiovascular Disease Prevention and Control. Geneva, Switzerland: World Health Organization, 2011

[19]. R.Kavitha and E.Kannan et al. “An Efficient Framework for Heart Disease Classification using Feature Extraction and Feature Selection Technique in Data Mining “, 2016

[20]. Sellappan Palaniappan, Rafiah Awang “Intelligent Heart Disease Prediction System Using Data Mining Techniques”, IEEE, July 2015

[21]. Marjia Sultana, Afrin Haider and Mohammad Shorif Uddin “Analysis of Data Mining Techniques for Heart Disease Prediction”, May 2015.