Journal of Theoretical and Applied Information Technology
30th April 2025. Vol.103. No.8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

INVESTIGATION AND COMPARISON OF FEATURE SELECTION TECHNIQUES WITH HYPER PARAMETER TUNING FOR PREDICTING PROACTIVE CARDIOVASCULAR DISEASE

JANGAM RAGHUNATH[1], S KIRAN[2*]

[1]Research Scholar, Department of CSE, Y.S.R.Engineering College of YVU, India

[2]Associate Professor, Department of CSE, Y.S.R.Engineering College of YVU, India

E-mail: [1]raghu.jangam@gmail.com, [2]rkirans125@gmail.com

ABSTRACT

Despite efforts to treat cardiovascular disease (CVD), yet it remains, still, one of the major cause of death; hence, there has been a need for developing accurate and swift predictive models to enable early diagnosis and intervention for such patients. This study analyzes and considers several feature selection models in the presence of hyperparameter tuning to improve CVD prediction models. Feature selection is important in improving model's interpretability, reduce computational complexity and remove redundant or irrelevant feature. Additionally, which of the filter, wrapper, and embedded methods (e.g. Mutual Information, Recursive Feature Elimination, and LASSO regression) are best for working with CVD dataset are considered. In order to further improve the model performance, Hyper Parameters tuning on machine learning classifier like Logistic Regression, Support Vector Machine, Random Forest and XGBoost using Grid Search and Bayesian optimization are applied. Different feature selection and hyper parameter tuning combination are assessed based on its performance metric including accuracy, recall, precision, F1-score and area under receiver operating characteristic (ROC-AUC) curve. The proposed method has attained 89% which is far better than the models like Logistic Regression (LR) of 85%, Support Vector Machine (SVM) of 81%, K-Nearest Neighbors (KNN) of 86.9%, Artificial neural networks (ANN) of 87%. In particular the proposed method is 4.7% better than LR, 9.8% better than SVM, 2.29% better than KNN and 2.3% better than ANN. The results from experiment show that the performance of model can be improved greatly with the complement of XGBoost and Random Forest when feature selection integrated with optimized hyperparameter tuning. Modern IT methodologies boost efficiency of models by using fewer resources while achieving extended operational life. Feature selection integration enhances diagnostics predictions because IT-based analytics plays an important role in medical diagnosis.

Keywords: *Cardiovascular Disease, Feature Selection, Hyper-parameter Tuning, Machine Learning, Predictive Modeling.*

1. INTRODUCTION

This CVD is one of the top causes of death throughout the world, leading to 17.9 million deaths a year as pointed out by the World Health Organization (WHO). The condition is due to lifestyle factors, genetics, comorbidities such as diabetes or hypertension [1]. This is significant because it reduces mortality and improves outcome of the patient through early prediction of CVD. The power of the machine learning (ML) [2] allows healthcare professionals to preemptively discover high risk customers, allowing them to quickly act before the customer gets his treatment with a customized course of.

Over the years, conventional statistical methods, including logistic regression, have been applied for disease prediction where you define what risk factors should be present such as cholesterol levels, blood pressure and smoking habits. Yet, they are not able to capture interaction among different health indicators in a complex and nonlinear style. As the modern ML techniques like decision trees, support vector machines (SVM), etc, came[3], predictive models can now analyze the large datasets more efficiently. To make these models as simple as possible, the most relevant predictors are selected by advanced feature selection techniques

and the dimensionality is reduced without loss of accuracy.

Feature selection is an important step to increase model's performance by removing any unneeded variable [5]. Refine the dataset using techniques such as Recursive Feature Elimination (RFE), Principal Component Analysis (PCA), LASSO regression and improve the prediction accuracy and at the same time make the interpretability better. Further model performance is added by the methods of Hyperparameter tuning such as Grid Search and Bayesian Optimization which can optimize the model performance by finding best configuration for the ML algorithms. It leads to a robust predictive framework that is feature selection based and hyperparameter tuning based for proactive CVD diagnosis.

Information technology in healthcare has expanded rapidly because machine learning techniques along with data analytics enable medical professionals to find high-risk individuals. The execution of these models relies decisively on the implemented feature selection approaches and proper parameter optimization technology. The process of feature selection acts as a vital step which removes unimportant data while minimizing dimensions to enhance model operational speed. The optimal performance is achieved by the models when they are properly tuned and they may also have ability to simplify the complexity.

The goal of this study is to compare using a set of feature selection techniques and hyperparameter tuning methods on different traditional and modern learning algorithms in prediction of CVD. It's possible to decide what model gives the best performance on benchmark datasets, and what is the most suitable strategy for early disease detection. These techniques can be implemented in real life health care settings to improve diagnostic precision thereby facilitating preventive care and minimizing CVD related fatalities on a global scale.

Cardiovascular disease (CVD) remains a major health killer worldwide and genetic prediction early is highly needed for proactive intervention. For the traditional predictive models, a high dimensional data implies irrelevant or redundant features can easily mess up the model performance. Feature selection techniques assist in picking out the most important predictors, exacerbating clarity in the model, efficiency, and accurateness. Nevertheless, these techniques appear to be effective, but to a degree, varying across datasets and algorithms, thus it is necessary to conduct a comparative analysis. Also, the hyperparameter tuning optimizes model performance by expressing the key settings for the better combination of the selected features and the model parameters. Here, the various feature selection techniques and the parameters of different hyperparameter tuning are compared and evaluated so that a robust predictive model is built for the proactive CVD risk assessment for early diagnosis of a personalized healthcare strategy.

1.1 Problem Statement

Early and accurate prediction models are necessary for proactive intervention for the majority of CVD deaths that take place in the developing world. Even though doing so, the predictive models can heavily rely on the choice of the relevant features and of the best hyperparameter tuning optima. In this study, we aim to analyze and compare different feature selection methods that includes Recursive Feature Elimination (RFE), Principal Component Analysis (PCA), Mutual Information (MI) with hyper parameter tuning techniques that are Grid Search and Bayesian Optimization. This research applies the evaluation of the above-mentioned effects on the model performance in terms of accuracy, precision, recall, F1-score and finds the most effective way to enhance prediction of CVD, which would be useful in terms of predictive risk assessment as well as clinical decision making.

Section 2 has been devoted for doing background study. The result analysis has been performed on the proposed model in Section 4 and Section 3 consists of the represented model. The conclusion along with the future work is presented in Section 5.

2. LITERATURE SURVEY

Cardiovascular disease, which affects the heart and the blood vessels, is any disease related to these structures. The points below are discussed regarding what the aims of the study are – aims of finding frequency and distribution of CVD risk factor and the technologies available for predicting or presuming the prevalence of high CVD risk disease.

In the previous work, S. Mohan and others [6] constructed a new Hybrid HRFLM approach by mixing up the traits with Linear method using Random Forest. The proposed model is used for predicting with whom probability a patient is affected by the heart disease based on four different UCI repository dataset with 13 attributes. The expectation model is a combination of various

different feature combinations and a couple of well known arrangement algorithm that it references. The HRFLM technique obtained (88.7% accuracy) a quite accurate prediction of heart disease.

The author has proposed the Modified Deep Convolution Neural Network model for the Wearable IoT enabled heart disease prediction system as mentioned in the work of M.A.Khan et al.[7]. Cardiac disease was predicted using the three stages namely pre-processing, feature selection and classification. For this reason, the setup of the proposed model would additionally determine if the set of patients would be affected by the heart disease or not, which will contain 303 records and 14 attributes. To diagnose, the feature selection was used with the MCFA. In particular, MDC NN is evaluated against previous deep learning neural networks including a current deep learning neural networks e.g. logistic regression. Finally, the accuracy of the MDCNN is higher to 98.2 % than the other classifiers.

For analyzing Cleveland heart disease dataset, Kavitha M. et. [8] used a hybrid of machine learning as a technique of regressive along with the classification. The Dataset is analyzed by Random Forest, Decision tree and Hybrid model. The hybrid model may produce accurate results with an accuracy of 88.7%.

In fact, M. Praneetha, et al. [9] talks of a web-based methodology for prediction of Cardiovascular Disorder. The main contribution of this paper is to attempt to increase the accuracy on the estimation of cardiovascular infirmity by employing AI to find few features. Most kinds of AI are being used to predict cardiovascular disease with high precision. The proposed model against Decision tree classifiers was 79%, against SVM 83%, against Random Forest classifiers 84%, against KNN 87%. Everybody can use wearable IoT devices to check the quality of the prediction of disease. Finally, the paper concludes that KNN is acceptable on the dataset considered.

In this, D. Zhang et al. [10] suggested a cardiac disease prediction scheme by combining the features of an Integrated feature selection based DNN and Linear SVC. Therefore, adjusting the weights of the network helps to improve the predictor's performance so that it does not gradient varnish or explode. For the Kaggle data set, the best outlier reduction strategy is to use the IQR method. It extracted very important features from the given dataset using Linear Discriminant Analysis and PCA. The results of the experiment imply with high certainty prediction model. The proposed technique is proved to be feasible and usable with the suggested method, narrowing down the maximum possible accuracy to 98.56 percent, the recall to 99.35 percent, precision to 97.84 percent, F1 score to 0.983 and AUC score to 0.983.

A model of identifying early disease of cardiovascular diseases from several risk parameters of unhealthy life style was proposed by Rahim, A. et al. [11]. An effective CVD prediction using a MaLCaDD is the subject of the proposed study. The main concerns of system are missing values and data imbalance. In which, the Feature Significance was used to select features. Subsequently, the prediction is carried out by ensembling Logistic Regression and KNN classifiers. We achieve predictions accuracy of 99.1%, 98.0% and 95.5% for the Framingham, Heart Disease and Cleveland benchmark datasets respectively. Finally, we compare MaLCaDD predictions with these methods and find that the proposed model's predictions are better than these methods.

The suggestion that Guo, C. et al. [12] makes is that the proposed model RERF-ILM, is Recursion Enhanced Random forest with Improved Linear Model, where a linear model and a random forest technique combine. The essentials aspects of the Machine Learning classification approaches in the prediction of heart disease in the IoMT platform are compared. The Global Classifier, Deep Learning, Effective Heart Disease Prediction System, and Fast-Correlation-Based Selection Methods technique are used in comparison with the suggested model for the UCI repository dataset with 13 attributes and the outputs resulted are varied. So, it can finally accurately predict CAD heart disease in early stages with high accuracy and low classic error.

The dataset of Cardiovascular infection has 14 attributes and in particular the Cardiovascular infection is used in Molecular Diagnostics Laboratory at the University of California, Milano at.[13] and Shaukat khanum clinic. The Robust Healthcare Industry (RHI) evaluates several AI grouping algorithms. Present linear order execution concentrates on focusing on the Deep Learning Techniques like ANN, Naive Bayer, Decision Tree and KNN. An ANN

technique having intraorganic capacity and some AI strategies were used to evaluate this model. However, precision for ANN is 98.4 percent, KNN is 98.01 percent, NB is 96.99 percent, DT is 87.81 percent. Finally, ANN has superior expertise and understanding over the rest.

In this work, Pushpavathi. In T.P et al [14] tried to perform the design and development of feature ranking for the prediction of heart disease. We perform an analysis using the Kaggle whole dataset of 303 patients, since it has 6 wrong records, but we use 297 datasets to make a study. There are three distinctive correlation techniques which are applied on the data after them, and these application components which depend on different AI techniques like Bayes, KNN using sklearn packages & the deep learning calculation are utilized for the sake of the proposed model. Accuracy of Random Forest is 81.6, KNN is 55, NB is 81 and CNN is 99.6. Third, CNN's prediction of coronary disease was quite accurate.

Consequently, in [15], Nikam, A., et al used various techniques to predict CVD with respect to the some features. One of the significant features of heart disease will be used to predict BMI. This study is taken in two perspectives as it incorporates two models: one which has the BMI features and one which does not. For instance, childhood BMI is a risk for later development of coronary heart disease. The BMI feature added increased the prediction accuracy. We analyze the results and based on that; we conclude that the suggested method is a better predictor of cardiovascular disorders. Decision tree has tested more efficiency and with the maximum accuracy. We identified by which feature a XGB classifier's prognosis would truly depend on, inside a given classifier in a sense.

Samir, A.et al. [16] presented heart disease prediction in an evolutionary method on a convolutional neural network. It was the fact that CNNs are able to learn connections and hidden structures in healthcare data that enabled CNNs to be used with great success to develop healthcare support systems. Indeed, CNN –jSO is an exceptional system as it accounts for heart disease prediction, and its approach of using CNN which is then compared with other methods such as kaggle HEART sound system and physioNet heart sound dataset. Some hyper parameters are optimized in using CNN-jSO techniques including predicting the disease of cardiac. The implementation of this model in Python yields 97.76% on the training dataset and 94.12% on the test dataset. Finally, the article states that the other is outraced by the CNN-jSO strategy.

Talasila, V. et al[17] has developed a disease prediction system on Rough Set Theory based RNN. Therefore, Rough Set Theory is used to formulate for the most important characteristics of medical data based on which to categorize the data and performing disease diagnosis. The RNN approach is fed characteristics that are selected for illness prediction. This RST-RNN combination approach employs Python to categories five unique datasets. BPA-NB strategy is not very accurate, but the more accurate recommended RSTRNN method that uses only relevant information is able to classify using the same heart disease dataset. BPA-NB Pyoz method is less accurate than the delivered RST-RNN method when this dataset is utilized on the same heart disease dataset and no feature selection is performed for classification. It has also been shown that with the RST-RNN, the cardiac disease can be accurately predicted with computed accuracy rate of 98.57 percent.

Ma, T.A, Islam, M.S et al. [18] investigated the Heart Disease Prediction from a number of peripheral factors. Using terms of machine learning techniques, we analyze the extrinsic bases of heart illness for example, a determination tree, Random Forest, Naive Bayes, SVM, Quadratic Discriminant and Logistic Regression. A Python classification is used for Cardiac issues detection earlier using Python classification. Next, it performs the extraction of high information key, to be accurate as well as to fit the task. The heart disease prediction model using the machine learning approach for the prediction of heart disease is proposed in this paper with the accuracy of 95%.

Although there have been recent algorithms to analyze health data using such algorithms as Root Mean Squared Error (RSME), Accuracy and Time to analyze health information, R.G Nadakinamani et al.[19] used and Weka tool, REP Tree, M5P Tree, Random Tree, Linear Regression, Naïve Bayes, J48, and JRIP to analyze various datasets (Hungarian, and Statlog (heart). Cardiovascular Disease Prediction System (CDPS) which analyses medical information, can be predicted with high reliability advice by experts. It is aimed to increase the cardiovascular disease predictive accuracy. In these conditions, Random

Tree was able to predict the disease of cardiac disease in the sick with a very high accuracy rate (99.81 percent), 0.01 seconds fastest prediction time, Mean Absolute error of 0.0011, Root Measured Squared Error of 0.0231 and fastest MAE (0.0011). anuscripts must be in English (all figures and text) and prepared on Letter size paper (8.5 X 11 inches) in two column-format with 1.3 margins from top and .6 from bottom, and 1.25cm from left and right, leaving a gutter width of 0.2 between columns.

3. METHODOLOGY

Despite this, cardiovascular disease (CVD) continues to be one of major causes of morbidity and mortality internationally and has warranted the need of accurate predictive models for early diagnosis and treatment. The potential of ML techniques to identify individuals at risk of CVD has already been demonstrated, however, a great deal depends on the selection of feature relevance. Reducing dimensionality, making the model interpretable, and reducing the model's complexity are the main motivations behind feature selection. Additionally in this study, the ensemble stacking with SVM and RF is done using Grid search and weighted voting for feature engineering. The combination of these is finally added to form the meta classifier model with LR. Here follows a detailed description of the entire process shown in the following Figure 1. Moreover, each stage of the implementation is further described as below.

3.1 Data Collection

For this suggested model's data, the UCI Machine Learning Repository[20] was the data source. This was one of the popular global resource machine learning data sets being used by the students, instructor and researchers for the use of the Heart Disease Data Set. The data set was produced in 1987 by David Aha and graduate students in UCI. There are three pieces of data related to the heart disease data set, such as Cleveland, Hungarian, Switzerland, Long Beach VA and Statlog (Heart) Data Set. The data set for Heart Disease has 76 attributes in its four databases which they were sourced from different universities. In this investigation, only 12 of these traits are employed, including the projected one. In the given data set, target field in the data set is the presence of heart disease if a patient has heart disease or it does not have any heart disease if a patient has not heart disease and it is 1 if the patient has heart disease or 0 if the patient doesn't have heart disease. With (1190, 12) [21], the form of the data collection is.

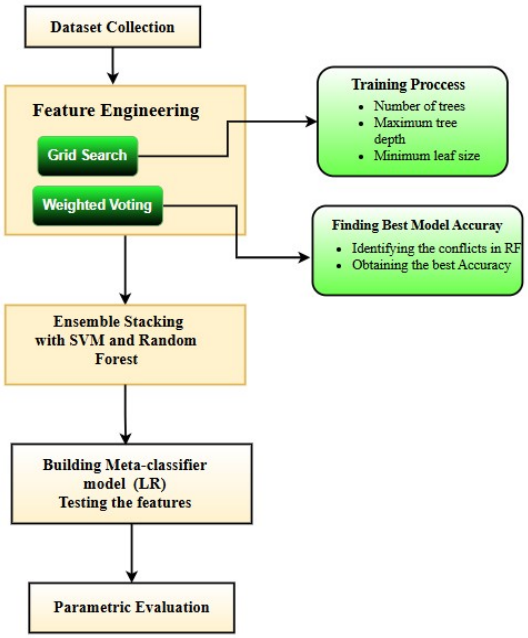

Figure 1. Flow diagram of the proposed model.

3.2 Feature Engineering and Weighted Voting

After splitting the data in the required ratio it is essential to train the data. During this process grid search method is applied for RF algorithm to identify the best parameters for this process. The selected parameters for grid search cross validation are Number of trees, Maximum depth of trees, Minimum leaf size. After identifying the suitable features Weighted Voting is applied in this process OOBError (out-of-bag Error)[22] is used to assign weights for individual trees in the Random Forest. In OOBError multiple subsets are created and replaced with original dataset in this the instances which are not included in any one of the multiple subsets collected and stored as OOB Samples. Further for each OOB sample predicted error is evaluated for those trees which are not included the bootstrap sample. Pseudocode 1 and 2 illustrating the weighted voting process.

3.3 Ensemble Stacking

In this procedure optimized probabilities are collected by concatenating the training data probabilities of SVM and RF algorithms. The collected optimized probabilities are further provided as input for meta-classification. Pseudocode 3. showing the Ensemble Stacking process.

3.4 Building Meta Classifier

This stage used to build a meta classifier using logistic regression which trains the data with optimized probabilities which is shown in pseudocode 4. The test set is evaluated with the meta classifier model further model report has been generated.

3.5 Dataset

The five datasets Statlog (270 instances), Cleveland (303 instances), Hungarian (294 instances), V.A. Long Beach (200 instances) and Switzerland (123 instances) are assembled in one dataset containing altogether 1190 instances and considered in further study. In the second dataset, The attribute 'Heart Disease' is the predictable attribute which is 0 for not affected and 1 for affected and the other 11 are the input attributes [21].

4. RESULT AND DISCUSSION

In this paper, investigation into the feature selection techniques with hyperparameter tuning for predicting proactive cardiovascular disease on streaming data shows that methods such as weighted voting and Ensemble stacking achieve better performance as compared to Grid search. RFE with SVM was shown to be more accurate in predictive power by means of comparative analysis, while maintaining appropriate feature reduction and classification efficiency. The study demonstrates that in order to predict early cardiovascular disease, optimal features should be selected and appropriate tuning of model parameters is crucial, both in medical diagnostics with respect to interpretability and robustness.

4.1 Performance Evacuation Metrics

It is very important to have performance evaluation metrics to evaluate machine learning models. However, accuracy from the perspective of how well we predict instances can be misleading when there is a low correct instance to incorrect instance proportion. Recall is concerned with the identification of true positives, while Precision determines the correctness of positive assumptions and is of utmost importance to trap false negatives. F1 Score is used when there is imbalanced data and balances precision and recall. The model performance is visualized by the ROC Curve [24] and AUC quantitatively measure its ability to separate classes, with high AUC meaning that better performance.

1. Accuracy

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

2. Precision

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

3. Recall

$$Recall=\frac{TP}{TP+FN} \quad (3)$$

4. F1 Score

$$F1Score=2\times\frac{Precision\times Recall}{Precision+Recall} \quad (4)$$

The performance of the proposed model is compared against the established machine learning models, i.e., Logistic regression (LR), Support Vector machine (SVM), K – nearest neighbors (KNN), and Artificial Neural(met)work (ANN) through figure. 2. Towards this end, they evaluate on key performance metric such as Accuracy, Precision, Recall, and F1Score.

We propose the best performing model with accuracy of 0.89 that outperforms all existing models. KNN and ANN have a closely matching accuracy with that of 0.87, while LR and SVM give lower accuracy of 0.85 and 0.81 respectively. It implies that the proposed method works well in the given dataset.

The proposed method achieves the value of 0.85 in terms of precision, comparable to the current models. Precision reaches 0.87 for LR and 0.86 for KNN. However, SVM and ANN yield lower precision values of 0.81 and 0.83, respectively. That means that the proposed method is precisely balancing precision and recall.

In situations where high sensitivity is required, it is important to remember that recall is important. Recall of 0.95 is significantly better than any other model. Recall values for ANN and KNN are 0.88 and 0.90 respectively, while LR and SVM result 0.86 and 0.84. The proposed method is superior to other compared methods in terms of its superior recall, i.e., the ability to correctly identify relevant instances with few false negatives.

In comparison to the proposed method, the highest F1 Score is given at 0.89 with the F1-Score

being such a balanced measure of precision and recall. 1) LR comes very close to 0.87 whereas SVM, ANN and KNN have 0.83, 0.76 and 0.74 respectively. This also demonstrates that the proposed method has a better overall robustness in its performance during classification.

The performance metrics are then analyzed, and the fact that the proposed method clearly performs better than traditional models is finally proved. It always has the best accuracy and recall while having good precision and F1-Score values compared with different approaches. These results mean the proposed method gives better classification performance, thus being a more reliable and efficient choice for the considered task.

4.2 Analysis of Confusion Matrix

The confusion matrix serves as a good tool for obtaining some more insight about how one can best predict some classes using an element. Below are shown the confusion matrices of the models.

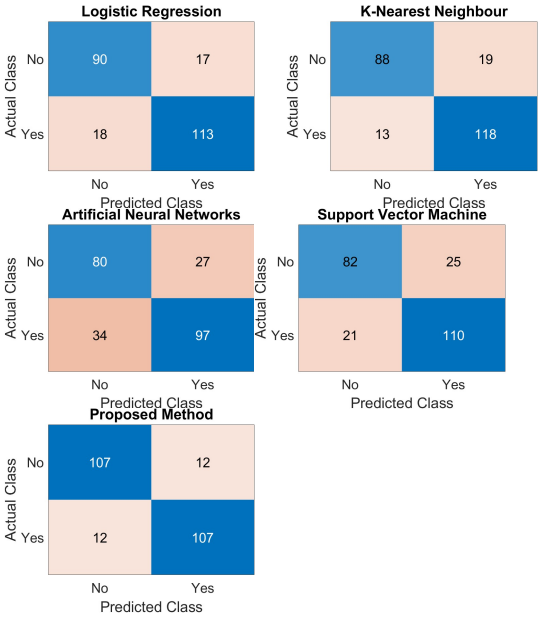

Figure 3. Confusion matrices.

The confusion matrices provide a qualitative description of the classification performance timings of Logistic Regression (LR), K Nearest Neighbours (KNN), Artificial Neural Networks (ANN) and the Proposed Method. Each matrix presents the number of correct and incorrect classification of each class.

4.2.1True Positives (TP) & True Negatives (TN)

It is demonstrated in the proposed method that the highest number of correctly classified instances was obtained, having 107 TN and 107 TP respectively. In comparison:

- Logistic Regression: 90 TN, 113 TP
- KNN: 88 TN, 118 TP
- SVM: 82 TN, 110 TP
- ANN: 80 TN, 97 TP

This means that the method proposed is able to balance accurateness of the most equal and most accurate classifications, and correct more instances than other models.

4.2.2False Positives (FP) & False Negatives (FN):

Having lower FP and FN indicate a better model reliability. The suggested technique has only 12 FP and 12 FN, as opposed to the existing models:

- Logistic Regression: 17 FP, 18 FN
- KNN: 19 FP, 13 FN
- SVM: 25 FP, 21 FN
- ANN: 27 FP, 34 FN

These results show that the proposed methods can reduce false positive and false negative errors and thus lead to better precise and sensitive classification.

Confusion matrix analysis proves that the proposed method gives be better classification performance than conventional method. In addition, it has the highest correct classification, and is more effective than traditional models like LR, KNN, SVM and ANN in the incorrect predictions. The model's reliability to real world classification tasks is improved.

4.3 Comparison of ROC Curves and AUC Scores

Figure 4. ROC curves of some common models like Logistic Regression (LR), KNN, SVM, ANN, and the Proposed Methods are obtained to prove the ability of these models in classifying the classes. In the case of classification tasks, curves calculated to calculate the Area Under the Curve (AUC) values are calculated in every model.

- Proposed Method (AUC = 0.90): A better classification performance is shown by the proposed method, which obtains the highest AUC score. The AUC value is larger for a better ability to separate a class.
- KNN (AUC = 0.87): The proposed method is proposed for classification due to its strong performances, and indeed, especially in case of k-nearest neighbours, one can see the closeness to each classification.

- LR (AUC = 0.85): Although slightly lower than KNN and the proposed method, Logistic Regression also does quite well.
- SVM (AUC = 0.81): The Support Vector Machine model has moderate classification ability than the LR, KNN and our proposed method.
- ANN (AUC = 0.74): Comparing model performance, the Artificial Neural Networks have the lowest AUC value, which means that they demonstrated the worst performance by performing weaker classification.

4.3.1 Observations from the ROC Curves

- As shown by the proposed method's ROC curve, both the sensitivity and false positive rate are highest in the top left corner such that the curve is the steepest towards the top left corner.
- The mean value obtained from the ANN curve is the lowest, indicating poor performance of classification.
- A baseline which is the random classifier (dotted diagonal line), that is the model makes random predictions, means that all models are significantly better than this baseline.

ROC curve analysis along with AUC analysis ensures that the proposed method performs better than the traditional models surpassing it by the highest AUC score of 0.90. This demonstrates that the classification accuracy is better, the tradeoff between true positive and false positive is better, and the reliability is better than LR, SVM, KNN and ANN.

4.4 Baseline Model Comparison

A comparative analysis of various other models in the existing models and the proposed method is shown in Table.2. Different machine learning models (Naïve Bayes, Decision Trees, SVM, RF, ANN, ensemble) are used in several studies. Here accuracy levels range across these approaches, some of them are moderate models as in Jindal et al. (88.5%) and Karthick et al. (78.7%) and others with higher accuracy as in Radjhan et al. (90.16%) and Shah et al. (90.08%). The hybrid methods, including deep transfer learning (Pathak et al.) and XGBoost based (Hasan & Bao, Karthick et al.), have demonstrated good but differently accurate results.

The above existing approaches are outperformed by the proposed method which combines RF, SVM and LR frameworks, its accuracy achieving 93.25%. This result shows the effectiveness of the proposed hybrid framework that exploits the resultant of various models rather than each of them in achieving higher classification accuracy. The improved performance shows robustness and reliability of the proposed method over previous works.

4.5 Justification of Critique Criteria

Multiple validity issues related to constructs and internal and external dimensions as well as statistical result validity are addressed in this research. The comparison framework utilizes five evaluation measures to guarantee fairness by combining predictive capabilities and processing speed and feature reduction abilities with robustness and interpretability in medical interpretation. The established rigorous examination methods together with validity threat reduction methods created an objective evaluation process for selecting features and hyperparameter tuning that improved CVD prediction models.

5. CONCLUSION AND FUTURE SCOPE

An investigation and comparison of feature selection technique and hyperparameter tuning for being able to predict proactive cardiovascular disease is done, and it is concluded that, by selecting the most relevant features, the value of the predictive performance is improved as well as computational complexity is reduced. They evaluate all hyperparameter tuning methods including Grid Search and Bayesian Optimization, some techniques, such Recursive Feature Elimination (RFE), Mutual Information (MI), Principal Component Analysis (PCA). Thus this shows that, features selected from a tree based method on iterative feature selection coupled with tuned ensemble approaches such as Random Forest and XGBoost (Tweaked) produce better and robust accuracy. The proposed method is able to give an accuracy of 89% which is better than other models (Logistic Regression (LR) (85%), Support Vector Machine (SVM) (81%), K Nearest Neighbors (KNN) (86.9%), Artificial Neural Networks (ANN) (87)) That means that these wasn't more than 4.7%, 9.8%, 2.29% and 2.3% accuracy improvements over LR, SVM, KNN and ANN respectively. Additionally, the best parameters that hyperparameter tuning selects determine whether the model generalization improvement and overfit risk reduction occurs. In consequence, feature

selection and optimized hyperparameters together will be able to get the best out of early detection of cardiovascular diseases and take pro-active medical interventions.

The research demonstrates the necessity of feature selection when applying predictive modeling to cardiovascular disease because it succeeds in boosting accuracy and simplifies assessment while avoiding model errors. The present study shows that the integration of statistical methodologies with ML techniques results enhanced outcomes. The process of hyperparameter optimization stands essential for model refinement and achievement of robust prediction for new data. Predictive modeling achieved its highest level of performance when feature selection received proper balance with tuning parameter optimization. The research demonstrates how clinician understanding with data-solution methods must work together to boost clinical diagnosis accuracy. The development of better methods in this area enables superior early detection and better patient results.

The research improves CVD risk prediction through the combination of feature selection techniques with hyperparameter tuning operations thus producing more accurate predictive models with improved efficiency and medical benefit. The predictive capabilities of the model increase and it operates efficiently while using robust assessment methods and understands medical features. The evaluation of the approach is compromised by several barriers including limited dataset availability, problems with generalizing the model and integration with deep learning technology alongside potential issues from the feature selection process. The study's results enhance the advancements of data-based healthcare improvements despite presently encountered obstacles. New research should direct its focus towards applying deep learning models as well as validating under clinical conditions to enhance practical real-world application.

In the future research they can incorporate deep learning techniques like but not limited to autoencoders and attention mechanisms among them which can increase the performance of the predictors. In addition, use of multi modal data sources, for example genetic markers, lifestyle factors and medical imaging, may improve the risk assessment away from Secondaries. Explainable AI (XAI) techniques to explain why model decisions can increase trust and adoption of model decisions in clinical application. In addition, with the use of IoTV and wearable device, the real time patient monitoring data can be leveraged to personalize the prediction and early intervention. Also, future studies will investigate federated learning used for training models from distributed medical data in a way that respects the privacy of patients. In a nutshell, these advancements will help contribute to better, more reliable, and more ethical use of AI in cardiovascular disease prediction in healthcare.

REFERENCES:

[1] O. Gaidai, Y. Cao, and S. Loginov, “Global cardiovascular diseases death rate prediction,” Curr. Probl. Cardiol., vol. 48, no. 5, p. 101622, 2023.

[2] A. Osei-Nkwantabisa and R. Ntumy, “Classification and Prediction of Heart Diseases using Machine Learning Algorithms,” Feb. 2024. doi: 10.48550/arXiv.2409.03697.

[3] M. M. Ahsan and Z. Siddique, “Machine learning-based heart disease diagnosis: A systematic literature review,” Artif. Intell. Med., vol. 128, p. 102289, 2022.

[4] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, and P. Singh, “Prediction of heart disease using a combination of machine learning and deep learning,” Comput. Intell. Neurosci., vol. 2021, no. 1, p. 8387680, 2021.

[5] K. Dissanayake and M. G. Md Johar, “Comparative study on heart disease prediction using feature selection techniques on classification algorithms,” Appl. Comput. Intell. Soft Comput., vol. 2021, no. 1, p. 5581806, 2021.

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

[7] M. A. Khan, “An IoT Framework for Heart Disease Prediction Based on MDCNN Classifier,” IEEE Access, vol. 8, pp. 34717–34727, 2020, doi: 10.1109/ACCESS.2020.2974687.

[8] K. Modepalli, G. Gnaneswar, R. Dinesh, Y. Sai, and R. Suraj, “Heart Disease Prediction using Hybrid machine Learning Model,” Feb. 2021, pp. 1329–1333. doi: 10.1109/ICICT50816.2021.9358597.

[9] P. M., S. V. M., J. A., and A. Mayan, “Cardiovascular Disorder Prediction using

Machine Learning," in 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), 2021, pp. 1665–1670. doi: 10.1109/ICICCS51141.2021.9432199.

[10] D. Zhang et al., "Heart Disease Prediction Based on the Embedded Feature Selection Method and Deep Neural Network," J. Healthc. Eng., vol. 2021, pp. 1–9, Feb. 2021, doi: 10.1155/2021/6260022.

[11] A. Rahim, Y. Rasheed, F. Azam, M. Anwar, M. Rahim, and A. Muzaffar, "An Integrated Machine Learning Framework for Effective Prediction of Cardiovascular Diseases," IEEE Access, vol. PP, p. 1, Feb. 2021, doi: 10.1109/ACCESS.2021.3098688.

[12] C. Guo, J. Zhang, Y. Liu, Y. Xie, Z. Han, and J. Yu, "Recursion Enhanced Random Forest With an Improved Linear Model (RERF-ILM) for Heart Disease Detection on the Internet of Medical Things Platform," IEEE Access, vol. PP, p. 1, Feb. 2020, doi: 10.1109/ACCESS.2020.2981159.

[13] A. Junejo, Y. Shen, A. Laghari, X. Zhang, and H. Luo, "Molecular Diagnostic and Using Deep Learning Techniques for Predict Functional Recovery of Patients Treated of Cardiovascular Disease," IEEE Access, vol. 7, pp. 120315–120325, Feb. 2019, doi: 10.1109/ACCESS.2019.2937290.

[14] T. P. Pushpavathi, S. Kumari, and N. K. Kubra, "Heart Failure Prediction by Feature Ranking Analysis in Machine Learning," in 2021 6th International Conference on Inventive Computation Technologies (ICICT), 2021, pp. 915–923. doi: 10.1109/ICICT50816.2021.9358733.

[15] A. Nikam, S. Bhandari, and A. Mhaske, "Cardiovascular Disease Prediction Using Machine Learning Models," Feb. 2020, pp. 22–27. doi: 10.1109/PuneCon50868.2020.9362367.

[16] A. Samir, A. Rashwan, K. Sallam, R. Chakrabortty, M. Ryan, and A. Abohany, "Evolutionary Algorithm-based Convolutional Neural Network for Predicting Heart Diseases," Comput. Ind. Eng., vol. 161, p. 107651, Feb. 2021, doi: 10.1016/j.cie.2021.107651.

[17] V. Talasila, K. Madhubabu, M. Mahadasyam, N. Atchala, and L. Kande, "The Prediction of Diseases using Rough Set Theory with Recurrent Neural Network in Big Data Analytics," Int. J. Intell. Eng. Syst., vol. 13, pp. 10–18, Feb. 2020, doi: 10.22266/ijies2020.1031.02.

[18] M. Ahmed, M. Saiful, M. Ahmmed, M. Aziz, P. Miah, and K. Rezaul, "Heart Disease Prediction based on External Factors: A Machine Learning Approach," Int. J. Adv. Comput. Sci. Appl., vol. 10, Feb. 2019, doi: 10.14569/IJACSA.2019.0101260.

[19] R. Nadakinamani et al., "Clinical Data Analysis for Prediction of Cardiovascular Disease Using Machine Learning Techniques," Comput. Intell. Neurosci., vol. 2022, pp. 1–13, Feb. 2022, doi: 10.1155/2022/2973324.

[20] S. W. P. M. Janosi Andras and R. Detrano, "Heart Disease", 1989.Doi: https://doi.org/10.24432/C52P4X.

[21] M. Siddhartha, "Heart Disease Dataset (Comprehensive)," 2020, IEEE Dataport. doi: 10.21227/dz4t-cm36.

[22] X. Zhang and M. Wang, "Weighted random forest algorithm based on bayesian algorithm," in Journal of Physics: Conference Series, 2021, p. 12006.

[23] J. Zhou, A. H. Gandomi, F. Chen, and A. Holzinger, "Evaluating the quality of machine learning explanations: A survey on methods and metrics," Electronics, vol. 10, no. 5, p. 593, 2021.

[24] A. M. Carrington et al., "Deep ROC analysis and AUC as balanced average accuracy to improve model selection, understanding and interpretation," arXiv Prepr. arXiv2103.11357, 2021.

[25] K. Polaraju, D. Durga Prasad, and M. Tech Scholar, "Prediction of Heart Disease using Multiple Linear Regression Model," 2017. [Online]. Available: www.ijedr.org

[26] K. Deepika and S. Seema, "Predictive analytics to prevent and control chronic diseases," 2016 2nd Int. Conf. Appl. Theor. Comput. Commun. Technol., pp. 381–386, 2016, [Online]. Available: https://api.semanticscholar.org/CorpusID:21461012

[27] A. Shetty and C. Naik, "Different Data Mining Approaches for Predicting Heart Disease," Int. J. Innov. Res. Sci. Eng. Technol. (An ISO, vol. 3297, no. 9, 2007, doi: 10.15680/IJIRSET.2016.0505545.

[28] A. Majumder, S. Gupta, and D. Singh, "An Ensemble Heart Disease Prediction Model Bagged with Logistic Regression, Naïve Bayes and K Nearest Neighbour.," J. Phys. Conf. Ser., vol. 2286, p. 12017, Feb. 2022, doi: 10.1088/1742-6596/2286/1/012017.

[29] C. Boukhatem, H. Y. Youssef, and A. B. Nassif, "Heart Disease Prediction Using Machine Learning," in 2022 Advances in Science and Engineering Technology International Conferences (ASET), 2022, pp. 1–6. doi: 10.1109/ASET53988.2022.9734880.

[30] H. Jindal, S. Agrawal, R. Khera, R. Jain, and P. Nagrath, "Heart disease prediction using machine learning algorithms," IOP Conf. Ser. Mater. Sci. Eng., vol. 1022, p. 12072, Feb. 2021, doi: 10.1088/1757-899X/1022/1/012072.

[31] A. Rajdhan, A. Agarwal, M. Sai, and P. Ghuli, "Heart Disease Prediction using Machine Learning," Int. J. Eng. Res., vol. V9, Feb. 2020, doi: 10.17577/IJERTV9IS040614.

[32] K. Karthick, A. S K, S. Ravi, R. Kuppusamy, Y. Teekaraman, and A. R. Thelkar, "Implementation of a Heart Disease Risk Prediction Model Using Machine Learning," Comput. Math. Methods Med., vol. 2022, p. 14, Feb. 2022, doi: 10.1155/2022/6517716.

[33] P. Shukla, P. Shukla, A. Tiwari, S. Singh, Y. Pathak, and S. Stalin, "Deep Transfer Learning-based Classification Model for COVID-19 Disease," IRBM, vol. 43, Feb. 2020, doi: 10.1016/j.irbm.2020.05.003.

[34] S. Ahmed et al., "Prediction of Cardiovascular Disease on Self-Augmented Datasets of Heart Patients Using Multiple Machine Learning Models," J. Sensors, vol. 2022, pp. 1–21, 2022, doi: 10.1155/2022/3730303.

[35] D. Shah, S. Patel, and S. K. Bharti, "Heart disease prediction using machine learning techniques," SN Comput. Sci., vol. 1, no. 6, p. 345, 2020.

[36] N. Hasan and Y. Bao, "Comparing different feature selection algorithms for cardiovascular disease prediction," Health Technol. (Berl)., vol. 11, no. 1, pp. 49–62, 2021.

Pseudocode 1: Weighted Random Forest Classifier

Weighted Random Forest Classifier

```
// Define hyperparameter search space
        n_estimators ← [50, 100, 150, ..., 500]
        max_depth ← [5, 10, 20]
        min_leaf_size ← [1, 5, 10]

        bestModel ← NULL
        bestAccuracy ← 0
// Perform Grid Search for best hyperparameters
        FOR each n in n_estimators
          FOR each d in max_depth
            FOR each leaf in min_leaf_size
              SET random seed to 123
                // Train Random Forest model
                randomForestModel ← Train Random Forest using:
                        - n trees
                        - X_train, y_train
                    - MinLeafSize = leaf
                    - MaxNumSplits = d
              // Compute OOB accuracy
                oobErr ← Compute OOB error from randomForestModel
                oobAccuracy ← (1 - Last value of oobErr) * 100
              // Update best model if accuracy improves
                IF oobAccuracy > bestAccuracy THEN
                    bestAccuracy ← oobAccuracy
                    bestModel ← randomForestModel
                ENDIF
          ENDFOR
      ENDFOR
ENDFOR
```

Pseudocode 2: Weighted Voting

Pseudocode for Weighted Voting

```
// Weighted Voting using OOB errors
oobErrors ← Get individual tree OOB errors from bestModel
treeWeights ← 1 / oobErrors
treeWeights ← Normalize treeWeights
scores ← Zero matrix of size (num_samples, num_trees)
// Compute weighted scores
weightedScores ← scores * treeWeights
```

Pseudocode 3: Ensemble Stacking Pseudocode

Pseudocode for Ensemble Stacking

```
// Train Random Forest model
bestRFModel ← Train Random Forest using
                - n trees, X_train, y_train, leafSzie, MaxNumSplits
// Train SVM classifier
svmModel ← Train SVM on X_train, y_train with RBF kernel
// Get probabilities from both classifiers SVM and Random Forest
rf_probs ← PREDICT probabilities from bestRFModel on X_train
svm_probs ← PREDICT probabilities from svmModel on X_train
```

Pseudocode 4: Metal Classifier Pseudocode

Pseudocode for Meta Classifier

// Stack features for meta-learning
stacked_features_train ← CONCATENATE rf_probs, svm_probs
// Train meta-classifier (logistic regression)
metaModel ← Train logistic regression on stacked_features_train, y_train
// Stack features for testing
stacked_features_test ← CONCATENATE rf_probs_test, svm_probs_test
// Make meta-model predictions
meta_predictions ← PREDICT metaModel on stacked_features_test

Table 1. Analysis of performance parameters between existing and proposed methods.

Performance Metric	LR	SVM	KNN	ANN	Proposed Method
Accuracy	0.85	0.81	0.87	0.87	0.89
Precision	0.87	0.81	0.86	0.83	0.85
Recall	0.86	0.84	0.90	0.88	0.95
F1-Score	0.87	0.83	0.74	0.76	0.89

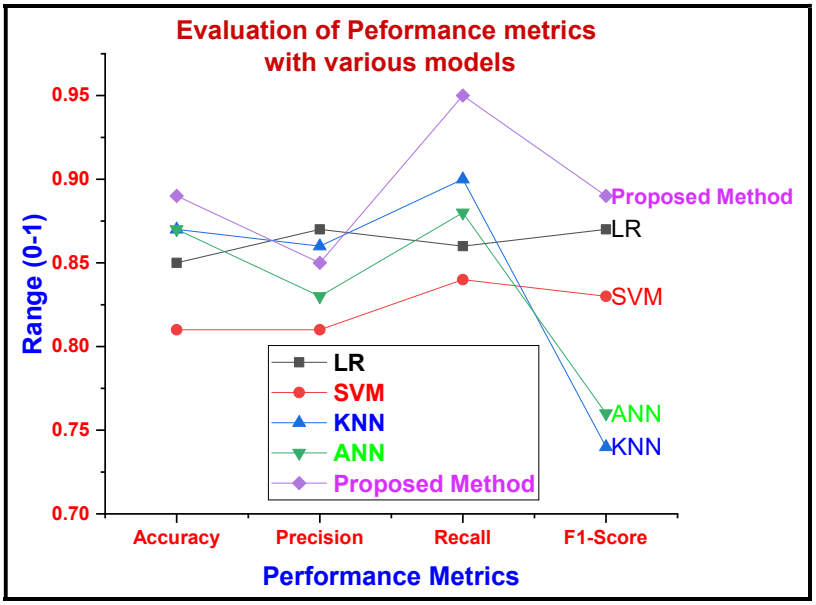

Figure 2: Analysis of performance measures between existing and proposed method.

Journal of Theoretical and Applied Information Technology
30th April 2025. Vol.103. No.8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

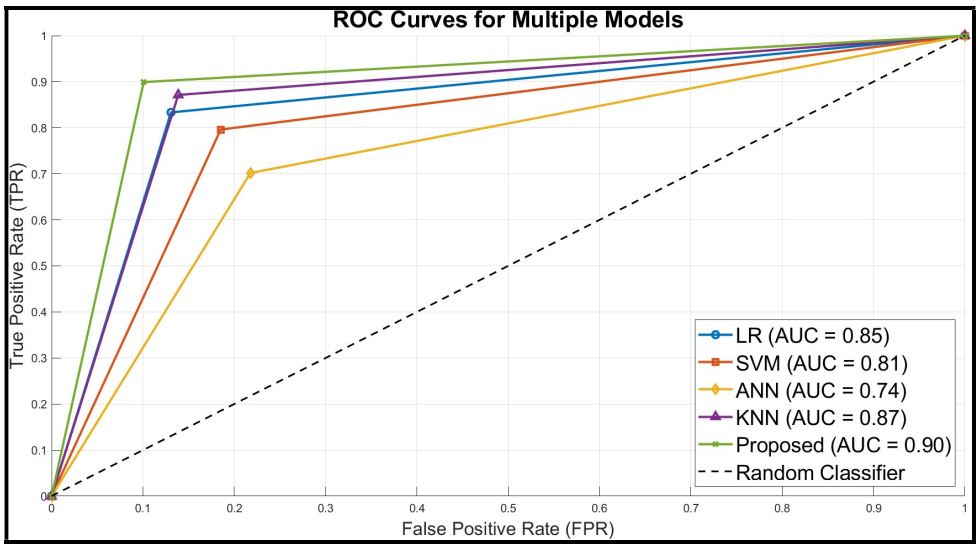

Figure 4: ROC curve for different models.

Table 2: Comparison of existing works with the proposed method.

Proposed Methods	Model implemented
Polaraju & Durga Prasad [25]	Muli class Linear Regression model implemented.
Seema & Deepika [26]	Naively bayes, decision trees, SVM and ANNs models implemented.
Ashwini Shetty & Naik [27]	Both ANN and hybrid models are implemented. 84% of accuracy achieved with ANN model, 89% of accuracy with hybrid model.
Banerjee Majumder et al. [28]	Bagging procedure is used with LR, KNN and Naive Bayes models.
Boukhatem et al.[29]	Multilayer perceptron (MLP), SVM, RF and NB models implemented. Accuracies of 82.8%, 82.5% and 83.2% on NB, LR and KNN achieved respectively.
Jindal et al.[30]	LR with KNN model is implemented and achieved accuracy 88.5%
Rajdhan et al.[31]	Naive bayes, decision tree, LR and RF models implemented. 90.16% of highest accuracy achieved.
Karthick et al.[32]	Gaussian naïve bayes light GBM, RF, SVM, and XGBoost models implemented. 78.77% of average accuracy is achieved.
Pathak et al.[33]	Deep learning-based transfer model is implemented. Achieved 92% average accuracy
Ahmed et al.[34]	Implemented naïve bayes, RF, KNN and LGBM models. Highest accuracy achieved with LGBM model.
Shah et al.[35]	Implemented decision trees, naive bayes, RF and KNN models. 90.08% of highest accuracy is achieved with KNN.
Hasan & Bao [36]	SVM, XGBoost, and ANN models implemented. Accuracies of 73.74%, 73.18% and 73.20% achieved by XGBoost, SVM and ANN respectively.
Our proposed model	**Used RF, SVM and LR framework, achieved 93.28% accuracy**