HEART DISEASE PREDICTION USING MACHINE LEARNING TECHNIQUES

K. Sravanthi[1], K. Rajasekhar[2]

[1]Student, ECE Department, DR.YSR ANU College of Engineering and Technology, Andhra Pradesh

[2]Assistant Professor, ECE Department, DR.YSR ANU College of Engineering and Technology, Andhra Pradesh

---

ABSTRACT - *cardiovascular diseases encompass a variety of heart-related ailments and remain the leading cause of global mortality. Early prediction and consistent treatment are crucial for minimizing potential harm. This study aims to compare various machine learning techniques to achieve optimal results. In addition to conventional models, novel approaches such as TabNet and Restricted Boltzmann Machine (RBM) will be explored to determine their superior predictive capabilities. AdaBoost classifier showcases better performance with 91% accuracy and HistgradientBoostClassifier with 89% accuracy among all other models included in this work. When AUC-ROC of both models is compared the latter is highest with 0.96. Whereas LightGBM and RBM are the least performed models with an average accuracy rate of 83%. While other competing models such as TabNet and XGBOOST are closer in their performances with average accuracy of 85% but AUC-ROC of these approaches is less. Logistic Regression produced an accuracy of 87.4% with good precision value.*

Key words: **AdaBoostClassifier, HistgradientBoostClassifier, RandomForestClassifier, Logistic Regression, LightGBM, XGBOOST, TabNet, RBM**

1. INTRODUCTION

The heart plays a significant role in the human body. It coordinates with all other organs by pumping blood and maintaining blood pressure. However, due to a lack of proper exercise and diet maintenance, heart problems have become more common among younger individuals. Every year, approximately 370,000 people die due to heart disease [21]. It is crucial to predict heart disease as it involves several risk factors, such as diabetes, high blood pressure, high cholesterol levels, abnormal pulse rate, and other related conditions [1]. Therefore, the implementation of machine learning techniques becomes essential for more accurate predictions and early detection [2] The main objective of this project is to predict whether a person is at risk of getting heart disease in the next 10 years. Through the comparison of conventional models and the exploration of novel approaches such as TabNet and RBM [3], [2] we will implement these methods to enhance the accuracy of our predictions. The computation of metrics such as AUC-ROC, Accuracy, Recall, F1-SCORE are used for the performance evaluation of each model. This helps early detection and prevention of heart disease.

Literature Review:

Aravind Sasidaran Pillai [1] proposed a tabular neural network model for cardiac disease prediction and compared it with base models such as random forest, XGBoost, logistic regression, and gradient boost using the UCI heart disease dataset. The successful outcomes of the study were validated using evaluation metrics like ROC curves, accuracy, precision, sensitivity, specificity, and confusion matrices. The TabNet model performed exceptionally well, achieving an accuracy of 94.4%.

Himanshu Sharma et al. [2] explored the performance of different machine learning algorithms based on parameters like parametric accuracy. They also provided descriptions of deep learning algorithms such as stacked RBM, autoencoder-decoder techniques, recurrent neural networks (RNNs) like LSTM and GRU, highlighting their remarkable performance in sequence-based tasks.

Sercan O et al. [3] demonstrated that the TabNet model, implemented on real-world datasets, outperformed other models with an impressive accuracy of 96%. They also proposed tabular self-supervised learning in TabNet, considering unsupervised learning scenarios using decoder architecture to reconstruct tabular features from encoded representations.

Subhashish Mohapatra et al. [4] proposed various machine learning approaches for the early prediction of heart disease. They utilized GridSearchCV to select the best hyper parameters through cross-validation. Random Forest emerged as the best-performing model, achieving an accuracy of 86.89% before hyperparameter tuning. After tuning, k-nearest neighbors yielded an accuracy of 88.52%.

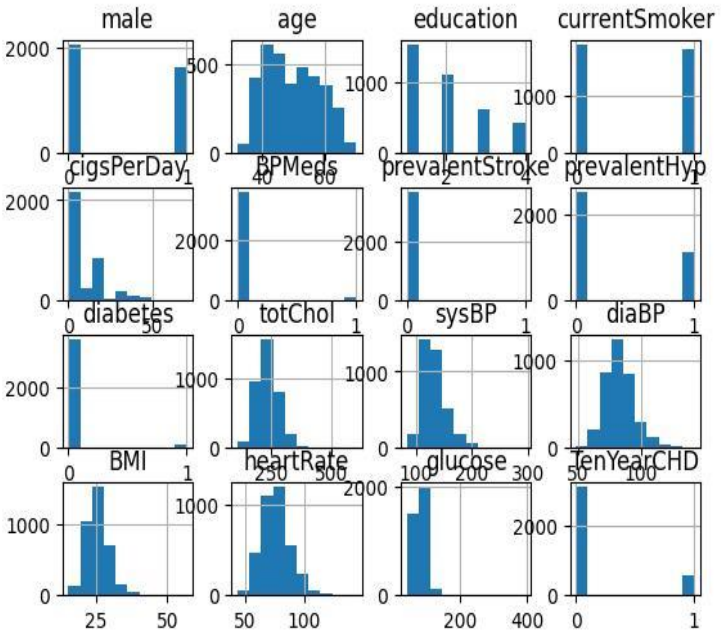
Vardhan Shorewala et al. [6] emphasized the importance of data preprocessing, feature analysis, and modeling techniques in heart disease classification. They discussed various techniques, including traditional classifiers, neural networks, and ensemble methods, for accurate predictions. The dense neural network showed the best performance with 73.93%, while the stacking evaluation of bagged decision trees yielded 74.8%.

Saroj Kumar Pandey et al. [7] utilized the MIT-BIH AD database to classify ECG signals. They implemented a stacked RBM model, which demonstrated excellent performance in the classification of arrhythmias. The model's parameters, including weights and biases, were adjusted using the contrastive divergence learning algorithm.

G.B.C Latha et al. [8] proposed different models for improving prediction. The ensemble model is one among the specified models that improves the accuracy of weak learners by up to 7%. Furthermore, feature selection was observed to improve accuracy by up to 4.63% for the stack of Naïve Bayes, Bayes Net, C4.5, PART, and MLP with Random Forest.

2. MATERIALS AND MODELS

This study utilized the Framingham Heart Study dataset, which initially consists of 4240 records with 16 variables. After dropping missing values, a modified dataset included a total of 3658 records for further analysis. The dataset was split into two subsets, with 20% of the data reserved for testing and the remaining 80% for training purposes.



Name	Description
age	Age (in yrs)
education	Education (less than high school=0, high school=1)
CurrentSmoker	Whether the person is smoking currently
CigsperDay	Number of cigarettes a person smoke per day
BPMeds	Whethr a patient is taking blood pressure medicine or not
PrevalentStroke	Indicates any previous stroke experienced by patient
PrevalentHyp	Indicates the previous record of hypertension
Diabetes	Whether patient is having diabetes or not
totChol	Indicates the total cholesterol level of patient
sysBP	Systolic blood pressure (90 - 120 mm Hg)
diaBP	Diastolic blood pressure (60 – 80 mm Hg)
BMI	Body mass index value
heartRate	It is discrete value, but taken continuous range of values as there are many possibilities
glucose	Blood glucose level
TenYearCHD	Whether the person is in chance of getting heart disease in coming 10 years

Table: Tabular description of Data

3.CONVENTIONAL MODELS:

Random Forest: Random Forest is a heterogeneous ensemble learning method that combines multiple decision trees to achieve accurate predictions.

Each decision tree within the ensemble is trained individually using different subsets of training data and random subsets of features.

HistGradientBoosting: It is a gradient boosting algorithm that takes advantage of utilizing histograms in the training process. It efficiently determines the best split points during tree construction by creating feature histograms. This makes HistGradientBoosting more suitable for handling large datasets with numerical features.

XGBOOST: It shows exceptional performance by utilizing advanced techniques like boosting and regularization. XGBoost is a highly optimized technique with impressive speed and scalability, making it compatible with complex structured data. These features make XGBoost a preferable choice in various machine learning techniques.

Logistic Regression: It is a statistical model employed for binary classification tasks, aiming to predict the probability of an instance belonging to a specific class. It establishes the connection between the input variables and the output class in the logistic function. Logistic regression is widely embraced due to its simplicity, interpretability, and capability to handle high-dimensional data. It is frequently utilized as a foundational model for classification tasks, serving as an initial step in comprehending the influence of features on the predicted outcome.

AdaBoost (Adaptive Boosting): AdaBoost is an ensemble learning method that combines multiple weak classifiers to create a strong classifier. It iteratively trains weak classifiers on different subsets of the training data, assigning higher weights to instances that are misclassified. The final prediction is based on the weighted combination of the weak classifiers. AdaBoost is effective in situations where weak classifiers can be combined to achieve higher accuracy.

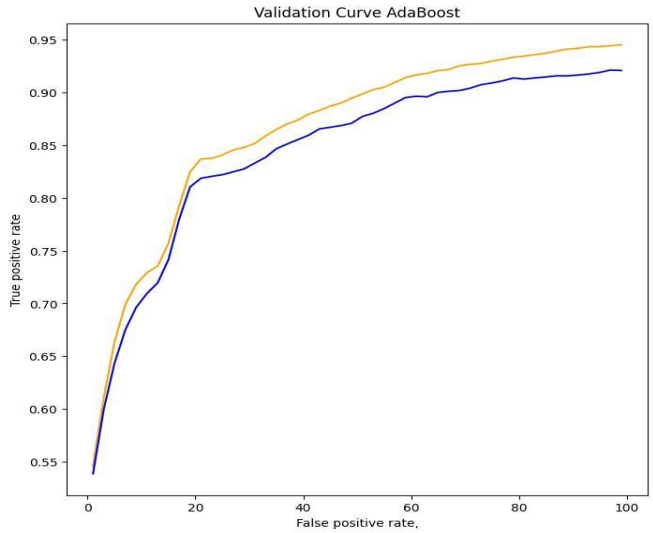


Fig 3.1 AdaBoostClassifier Validation Curve

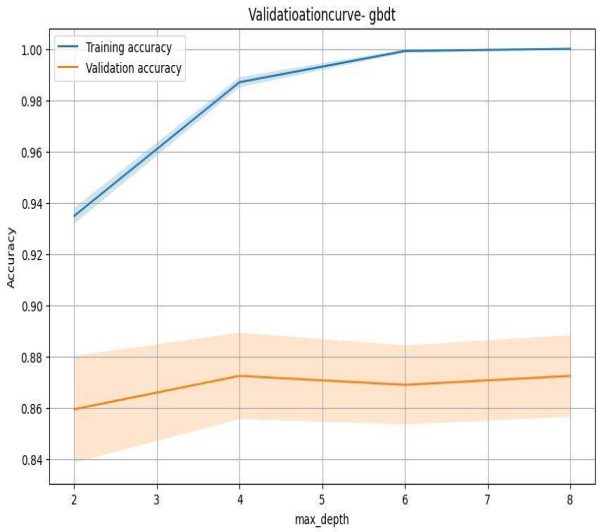


Fig:3.2 HistGradientBoostingClassifier validation Curve

LightGBM: LightGBM belongs to the gradient boosting family and stands out due to its unique learning approach, employing leaf-wise tree growth. This approach creates deeper trees with fewer nodes and leaves compared to traditional gradient boosting algorithms.

During the training process, LightGBM incorporates an optimization technique called Gradient-based One-Side Sampling. This technique selectively samples instances with larger gradients while reducing the number of instances considered for training. Remarkably, this sampling strategy maintains promising accuracy without compromising performance.

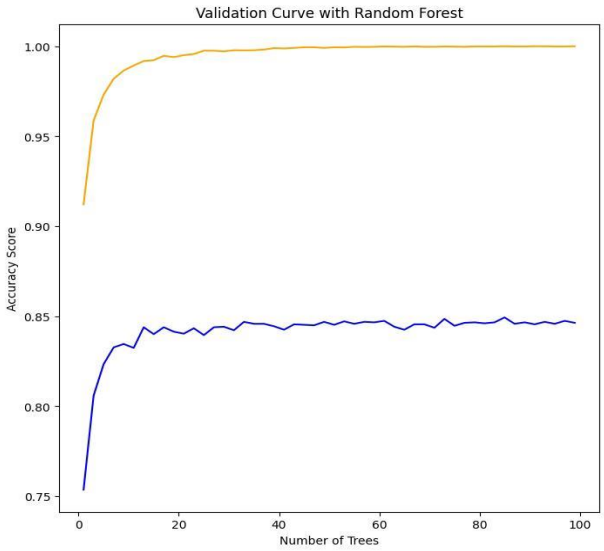


Fig 3.3 Random Forest Validation Curve

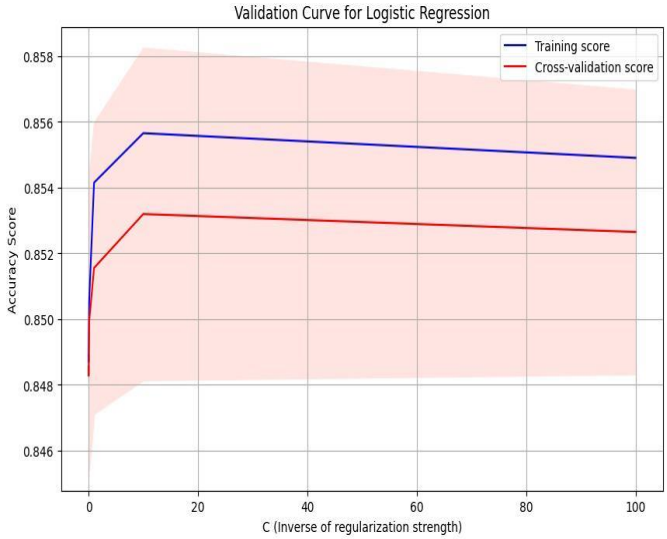


Fig 3.4 Logistic Regression Validation Curve

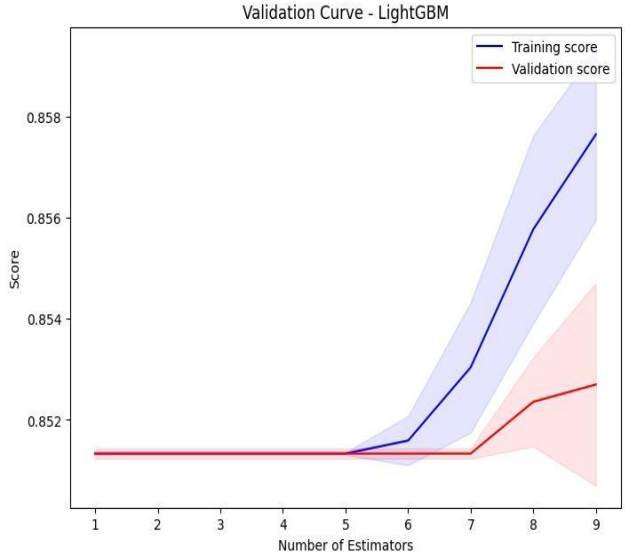


Fig 3.5 light GBM validation curve

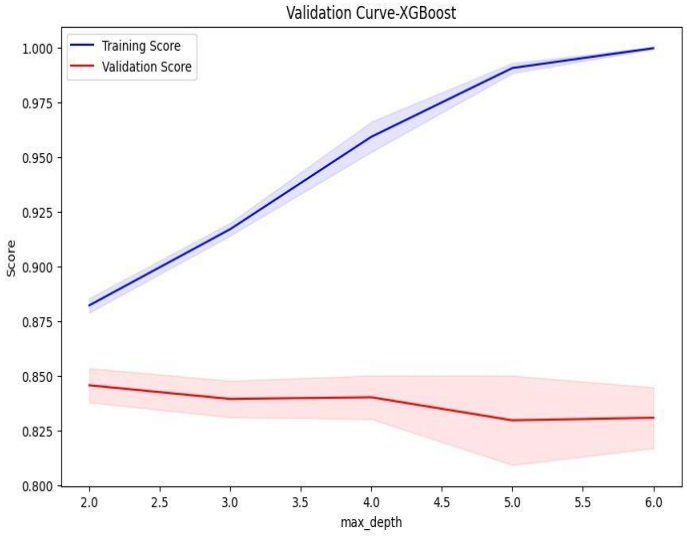


Fig 3.6 XGBOOST Validation Curve

Evaluation Metrics:

Accuracy measures the overall correctness of the model's predictions, while precision focuses on the proportion of true positive predictions among all positive predictions. Recall evaluates the model's ability to identify positive instances correctly, and the F1 score combines precision and recall into a single metric. AUC-ROC assesses the model's ability to distinguish between positive and negative instances across various probability thresholds.

3.1 TabNet

TabNet is an innovative deep neural network designed specifically for structured and tabular data. It offers a unique approach by combining traditional deep learning techniques with soft function selection, mimicking the behavior of decision trees. This enables TabNet to effectively handle tabular datasets and achieve interpretability. A key advantage of TabNet is its ability to perform feature selection and inference within a single architecture, making it efficient and interpretable. The model utilizes sequential attention to determine the most relevant features at each decision step, allowing it to focus on the most salient information for better learning. One notable feature of TabNet is its capability to directly process raw tabular data without requiring extensive preprocessing. This flexibility facilitates seamless integration into end-to-end learning pipelines and reduces the need for manual data transformations. TabNet offers two levels of interpretability. Local interpretability enables visualizing the importance of individual features and their combinations, aiding in understanding the model's decision-making process. Global interpretability quantifies the contribution of each feature to the overall model, providing insights into feature impact. In terms of model training, TabNet employs techniques such as early stopping to prevent overfitting. It splits the training data into validation and test sets and utilizes metrics like ROC and accuracy to determine the optimal stopping point. Additionally, TabNet supports the use of categorical features through supplied feature indices. These features and capabilities make TabNet a promising option for handling structured and tabular data, offering both efficiency and interpretability. Its unique approach to feature selection and its ability to process raw data make it well-suited for various applications in the field of machine learning and data analysis.

Model Training for TabNet

During the model training process, we randomly divided the training data into three sets: 24% for validation, 56% for testing, and the remaining portion for actual training. The base models were trained using default hyper parameters. Specifically, the TabNet model is trained for a maximum of 1000 epochs, but early stopping occurred at 120 epochs based on the performance metrics of ROC and accuracy. The PyTorch optimizer function, Adam, was used with an initial learning rate of 0.01. The 'sparsemax' masking function was employed for feature selection.

Fig 3.1 .1 TabNet training loss

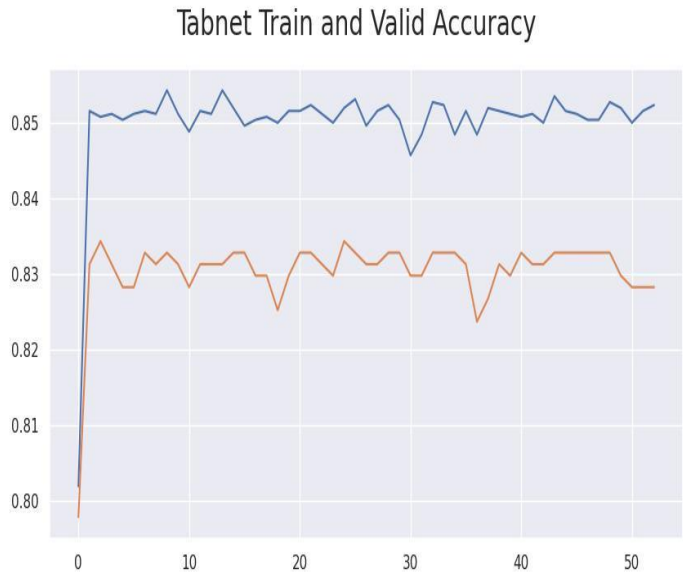


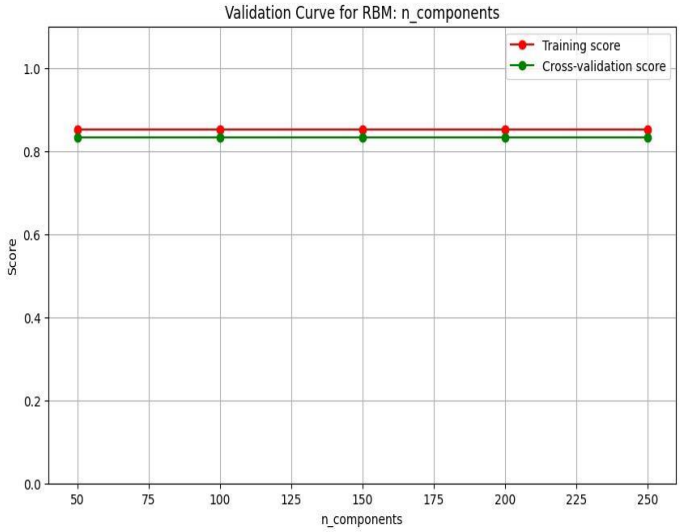
Fig 3.1.2 TabNet validation Curve

3.2 Restricted Boltzmann Machine (RBM):

An RBM is a neural network model utilized for unsupervised learning. It comprises two layers: the visible layer and the hidden layer. These layers are connected through weights, but there are no direct connections between the individual nodes within each layer. RBM learns to model the probability distribution of the input data by adjusting the weights and biases to minimize the reconstruction error. RBMs are particularly useful for tasks like dimensionality reduction, feature learning, and collaborative filtering. They are powerful tools for extracting meaningful representations from complex datasets. A stacked RBM model, also known as a deep belief network (DBN), is formed by arranging multiple RBMs in a hierarchical structure. This involves training RBMs one layer at a time and stacking them together to create a deep neural network. Each RBM in the stack acts as a hidden layer for the subsequent RBM. This arrangement enables the model to learn hierarchical features, capturing increasingly complex patterns and representations in the data.

Model Training For RBM

The data is preprocessed using standard scalar which computes mean and standard deviations. This standardized data will reduce mean which is then utilized for both train and test data which will create equal contribution of every feature. A pipeline is created to combine RBM and Logistic Regression models which RBM extracts features, and Logistic Regression further classifies. The model then trained with initial learning rate of 0.1 and up to 10 iterations are created to learn



weights and biased of hidden layer and visible layer through learning process.

Fig 3.2 Validation Curve for RBM

4.RESULTS

AdaBoost achieved the highest accuracy with models, while HistGradientBoost exhibited the best performance in distinguishing between the classes, as indicated by its AUC-RUC of 0.96. 91.2% among the conventional.

Model	auroc	Accuracy	Precision	Recall
AdaBoost	0.91	91.2	0.91	0.91
Histgradientboost	0.96	89.6	0.88	0.91
Logistic Regression	0.53	87.4	0.88	0.08
Random Forest	0.68	86.8	0.58	0.16
XGBOOST	0.56	85.2	0.39	0.17
LightGBM	0.53	83.3	0.50	0.09

Table 4.1- Models comparison by AUCROC accuracy, precision, recall

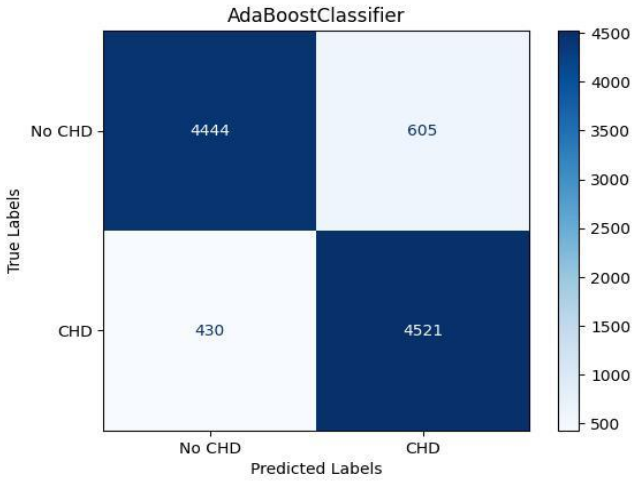


Fig 4.1Confusion Matrix for AdaBoost

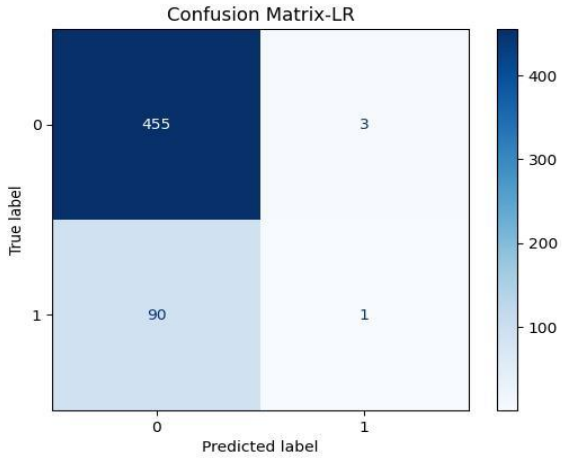


Fig 4.2 Confusion matrix for Logistic Regression

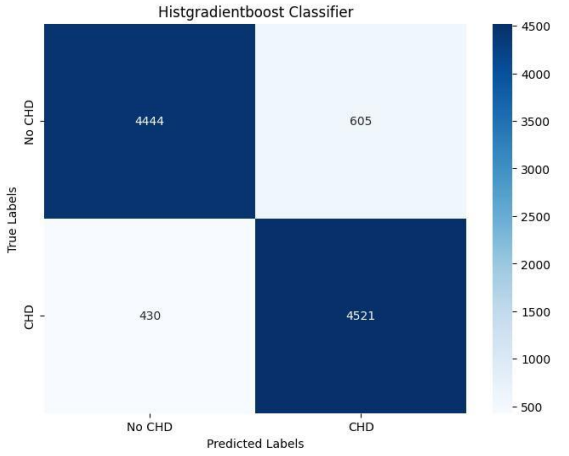


Fig 4.3 Confusion Matrix for HistGradientBoost

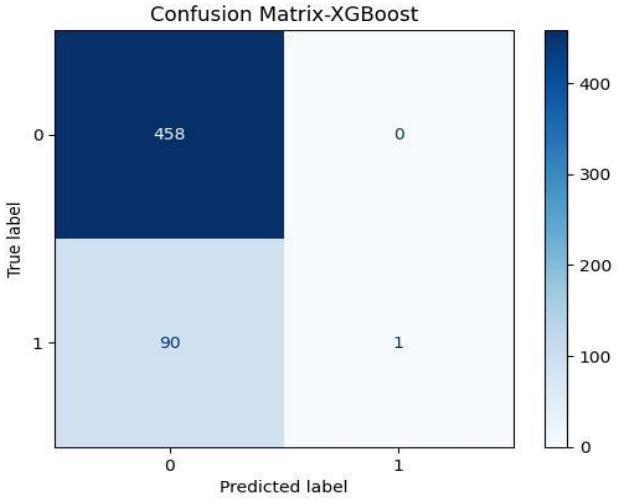


Fig 4.4 Confusion Matrix for XGBoost

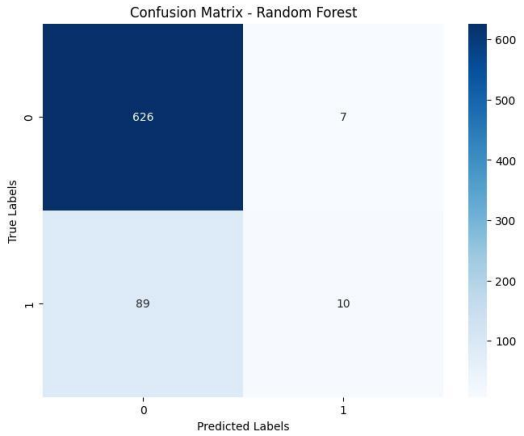


Fig 4.5 Confusion Matrix for Random Forest

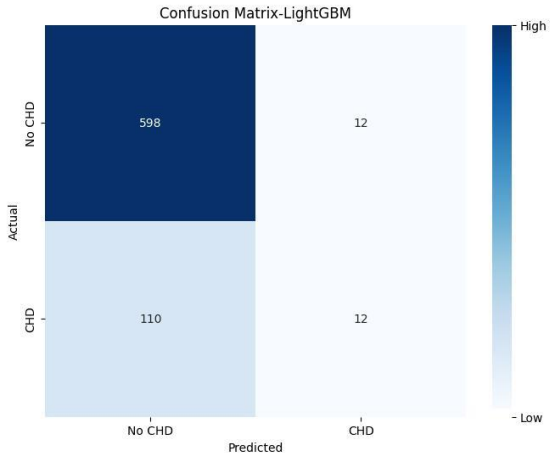
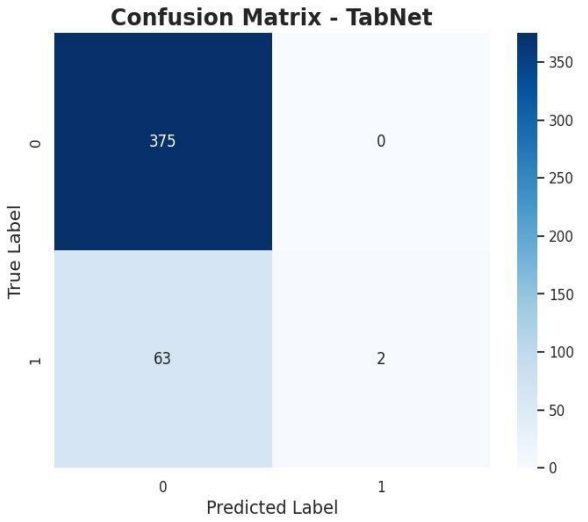


Fig 4.6 Confusion matrix for LightGBM

When compared the performance of models using confusion matrix, RBM correctly identifies the 610 true negative cases, while TabNet identifies 375 cases. However, RBM fails to identify any true positive cases, whereas TabNet successfully identifies 2 true positive cases.



. **Fig** 4.7Confusion Matrix for TabNet.

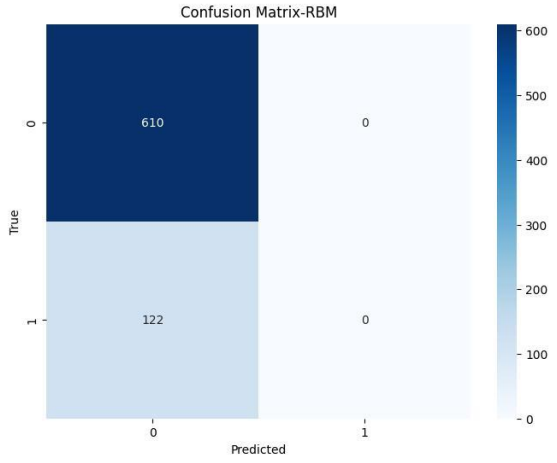


Fig 4.8 confusion matrix for RBM

In the novel approach, TabNet Demonstrates an Accuracy of 83.6%, indicating that it correctly predicts the outcome for 83.6% of the instances. Additionally, TabNet shows better performance in terms of the AUC-ROC value of 0.74, which implies that it has good discrimination power in distinguishing between positive and negative instances by AUC-ROC, accuracy, precision, recall.

Model	auroc	accuracy	Precision	recall
TabNet	0.5 0	84.5	0.2	0.015
RBM	0.50	83.3	0.0	0.0

Table4.2 Novel approaches comparison

5. DISCUSSION

Feature Importance plays a vital role in problem identification. Feature Importance for TabNet and RBM are plotted to identify which features are effectively contributing to decision-making. There is future scope for research to include the most influencing features.

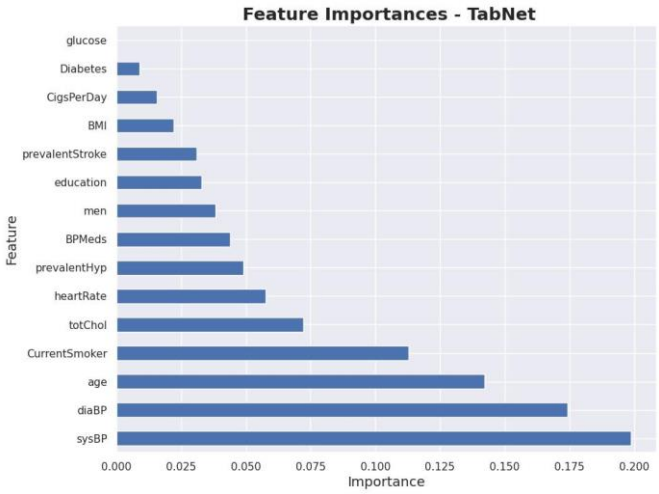


Fig 5.1 TabNet feature importance

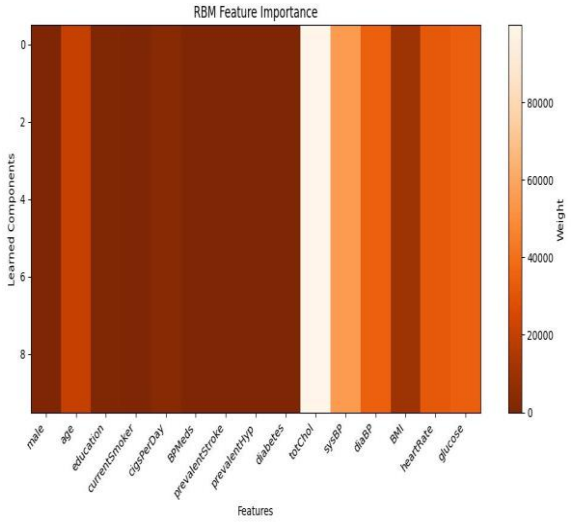
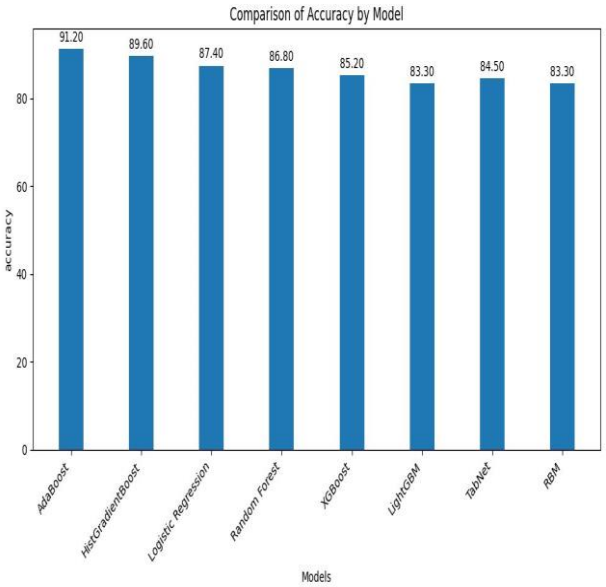
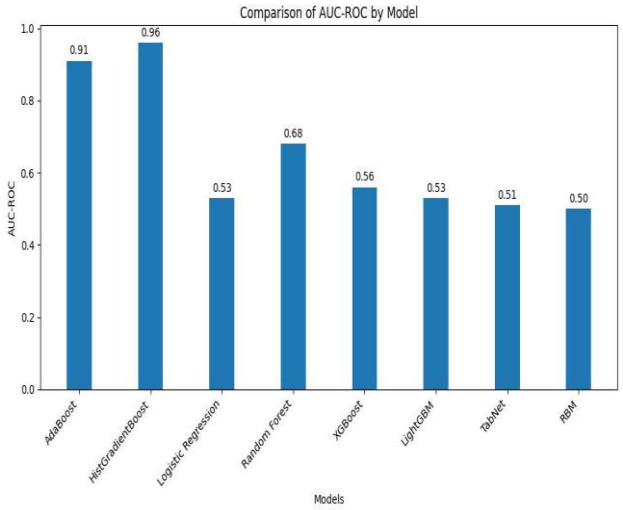


Fig 5.2 RBM feature importance





The bar graphs represented below indicates model outcomes in terms of accuracy and AUC-ROC.

Fig 5.3comparison of models by accuracy

Fig 5.4 comparison of models by AUC-ROC

6.CONCLUSION

It is needed to predict and identify root cause of disease for Saving many lives. Predicting the disease is somewhat typical task as the reason for disease varies from person to person. This work includes the risk factors that are responsible for developing heart disease using Machine Learning algorithms. This is a comporative approach for early prediction. Histgradientboost showcased best performance with AUC-ROC of 0.96 and TabNet performed well achieving an accuracy of 84.5% in novel approaches. There is future scope for deep classification of features using TabNet and improve performance of algorithms.

7. REFERENCES

1.Pillai, Aravind Sasidharan. “Cardiac disease prediction with tabular neural network.” (2022).

2.Sharma, Himanshu, and M. A. Rizvi. “Prediction of heart disease using machine learning algorithms: A survey.” International Journal on Recent and Innovation Trends in Computing and Communication 5.8 (2017): 99-104.

3. Arik, Sercan Ö., and Tomas Pfister. “Tabnet: Attentive interpretable tabular learning.” Proceedings of the AAAI conference on artificial intelligence. Vol. 35. No. 8. 2021.

4. Mohapatra, Subasish, et al. “An Approach for Heart Disease Prediction Using Machine Learning.” Intelligent Systems: Proceedings of ICMIB 2020. Springer Singapore, 2021

5.Shah, Maunish, et al. “heart disease prediction and recommendation.” Heart Disease 6.04 (2019).

6.Shorewala, Vardhan. “Early detection of coronary heart disease using ensemble techniques.” Informatics in Medicine Unlocked 26 (2021): 100655.

7.Pandey, Saroj Kumar, et al. "Automated arrhythmia detection from electrocardiogram signal using stacked restricted Boltzmann machine model." SN Applied Sciences 3.6 (2021): 624.

8.Latha, C. Beulah Christalin, and S. Carolin Jeeva. "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques." Informatics in Medicine Unlocked 16 (2019): 100203

9. Lu, Peng, et al. "Research on improved depth belief network-based prediction of cardiovascular diseases." Journal of healthcare engineering 2018 (2018).

10.Dutta, Aniruddha, et al. "An efficient convolutional neural network for coronary heart disease prediction." Expert Systems with Applications 159 (2020): 113408.

11. Yazdani, Armin, et al. "A novel approach for heart disease prediction using strength scores with significant predictors." BMC medical informatics and decision making 21.1 (2021): 194.

12. Krittanawong, Chayakrit, et al. "Machine learning prediction in cardiovascular diseases: a meta-analysis." Scientific reports 10.1 (2020): 16057.

13. Khazaee, Ali. "Heart beat classification using particle swarm optimization." International Journal of Intelligent Systems and Applications 5.6 (2013): 25.

14. Baccouche, Asma, et al. "Ensemble deep learning models for heart disease classification: A case study from Mexico." Information 11.4 (2020): 207

15. Shaikh, Rumana M. "cardiovascular diseases prediction using machine learning algorithms." Turkish Journal of Computer and Mathematics Education (TURCOMAT) 12.6 (2021): 1083-1088

16. Prakaash, A. S., et al. "Design and development of modified ensemble learning with weighted RBM features for enhanced multi-disease prediction model." New Generation Computing 40.4 (2022): 1241-1279

17. Muhammad, Yar, et al. "Early and accurate detection and diagnosis of heart disease using intelligent computational model." Scientific reports 10.1 (2020): 19747.

18. Gonsalves, Amanda H., et al. "Prediction of coronary heart disease using machine learning: an experimental analysis." Proceedings of the 2019 3rd International Conference on Deep Learning Technologies. 2019.

19. Singh, Navdeep, Punjab Firozpur, and Sonika Jindal. "Heart disease prediction system using hybrid technique of data mining algorithms." International Journal of Advance Research, Ideas and Innovations in Technology 4.2 (2018): 982-987

20. Mohan, Senthilkumar, Chandra Segar Thirumalai, and Gautam Srivastava. "Effective heart disease prediction using hybrid machine learning techniques." IEEE access 7 (2019): 81542-81554.

21. https://towardsdatascience.com/heart-disease-prediction-73468d630cfc