Kalpa Publications in Computing

Volume 19, 2024, Pages 420–431

Proceedings of 6th International Conference
on Smart Systems and Inventive Technology

An Analysis of Performance and Dataset Dynamics in the Early Detection of Cardiovascular Diseases

S Kartikaaditya[1], T Divya Teja Reddy[2], S Reshmi Panda[3], P Sanjay Vardhan[4], and Sarath S[5]

[1, 2, 3, 4, 5] Department of Computer Science and Engineering, Amrita School of Computing, Amrita Vishwa Vidyapeetham, Amritapuri, India.

kartikaditya6@gmail.com[1] divyatejareddytadi@gmail.com[2]
sahukarreshmipanda@gmail.com[3] sanjayvardhanpadala@gmail.com[4]
sarathS@am.amrita.edu[5]

Abstract

Cardiovascular Diseases (CVD) are the most prevalent global health concern that demands prompt attention given their substantial role in increasing mortality figures. Owing to the need for early detection to alleviate the inimical effects of CVD, this study makes extensive use of machine learning techniques including Support Vector Machine (SVM), AdaBoost, XGBoost, and Decision Tree in the early prediction of cardiovascular diseases. The robustness of the model will be enhanced by assessing three diverse datasets enriched with various types of patient information to derive the most efficient model. Through this study we conduct thorough performance evaluations, considering various evaluation metrics such as Accuracy, Sensitivity and False positive rate, aiming to identify the most effective machine learning model for early CVD detection. The results help shed light on important findings that can lead to improved outcomes, which help in the fight against cardiovascular diseases.

Keywords: Cardiovascular disease, classification algorithms, decision tree, model evaluation.

1 Introduction

Being a formidable global health challenge, CVD have been casting a long shadow over public health. Some of the most significant behavioral risk factors that fuel cardiovascular diseases are unhealthy diet, physical inactivity, high blood pressure, and tobacco use, and can also be due to various other detrimental lifestyle choices. Timely intervention can significantly reduce the burden on healthcare systems while improving the quality of life of individuals affected. The stealthy onset of the majority of CVD in their nascent phase emphasizes the critical need for

early detection, as delayed diagnosis often results in irreparable damage and an increased vulnerability to dire consequences.

To advocate for effective solutions, this investigation is centered on four models Support Vector Machine (SVM), AdaBoost, XGBoost, and Decision Tree to identify the most efficient model for early detection of CVD. Support Vector Machines (SVM) are known for their ability to handle complex data patterns. AdaBoost is an ensemble method that combines weak learnersfor prediction. XGBoost is an efficient gradient-boosting algorithm. Decision Tree is a versatile machine-learning technique, that excels in complex decision-making.

Following meticulous preprocessing of the health datasets, which includes data cleaning, feature scaling, and missing value resolution, each model is applied to a set of three varied datasetsof patient information. This study aims to analyze the performance of these models on three key datasets through an extensive analysis report comparing these models using quantitative evaluation metrics such as precision and False positive rate find the most optimal one for early detection. Furthermore, this study investigates how the characteristics and the diversity of the datasets affect the performance of the predictive models.

In the context of early detection of Cardiovascular Prediction, the effectiveness of the model to accurately predict the True positives is crucial. This Comprehensive analysis focuses on Sensitivity(True Positive Rate) demonstrating the importance of predicting True positives Thereby, Contributing to the field by comparing various predictive models, facilitating the early andprecise prediction of people at risk for cardiovascular disease.

2 Related work

Many previous studies have underscored the vital need for early detection frameworks in addressing the prevalent global health threats posed by Cardiovascular diseases. The revelations brought by the research in the field of machine learning for the prediction of CVD have shown an increase in inclination towards using notable machine learning techniques such as Support Vector Machines (SVM), XGBoost (XGB), AdaBoost, and decision tree models thus demonstrating the potential of these algorithms to unlock new insights in cardiovascular health interpretation. Recent studies have emphasized the importance of varied datasets in developing robust machine-learning models. Prior research in this field showed that the boosted model proved tobe more efficient than the base classifiers which involved AdaBoost and XGBoost[1].There is alsoa study indicating that using gradient boosting for a more robust sequential tree ensemble would be efficient to predict binary classes[2]. The recent work on heart disease prediction using the Stacking Classifiers Model[3] provides valuable benchmarks for future research in cardiovascular health analytics. This particular study has shown that good results were achieved through SVM and SVM is claimed to be a good tool for medical diagnosis[4].A study revealed that SVM based method under 5-fold cross-validation exhibits better performance than logical regression and random forest in predicting cardiovascular disease.[5].Gaining valuable insights from the extensive research, This study focuses on comparing now models to find the most effective approaches in CVD prediction."

3 Methodology

The experimental setup (fig1) comprises 4 machine-learning models: Support Vector Machine (SVM), XGBoost, AdaBoost, and Decision Trees, Following the necessary preprocessing of the considered datasets, each model is applied to a set of three varied datasets of patient information. The aim is to evaluate the performance of each model in predicting the presence or absence ofa tendency to cardiovascular diseases utilizing various features available in the datasets.

3.1 Datasets

The motive behind picking diverse datasets is to entrap a varied range of characteristics of the patient information. In this study, three different datasets, the cardio-train dataset, data-cardiovascular-risk-data, and CVD Prediction dataset were considered for extensive evaluation of models. The datasets contain categorical as well as numerical values. Hence, feature encoding is performed on the categorical values for swift analysis of model robustness. Also, null values and duplicate records are handled. The model accuracy is calculated with the help of target features extracted from each of the datasets, such as TenYearCHD in the data-cardiovascular-risk dataset, which refers to the 10-year risk of coronary heart disease CHD and similarly Heart Disease in the CVD prediction dataset and cardio in cardio-train dataset. These accuracies are used for a comparative study of the model's robustness. Dataset Dimensions is shown in Table 1.

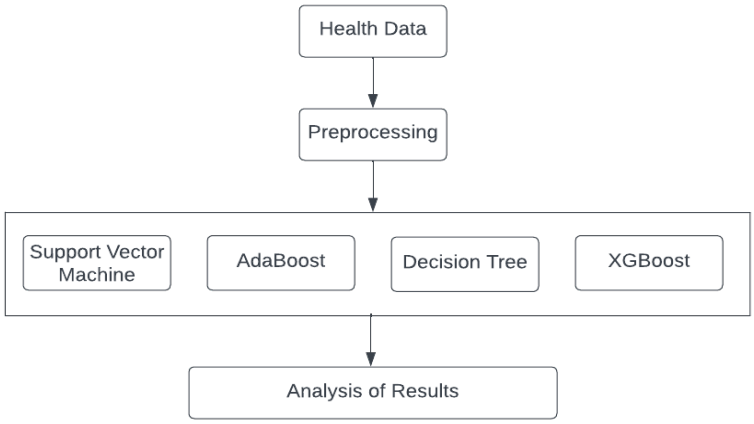

Figure 1: Work flow of analysis

Dataset	rows	columns
cardiovascular-risk-data	3390	17
CVD Prediction Dataset	918	12
cardio-train	70000	13

Table 1. Dataset Dimensions

3.2 Machine learning Models

Support Vector Machines (SVM) Support Vector Machines (SVM)[6] is a powerful supervised learning approach that can be used for both classification and regression problems. SVM, which are well-known for their adaptability and efficacy in dealing with both linear and non-linear interactions, strive to find the optimal hyperplane for distinguishing separate classes within a dataset. This study uses SVM because of its ability to achieve high accuracy while preventing false positives and negatives. This model excels in high-dimensional environments and is particularly useful when dealing with difficult decision boundaries.

AdaBoost AdaBoost[7], also called Adaptive boosting is a supervised machine learning algorithm that combines multiple weak learners such as logistic regression and Decision tree, into one strong learner to classify data. It is an ensemble machine learning algorithm that is usedfor various regression and classification tasks. It repeatedly focuses on misclassified instancesand gives weights to each, progressively improving the model's performance. AdaBoost's iterative adjustment of misclassified instance weights and combination of weak learners lead to high accuracy, crucial in medical diagnosis, particularly for early detection where subtle differences between healthy and diseased individuals are challenging to discern.

Decision Tree Decision Tree is a versatile machine-learning technique. They are built to mimic complex decision-making. They achieve this by continuously categorizing data, on feature values. The algorithm decides on the division of internal nodes by using methods such as information gain or Gini inequality. The adaptation process is repeated until it has reached a stopping criteria. Decision tree is used in this study due to its ability to provide a transparent representationof the decision-making process which plays a vital role in gaining trust in the medical field.The key feature of ranking importance helps in recognizing significant risk factors. DecisionTree strives to capture non-linear relationships of many risk factors making it preferable for detecting cardiovascular disease prediction.

XGBoost Extreme Gradient Boosting (XGBoost)[8] is a supervised machine learning model that achieves high accuracy by combining various weak decision trees. It can be highly efficient in cardiovascular disease prediction due to its exceptional performance in handling large and complex healthcare datasets. XGBoost combines predictions of weak decision trees to capture complex relationships in the cardiovascular dataset. The presence of regularization techniques as an inbuilt feature prevents over-fitting. XGBoost is not only capable of handling missing data to attain accurate prediction but also has the ability of important analysis to help in recognizing major factors leading to cardiovascular diseases[9-14].

4 Experimental Results

In this Section, training outcomes are summarized using important evaluation metrics along-side bar chart providing a visual representation of distinct models performance on the datasets. Figure2 depicts the performance of SVM, XGBoost, AdaBoost, and Decision Tree models. This

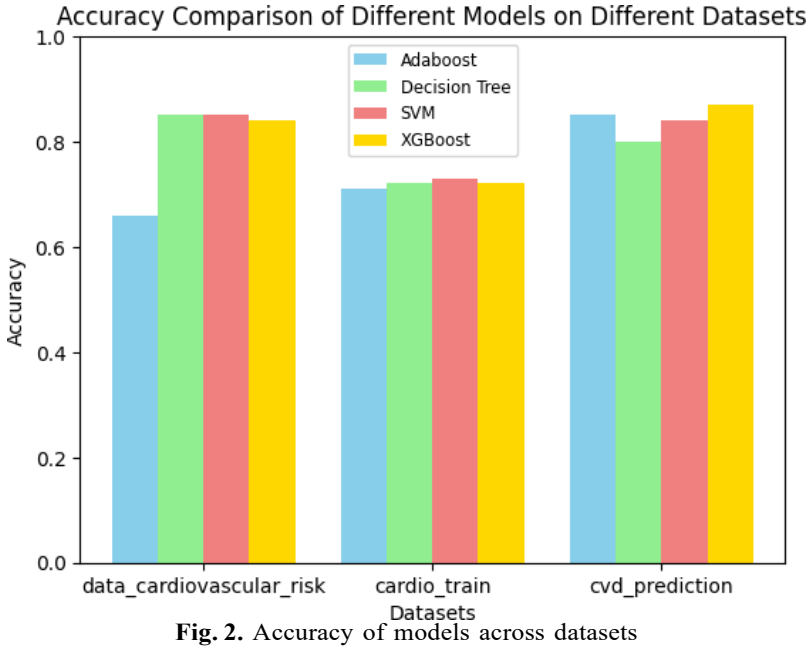

Fig. 2. Accuracy of models across datasets

graphical representation of accuracy demonstrates a clear and concise understanding of each model's performance across the datasets. From the figure2 it is evident that XGBoost consistently performed in all the datasets. while AdaBoost also shows competitive results. SVM and Decision Tree models exhibit a similar accuracy measure for one dataset. It is notable from the representation that there are variations in model performance across the datasets, for instance, XGBoost exhibits high accuracy on one dataset but this is not the case when other datasets were considered SVM and AdaBoost outperformed XGBoost, indicating that different models excel across various datasets. An extensive analysis is performed to analyze the factors contributing to these variations of performance among the datasets. Additionally, our analysis will delve intothe reasons behind the superior performance of a particular model on a specific dataset. Laterin this section, a thorough analysis of model's performance grouped by datasets is presented.

4.1 Cardiovascular risk Dataset

model	accuracy
AdaBoost	0.66
Decision Tree	0.85
SVM	0.85
XGBoost	0.84

Table 2. Performance of models on Cardiovascular risk dataset

Cardiovascular Risk Dataset is an imbalanced dataset. From the table2 it is evident that Decision Tree and Support Vector Machine exhibit similar accuracy, outperforming the other models on this dataset. while XGBoost performs competitively well. AdaBoost exhibited the least accuracy. One of the reasons for the excellent performance of Decision tree could be its ability

adjust to imbalanced data. Most of the disease prediction models comes under imbalanced classification models higher accuracy alone can't determine the performance, Additional analysis is required to determine its efficiency in predicting the CVD.

	predicted 0	predicted 1
actual 0	580	1
actual 1	96	1

Table 3. Confusion Matrix SVM

Table 3 provides details about the prediction positive and negative class. A high true negative value(580) is observed, indicating the model's precision in predicting the negative class correctly. Despite the model's difficulty in capturing the minority positive class, SVM predictedmajority class accurately. The imbalanced distribution of the classes, with a much larger number of scenarios in the negative class, could result in a high accuracy score. In the context of a dis-ease prediction model predicting the positive class is crucial, SVM fails to predict the positiveclass correctly. The SVM model achieved 0.85 accuracy, which was largely due to its abilityto accurately predict the majority class (TenYeaCHD=0). Out of 97 CVD data points SVM predicts only 1 data point as CVD. This highlights SVM unsuitability for early detection of Cardiovascular diseases.

From the Table4,while the true positive count (17) represents correct predictions for TenYear CHD patients. The relatively low false negative count (80) represents cases in which the model

	predicted 0	predicted 1
actual 0	558	23
actual 1	80	17

Table 4. Confusion Matrix Decision Tree

incorrectly predicted no TenYearCHD when it was present, and the false positive count (23) represents scenarios in which the model incorrectly predicted TenYearCHD. One of the reasons for Decision tree model to exhibit higher accuracy could be High true negative value. Accuracy is calculated as the ratio of correct predictions to the total number of instances, With a high true negative count, the accuracy is naturally elevated. A higher accuracy doesn't mean the model is suitable for Cardiovascular prediction tasks.

	predicted 0	predicted 1
actual 0	557	24
actual 1	82	15

Table 5. Confusion Matrix XGBoost

From the confusion matrix presented in Table 5 . XGBoost performed relatively better than Decision Tree and SVM in accurately predicting the positive class with true positive rate(15). A high true positive rate 557 suggests the SVM's ability to correctly classify negative class. The accuracy of model depends on the distribution of the target variable class (0 and 1). In this scenario, the dataset is highly imbalanced with negative class(0) overtaking positive class(1) bya large margin. Hence the accuracy is elevated.

	predicted 0	predicted 1
actual 0	388	193
actual 1	38	59

Table 6. Confusion Matrix AdaBoost

Analyzing the confusion matrix in the table6, AdaBoost had a comparatively large true positive value, suggesting the model's ability to predict the positive class (TenYearCHD = 1) accurately. A comparatively less true negative value represents the model's capability to predict negative class. The false positive count (FP = 193) represents the instances when the model predicted TenYearCHD when it did not exist. This represents a significant drawback in model performance. Although AdaBoost performed well in predicting positive class, the significant reduction in the count of true negatives and elevation in the count of false prediction comprised the overall accuracy of the model.

Among the analyzed models from the figure3 and table7, AdaBoost outperformed others witha more balanced performance across the positive and negative classes. SVM excels at avoiding false positives but has a very low TPR. SVM, Decision Tree, and XGBoost failed to predictthe true positive values making them unsuitable for imbalanced datasets. While XGBoost has a higher TPR than SVM, it still has difficulty correctly identifying positive instances. Consideringthis imbalanced dataset AdaBoost stands out as the best-performing model with high True prediction rates and low false prediction rates.

Model	TPR	TNR	FPR	FNR
AdaBoost	0.608	0.668	0.332	0.392
SVM	0.010	0.998	0.002	0.990
XGBoost	0.155	0.959	0.041	0.845
Decision Tree	0.175	0.960	0.040	0.825

Table 7. True positive rates, False positive rates, true negative rates, false negative rates

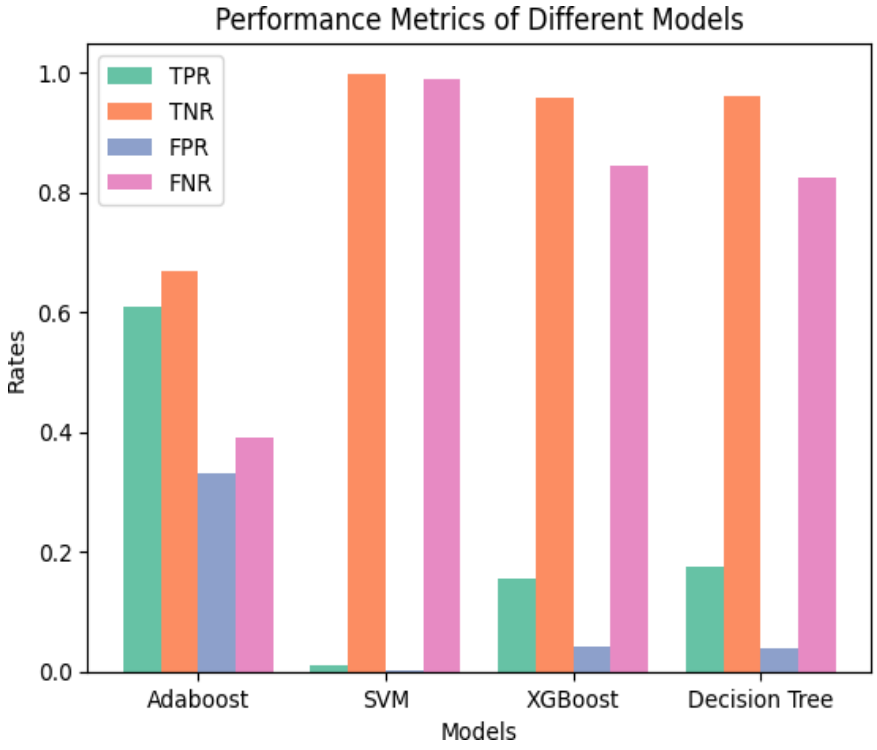

Figure 3: Performance Metrics of Models on Cardiovascular risk dataset

4.2 Cardiovascular train Dataset

model	accuracy
AdaBoost	0.70
Desicion Tree	0.72
SVM	0.73
XGBoost	0.72

Table 8. Performance of models on Cardiovascular train dataset

Data presented in the table8 clearly indicates that SVM has secured the highest accuracy, This superior performance is because of SVM's ability to find the best-fitting decision boundaries which enables enhanced classification accuracy. While Decision Tree and XGBoost have shown comparable accuracy. AdaBoost achieved slightly lower accuracy.

Table 9 is the confusion matrix of SVM, the model accurately predicted the absence of car-cardiovascular disease in 5231 (True Negative) cases and correctly recognized the presence in 4926

	predicted 0	predicted 1
actual 0	5321	1667
actual 1	2086	4926

Table 9. Confusion Matrix SVM

(True positive) cases. Although it also incorrectly predicted the absence when the disease was present in 1667 (False Negative) cases. From the table, we can clearly understand SVM's robust performance in classifying the presence or absence of cardiovascular disease. SVM successfully classified 5231 true negatives and 4926 true positives which reflects the model's accurate identification of both disease and non-disease cases. SVM's ability to handle the non-linear relationship between features and outcomes such 'as age',' height', and 'weight' might influence cardiovascular health. This shows its ability to generalize and also demonstrates the capability to navigate non separable data. Equipped with kernel trick which enables flexible decision boundaries by transforming data in higher dimensional spaces. While AdaBoost can handle non-linearity it isnot as efficient as SVM in cases of complex relationships between features, If they are featured with high variability AdaBoost may struggle to effectively correct errors subsequent iterations. Decision trees can easily over-fit and might not be able to detect complex interactions between features it may also struggle with the continuous nature of features such as 'age' and 'weight' which may again lead to complex decision boundaries leading to overfitting. XGBoost's performance is sensitive to features with weak predictive power as weak features may not be able to contribute towards the boosting process

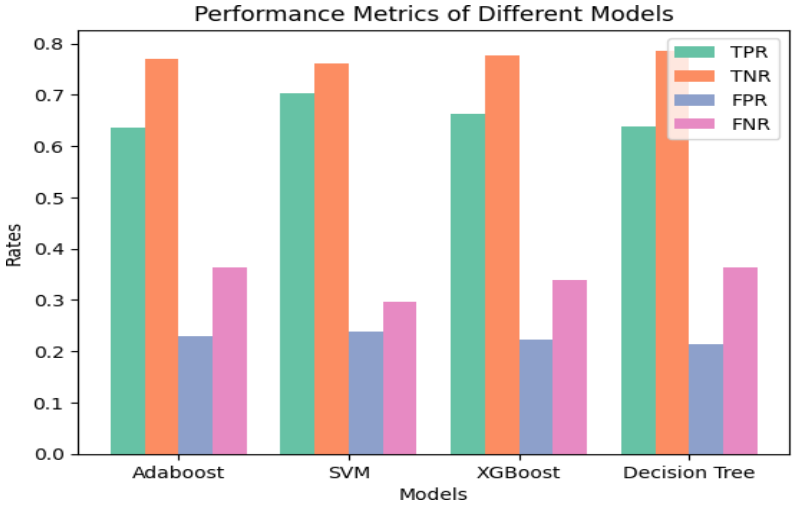

Figure. 4. Performance metrics of models on cardiovascular train dataset

From the figure4 it is clear that the SVM model has comparatively high true prediction and low false prediction values. while other models also have competitive prediction rates SVM model outperforms, making it suitable for Cardiovascular prediction tasks. Our main motto is to Predict CVD'S early, it is important for the model to have high true positive rate and less false negative rate. AdaBoost, XGBoost has similar true positive rate to SVM but false negative rate of AdaBoost and XGBoost is slightly higher which made AdaBoost, XGBoost unsuitable for this dataset.

4.3 CVD Prediction Dataset

Based on the data presented in Table10, XGBoost has the highest accuracy. It is an optimized distributed gradient boosting library, efficient in handling various data types and known for its high performance and speed.

model	accuracy
AdaBoost	0.85
Desicion Tree	0.80
SVM	0.84
XGBoost	0.87

Table 10. Performance of models on CVD Predition dataset

XGBoost uses ensemble learning approach which utilizes various weak learners to create a more optimum model. Although AdaBoost and SVM accuracies are comparable while Decision tree has shown the least accuracy among the models

	predicted 0	predicted 1
actual 0	67	10
actual 1	13	94

Table 11. Confusion Matrix XGBoost

In the table11, it is evident that XGBoost displays superior performances in cardiovascular disease detection as shown by its confusion matrix. The confusion matrix of XGBoost demonstrates a well-balanced classification with 67 true negatives 94 true positives and 67 true negatives. The model has attained high accuracy of 0.88 indicating its efficiency in correctly classifying data among both classes. XGBoost's ensemble nature combines the strengths of various decision trees to produce an optimum prediction model. The boosted decision tree used by XGBoost allows it to detect complex relationships in the data to adapt to non-linear patterns and handle feature interactions effectively. XGBoost success can be attributed to its sophisticated handling of the dataset's characteristics. It efficiently processed mixed data types such as 'Sex' (categorical) and 'FastingBS'(categorical) alongside 'Cholesterol' (continuous) seamlessly. The model's gradient boosting mechanism, adept at reducing errors iteratively, was particularly effective in navigating the non-linear relationships present in medical datasets, such as the interaction between lifestyle factors (e.g., 'ExerciseAngina') and physiological measurements (e.g., 'MaxHR').

While SVM is effective in high-dimensional spaces the model's performance on this dataset was likely limited by the challenge of choosing an appropriate kernel to detect complex interplay

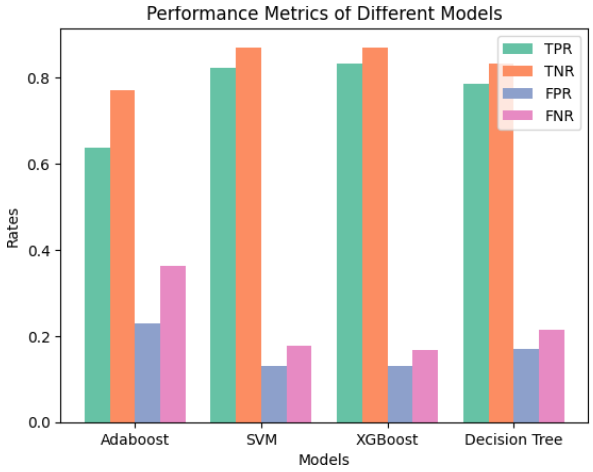

Figure. 5: Performance metrics of models on CVD prediction dataset

of features. Decision Tree's tendency to overfit was due to the dataset's diverse features which would create overly complex branches reducing its ability generalize test data. AdaBoost is sensitive to noisy variables in the dataset such as cholesterol which can vary significantly due todiet and might be handled less effectively in AdaBoost.

Referring the figure5 Decision tree also exhibits comparable accuracy with almost similar true positive rate. Although SVM exhibits almost similar True positive rates and false negative rates, XGBoost has slightly higher values of sensitivity, making it the best choice for this dataset

5 Conclusion

In conclusion, this extensive research into early prediction models for various cardiovascular datasets has yielded valuable insights. The performance of AdaBoost, SVM, XGBoost, and Decision Tree models underwent extensive evaluation in a variety of scenarios, including minimal data challenges, imbalances, and larger datasets. AdaBoost excels at handling imbalanced datasets, making it the best choice in scenarios where classes are unevenly distributed. SVM demonstrated high accuracy while effectively reducing false positives and negatives. Decision Tree proved to be dependable, with a well-balanced trade-off between sensitivity and specificity. XGBoost demonstrated exceptional balance with high true positive and true negative rates, establishing it as a top-performing model. Overall conclusion emphasizes the significant relationship between model performance and dataset characteristics. Choosing the best model for the dataset's specific attributes is critical for achieving the best results.

References

1. Shorewala, Vardhan. (2021). Early detection of coronary heart disease using ensemble techniques. Informatics in Medicine Unlocked. 26. 100655. 10.1016/j.imu.2021.100655.
2. Ibarra, Rodrigo & Leon, Jaime & Ávila, Iván & Ponce, Hiram. (2022). Cardiovascular Disease
Detection Using Machine Learning. Computación y Sistemas. 26. 10.13053/cys-26-4-4422.

3. Subasish Mohapatra, Sushree Maneesha, Prashanta Kumar Patra, Suhadarshini Mohanty, HeartDiseases Prediction based on Stacking Classifiers Model, Procedia Computer Science, Volume 218,2023, Pages 1621-1630.
4. Chen, Joy & Hengjinda, Pisith. (2021). Early Prediction of Coronary Artery Disease (CAD) by Machine Learning Method - A Comparative Study. Journal of Artificial Intelligence and Capsule Networks. 3. 17-33. 10.36548/jaicn.2021.1.002.
5. Sun, Weicheng & Zhang, Ping & Wang, Zilin & Li, Dongxu. (2021). Prediction of Cardiovascular Diseases based on Machine Learning. ASP Transactions on Internet of Things. 1. 30-35. 10.52810/TIOT.2021.100035.
6. Evgeniou, Theodoros & Pontil, Massimiliano. (2001). Support Vector Machines: Theory and Appli-cations. 2049. 249-257. 10.1007/3-540-44673-7 12.
7. Chengsheng, Tu & Huacheng, Liu & Bing, Xu. (2017). AdaBoost typical Algorithm and its application research. MATEC Web of Conferences. 139. 00222. 10.1051/matecconf/201713900222.
8. Chen, T., & Guestrin, C. (2016, August). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. https://doi.org/10.1145/2939672.2939785
9. A. Ravindran, G. G. Krishna, Sagara and S. S., "A Comparative Analysis of Machine Learning Algorithms in Detecting Deceptive Behaviour in Humans using Thermal Images," 2019 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 2019, pp. 0310-0314,doi: 10.1109/ICCSP.2019.8697911.
10. R. S. Pillai and D. L. R, "A Survey on Citation Recommendation System," 2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT), Kannur, India, 2022, pp. 423-429, doi: 10.1109/ICICICT54557.2022.9917887.
11. Gayathri, R.G. and Nair, J.J., 2015. Towards efficient analysis of massive networks. International Journal of Applied Engineering Research, 10(69), pp.222-227.
12. Chaitanya, S. Sarath, Malavika, Prasanna and Karthik, "Human Emotions Recognition from Thermal Images using Yolo Algorithm," 2020 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 2020, pp. 1139-1142, doi: 10.1109/ICCSP48568.2020.9182148.
13. Y. K. M, S. C K, T. A R, S. B. Kumar and G. Sarath, "A Twitter-based Software Vulnerability Alert Framework using Natural Language Processing," 2023 8th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2023, pp. 702-707, doi:10.1109/ICCES57224.2023.10192794.
14. Jasmine Bhaskar, Sruthi, K., and Prof. Prema Nedungadi, "Enhanced sentiment analysis of in- formal textual communication in social media by considering objective words and intensifiers", in IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE), 2014, Jaipur, 2014