Heart Disease Prediction Using Hybrid RandomForest and SVM

K Shyam Sunder Reddy
Department of Information Technology
Vasavi College of Engineering Hyderabad, India shyamd4@staff.vce.ac.in
R Sushma
Department of Information Technology
Vasavi College of Engineering Hyderabad, India sushmarao2001@gmail.com

Abstract—One of the leading causes of death in the modern world is heart disease. Clinical data analysis has a significant problem when attempting to predict cardiovascular disease. In this paper, a hybrid machine learning model for heart disease prediction that combines random forest (RF) and support vector machine (SVM) techniques is proposed. Hybrid model is designed to improve the accuracy of existing methods by identifying significant features and applying different techniques. The proposed hybrid model which is the combination of Random Forest and Support Vector Machine (H-RFSVM) utilizes the strengths of both SVM and RF algorithms to increase the accuracy of heart disease prediction. In comparison to other algorithms, SVM Classifiers perform faster prediction and offer good accuracy. SVM performs well in high-dimensional spaces and with well-defined margins of separation. Random Forest is a machine learning algorithm used for classification tasks due to its high accuracy, feature importance and scalability. Random Forest is less sensitive to noise and outliers in the data and reduces overfitting by averaging multiple decision trees. Overall, the proposed hybrid machine learning model, H-RFSVM, can be used for accurately predicting heart disease, identifying important risk factors.*Index Terms*—Machine learning, Hybrid machine learning, Prediction, Support Vector Machine (SVM), Random Forest, Classification.

I. INTRODUCTION

Heart disease is a significant public health concern and a leading cause of mortality worldwide. It refers to a group of conditions have an impact on the blood arteries and heart, like coronary artery disease and heart failure. Detecting and predicting heart disease at an early stage are crucial for effective treatment and prevention. Machine learning techniques have been used to develop predictive models that can analyze many patient-related factors and accurately predict the likelihood of developing heart disease. These models typically rely on algorithms such as logistic regression, decision trees, support vector machines, random forests, and neural networks. In recent years, machine learning techniques have shown great potential in predicting heart disease by analyzing various patient-related factors such as age, gender, blood pressure, cholesterol levels and other conditions. The use of machine learning techniques in heart disease prediction has become increasingly popular due to their ability to handle large and complex datasets, identify important risk factors, and accurately predict outcomes.

In addition to algorithmic approaches researchers are also exploring new data sources and techniques to improve heart disease prediction. For example, some studieshave incorporated genetic data, imaging data, and clinical biomarkers into predictive models.

II. PROBLEM STATEMENT – OVERVIEW

The research problem identified for heart disease prediction using machine learning models is the need to improve the accuracy and interpretability of the model. Although thesemethods have the potential to increase prediction accuracy, there are still a number of issues that need to be resolved. One of the key challenges is feature selection. Machine learning models require relevant features to make accurate predictions, but including irrelevant features can lead to overfitting. Another research problem is the selection of the optimal hyperparameter for both the random forest and logistic regression. Hyperparameter determine the specific settings of the algorithms and can have a significant impact on the model's performance. The current state of the art towards the problem of heart disease prediction using hybrid machine learning models that combine random forest (RF) with support vector machine (SVM) algorithms. These models have shown promising results in terms of accuracy and have been compared to other hybrid models such as random forest with

logistic regression. The SVM algorithm is known for its ability to handle complex data and handle non-linear relationships, which makes it a good complement to the logistic regression. The combination of two algorithms has been shown to improve feature selection, reduce overfitting, and improve the interpretability of the model.

III. SCOPE AND OBJECTIVE

The scope of this work is to explore the potential of combining the strengths of both algorithms to overcome the limitations of machine learning approaches. The primary objective of the proposed work is to enhance the accuracy of classification tasks by utilizing SVM as a base estimator for each tree in the Random Forest, which will enable the model to handle complex datasets with high-dimensional features. The proposed work will provide a significant contribution to the field of machine learning by introducing a new hybrid model that can produce more accurate and reliable results for complex classification tasks.

IV. LITERATURE SURVEY

"Heart Disease Prediction using Random Forest" was published in Journal of Physics: Conference Series in 2021 by the authors A. O. Adeniji, A. S. Akinola, A. O. Oluwatayo, and A. O. Oluwadare. The study focuses on using a random forest algorithm to predict the likelihood of heart disease in patients based on a dataset of clinical and demographic features. The authors use feature selection techniques to identify the most relevant features for predicting heart disease. They find that the random forest model outperforms other models in terms of accuracy and specificity.

"Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques" was included in the 2019 IEEE International Conference on Computational Intelligence in Data Science (ICCIDS) proceedings and published by SenthilkumarMohan, Chandrasegar Thirumalai, and Gautam Srivastava . The study focuses on using a hybrid machine learning approach for predicting heart disease by combining decision tree, random forest, and K-nearest neighbors algorithms.

"Heart disease prediction using machine learning algorithms" was published in the IOP Conference Series: Materials Science and Engineering was published by Harshit Jindal, Vikas Jindal, and Sonali Agarwal in 2021. The study aims to develop a heart disease prediction model using various machine learning algorithms such as decision tree, random forest and support vector machine, with an accuracy of 86.5, they discover that the random forest algorithm performs the best.

"Prediction and Analysis of Heart Disease using SVM Algorithm" was published in International Research Journal of Engineering and Technology was published by Madhura Patil, Rima Jadhav, Vishakha Patil, and Mrs. Geeta Chillarge in 2019. The study focuses on using a support vector machine (SVM) algorithm to predict heart disease based on a dataset of clinical and demographic features. The authors evaluate the performance of the SVM model against other machine learning algorithms and find that the SVM model achieves higher accuracy and specificity. They suggest that the SVM algorithm is a promising approach for heart disease prediction and can aid in the early diagnosis and treatment of the disease.

"Logistic Regression Models in Predicting Heart Disease" was published in the Journal of Physics: Conference Series by Yingjie Zhang, and Li Li in 2021. The study focuses on using logistic regression models to predict the occurrence of heart disease based on clinical and demographic features. The authors evaluate the performance of different logistic regression models, including stepwise regression and ridge regression.

V. PROPOSED WORK

The proposed work will provide a significant contribution to the field of machine learning by introducing a new hybrid model that can produce more accurate and reliable results for complex classification tasks. The hybrid method combines Random Forest (RF) with Support Vector Machine (SVM) to further improve the accuracy. This method essentially substitutes the Support Vector Machine algorithm for Logistic Regression because it is a better classifier. On different subsets of the data, several decision trees are

constructed, and each decision tree leaf is fitted with an SVM. The resulting SVM models are then combined to form an ensemble classifier. The outcomes of the several decision trees are merged (0 or 1), with 1 denoting the presence of heart disease.

A. BLOCK DIAGRAM

The RFSVM (Random Forest with Support Vector Machine) method, which replaces logistic regression with SVM along with Random Forest, aims to increase the accuracy of the proposed model. Unlike logistic regression, SVMs have the ability to prevent the model from being sensitive to outliers in the data, resulting in a model that can make accurate predictions for future analyses. Additionally, SVMs can handle nonlinearity in the data by employing nonlinear kernel functions rather than a simple linear kernel.

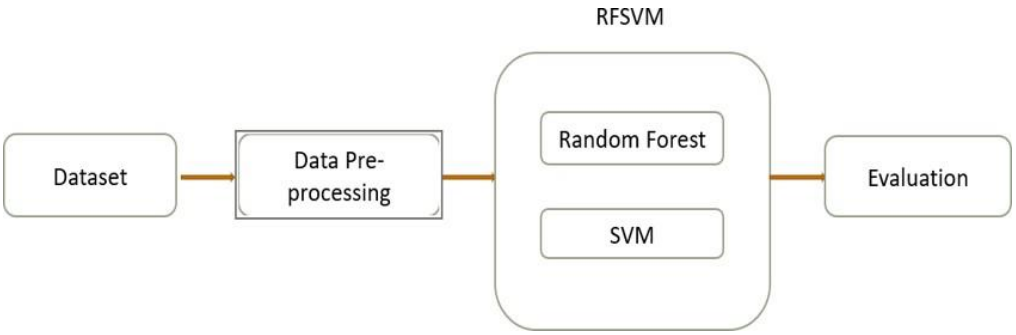

Fig. 1. Block Diagram

B. ALGORITHM

In this paper, a hybrid machine learning model that combines support vector machine (SVM) and random forest (RF) algorithms is proposed for heart disease prediction. The hybrid model is designed to improve the accuracy of existing methods by identifying significant features and applying different classification techniques. The proposed hybrid model which is the combination of Random Forest and Support Vector Machine (H-RFSVM) utilizes the strengths of both SVM and RF algorithms to increase the accuracy of heart disease prediction. In comparison to other algorithms, SVM classifiers perform faster prediction and offer good accuracy. SVM performs well in high-dimensional spaces and with well-defined margins of separation. Random Forest is a machine learning algorithm used for classification tasks due to its high accuracy, feature importance and scalability. Random Forest is less sensitive to noise and outliers in the data and reduces overfitting by averaging multiple decision trees. Overall, the proposed hybrid machine learning model, H-RFSVM, can be used for accurately predicting heart disease, identifying important risk factors.

Random Forest: Random forest is a type of ensemble learning method that builds multiple decision trees andcombines their results to improve the accuracy and reduce overfitting. In our hybrid model, random forest is used toextract important features from the dataset.

SVM: In this hybrid model of Random Forest and SVM,the SVM is being used as a base estimator for each tree inthe Random Forest. Each tree is constructed using a different subset of the training data, and a different subset of features is considered for each split. The SVM model in this hybrid model is used to find the threshold values of selected features to build each tree of the Random Forest.

The SVM model is initialized with the rbf kernel function and trained on selected features. The model is then used to predict class labels for the test set. The predictions made by all the trees are combined to form the final prediction.

VI. DATASETCleveland Dataset:

Information about heart disease was gathered from the UCI machine learning library. There are four databases: the VA Long Beach, Switzerland, Hungary, and Cleveland. The Cleveland database, which has extensive and full records, was chosen for this study since it is frequently used for ML projects. There are 303 records in the collection. Despite the Cleveland data set's 76 attributes, only 14 of them are collectedby the dataset that is available in the repository.

Attribute	Description	Type
Age	Patient's age in completed years	Numeric
Sex	Patient's Gender (male represented as1 and female as 0)	Nominal
Cp	The type of Chest pain categorized into 4 values: 1. typical angina, 2. atypical angina, 3. non-anginal pain and 4. asymptomatic	Nominal
Trestbps	Level of blood pressure at resting mode (in mm/Hg at the time of admitting in the hospital)	Numeric
Chol	Serum cholesterol in mg/dl	Numeric
FBS	Blood sugar levels on fasting > 120 mg/dl; represented as 1 in case of true, and 0 in case of false	Nominal
Resting	Results of electrocardiogram while at rest are represented in 3 distinct values: Normal state is represented as Value 0, Abnormality in ST-T wave as Value 1, (which may include inversions of T-wave and/or depression or elevation of ST of > 0.05 mV) and any probability or certainty of LV hypertrophy by Estes' criteria as Value 2	Nominal
Thali	The accomplishment of the maximum rate of heart	Numeric
Exang	Angina induced by exercise. (0 depicting 'no' and 1 depicting 'yes')	Nominal
Oldpeak	Exercise-induced ST depression in comparison with the state of rest	Numeric
Slope	ST segment measured in terms of the slope during peak exercise depicted in three values: 1. unsloping, 2. flat and 3. downsloping	Nominal
Ca	Fluoroscopy coloured major vessels numbered from 0 to 3	Numeric
Thal	Status of the heart illustrated through three distinctly numbered values. Normal numbered as 3, fixed defect as 6 and reversible defect as 7.	Nominal
Num	Heart disease diagnosis represented in 5 values, with 0 indicating total absence and 1 to 4 representing the presence in different degrees.	Nominal

TABLE I
UCI Dataset attributes detailed information

AGE	Numeric [29 to 77;unique=41;mean=54.4;median=56]
SEX	Numeric [0 to 1;unique=2;mean=0.68;median=1]
CP	Numeric [1 to 4;unique=4;mean=3.16;median=3]
TESTBPS	Numeric [94 to 200;unique=50;mean=131.69;median=130]
CHOL	Numeric [126 to 564;unique=152;mean=246.69;median=241]
FBS	Numeric [0 to 1;unique=2;mean=0.15;median=0]
RESTECG	Numeric [0 to 2;unique=3;mean=0.99;median=1]
THALACH	Numeric [71 to 202;unique=91;mean=149.61;median=153]
EXANG	Numeric [0 to 1;unique=2;mean=0.33;median=0.00]
OLPEAK	Numeric [0 to 6.20;unique=40;mean=1.04;median=0.80]
SLOPE	Numeric [1 to 3;unique=3;mean=1.60;median=2]
CA	Categorical [5 levels]
THAL	Categorical [4 levels]
TARGET	Numeric [0.00 to 4.00;unique=5;mean=0.94;median=0.00]

TABLE II
UCI Dataset range and datatype

VII. RESULTS

A. *EVALUATION METRICS*

	precision	recall	f1-score	support
0	0.91	0.89	0.90	47
1	0.83	0.86	0.85	29
accuracy			0.88	76

Fig. 2. Performance metrics of Hybrid Random Forest and Support Vector Machine

B. *ANALYSIS*

The ROC curve is a graphical representation of the performance of a binary classifier system. In this case, the blue line represents the ROC curve for our logistic regression model (HRFLM), and the dashed red line represents the ROC curve for a random classifier (a model that predicts 0 or 1 with equal probability). The False Positive Rate (FPR) and True Positive Rate (TPR) are represented by the x- and y-axes, respectively, of the ROC curve. The area under the ROC curve (AUC) is a measure of the classifier's ability to distinguish between the positive and negative classes. A perfect classifier has an AUC of 1, whereas a random classifier has an AUC of 0.5. So, in our plot, the blue line represents the ROC curve for our logistic regression model (HRFLM), which is better than the random classifier represented by the dashed red line. The AUC value for our HRSVM model is 0.88, which indicates that it is a good classifier.

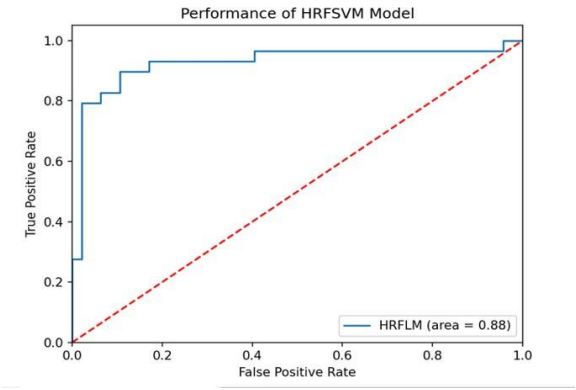

Fig. 3. Performance of HRFSVM Model

VIII. CONCLUSION

The proposed model of heart disease prediction using random forest and SVM is a hybrid approach that combines the strengths of both algorithms. Random Forest is known for its ability to handle high-dimensional datasets, handle missing data, and avoid overfitting. SVM, on the other hand, is known for its high accuracy, robustness, and effectiveness in handling both linear and nonlinear data. Moreover, in our proposed model, SVM is used as a base estimator for each tree in the Random Forest, and each tree is constructed using a different subset of the training data, and a different subset of features is considered for each split. This approach not only reduces overfitting but also improves the generalization of the model. Overall, our proposed model of heart disease prediction using Random Forest and SVM is a unique and innovative approach that combines the strengths of two powerful machine learning algorithms. Its performance is compared with other existing models using common evaluation metrics such as accuracy, sensitivity, specificity, and AUC.

IX. FUTURE SCOPE

The future scope for "Heart Disease prediction using a hybrid model" of SVM and RF is promising. Several potential areas of improvement and research can be explored, such as:

- Exploring the use of more advanced machine learning algorithms, such as deep learning, to improve the accu- racy of the heart disease prediction models. Deep learning models can learn complex patterns and relationships from large amounts of data and may provide more accurate predictions.
- Developing a user-friendly interface that can be used by medical practitioners to predict the likelihood of heart disease quickly and accurately. This can help save time and resources and improve patient outcomes.

Overall, the future scope for heart disease prediction using a hybrid model of SVM and RF is promising, and further research in this area can significantly improve our ability to predict and prevent heart disease.

REFERENCES

[1] Senthilkumar Mohan; Chandrasegar Thirumalai; Gautam Srivas- tava(2019):"Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques", IEEE Access (Volume: 7) , ISSN: 2169-3536.
[2] R. Goel and A. Jain. (2018) "The Implementation of Image En- hancement Techniques on Color n Gray Scale IMAGEs," 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC), pp.204-209, doi: 10.1109/PDGC.2018.8745782
[3] Archana Singh, Rakesh k. (2020)." Heart disease Prediction Using machine Learning Algorithms" International Conferences on Electrical and electronics Engineering (ICE3)
[4] Yanwei Xing, Jie Wang and Zhihong Zhao,Yonghong Gao. Combination data mining methods with new medical data to predicting outcome of coronary heart disease.

[5] Shadab Adam Pattekari and Asma Parveen "Prediction System for Heart Disease Using Na¨ıve Bayes" International Journal of Advanced Computer and Mathematical Sciences ISSN 2230-9624. Vol 3, Issue 3, 2012, pp 290-294.
[6] Mrs.G.Subbalakshmi (M.Tech), Mr. K. Ramesh M.Tech, Asst. Pro- fessor Mr. M. Chinna Rao M.Tech,(Ph.D.) Asst. Professor, "Deci- sion Support in Heart Disease Prediction System using Na¨ıve Bayes" G.Subbalakshmi et al. / Indian Journal of Computer Science and Engineering (IJCSE)2011.
[7] Jesmin Nahar, Tasadduq Imama, Kevin S. Tickle, Yi-Ping Phoebe Chen "Association rule mining to detect factors which contribute to heart disease in males and females" Expert Systems with Applications 40 (2013) 1086–1093.