Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 11, November-2022

Cardiac Disease Prediction with Tabular Neural Network

Aravind Sasidharan Pillai
University of Illinois,Urbana-Champaign

***Abstract*:- Cardiac disease, which includes a variety of diseases that affect the heart, is a leading cause of death worldwide. One of every four deaths in the United States is due to heart disease. This means that approximately 610,000 people die from the disease each year. Heart disease is much easier to treat if detected early. Machine learning can play a crucial role in early detection and save lives.**

**This research aims to develop an artificial intelligence-based system that identifies patients
who are more likely to develop heart disease based on their medical history. The heart disease dataset from the UCI Machine Learning Repository was used for training and validation. Traditional classification techniques such as logistic regression, random forest, gradient boosting, and extreme gradient boosting were used as base models, and the results were compared with the Tabnet model. Tabnet is a new robust, interpretable, deep-learning architecture for tabular data. TabNet uses sequential attention to choose which features to conclude from at each decision step, focusing its learning ability on the most salient features, allowing for interpretability and more efficient learning.**

Promising results were obtained and validated using ROC curves, accuracy, precision, sensitivity, specificity, and confusion matrices. The Tabnet deep learning model outperformed the others, achieving 94% accuracy, ROC score of 0.94, and specificity and sensitivity greater than 0.93.

1. INTRODUCTION

The term heart disease refers to different types of heart conditions. The most common type of heart disease is coronary artery disease, which affects blood flow to the heart. Reduced blood flow can lead to heart attacks. In addition, high blood pressure, high cholesterol, and smoking are heart disease risk factors. About half of people in the United States have at least one of these three risk factors. Diabetes, overweight and obesity, unhealthy diet, lack of exercise, and excessive alcohol consumption are other risk factors[1]. Heart disease may be asymptomatic and not diagnosed until there are signs or symptoms of a heart attack, heart failure, or arrhythmia. Therefore, it is vital to have adequate heart health monitoring tools.

Artificial intelligence (AI) and machine learning are revolutionizing medical diagnostics. Machine learning can recognize patterns in diagnostic data that a human doctor or medical technician might miss and point doctors in the right direction for diagnosis. A medical diagnosis identifies a disease or condition that describes a person's symptoms and signs. Diagnostic information is usually gathered from the patient's medical history and physical examination. For this study, we publicly utilized the Cleveland heart failure dataset on the machine learning UCI repository.

Related Works: Much work has gone into developing diagnostic systems for the early detection of heart disease using multiple clinical criteria. Many methods are used to identify patients, including logistic regression, decision trees, random forests, support vector machines, and artificial neural networks. Harshit Jindal et al.[2] used Logistic regression and KNN to achieve an accuracy of 87%. S. Mohan et al.[3] reached an accuracy level of 88.7% through the hybrid random forest with a linear model. Amin et al.[4] used a different combination of features and seven classification techniques. Experiment results show that the heart disease prediction model developed based on the identified significant features yielded an accuracy of 87.4%. S.Bashir et al.[5] applied Decision Tree, Logistic regression, Logistic regression SVM, Naïve Bayes, and Random forest, applied individually in Rapid miner on UCI heart disease data set and obtained an accuracy of 84.85%. Maji S et al. proposed a hybrid model combining the decision tree technique and the C4.5 algorithm and connected it with ANN, achieving an accuracy of 78.14%.

A thorough literature review was performed, and several other disease classification studies [6]–[9] were analyzed to construct the baseline models.

2. MATERIALS AND METHODS.

Our approach primarily compares various classification models, fine-tuning parameters, and proposes algorithms to achieve the best overall accuracy, sensitivity, and specificity. In addition, we did thorough data analysis to understand and summarize dataset characteristics.

2.1 Dataset:

This paper uses the UCI Machine Learning Repository Cleveland database containing 303 samples from patients with 14 different characteristics, as shown in Table 1. The dataset is split into a test set and a training set, with 70% used for training and the remaining 30% used for validation and testing. The dataset contains 303 patient records, 6 of which have missing values. These six records were removed from the dataset, and the remaining 297 patient records were retained for processing.

Name	Description
age	age in years
sex	sex (1 = male; 0 = female)
cp	chest pain type
trestbps	resting blood pressure in mm Hg
chol	serum cholesterol in mg/dl

FBS	fasting blood sugar > 120 mg/dl)
restecg	resting electrocardiographic results
thalach	maximum heart rate achieved
exang	exercise-induced angina
oldpeak	ST depression induced by exercise
slope	the slope of the peak exercise ST
ca	number of vessels colored by fluoroscopy
thal	Thallium stress test results
target	the predicted attribute

Table 1. Cleveland data characteristics.

The target variable is used to check for the presence of heart disease. If the patient has heart disease, the value is set to 1. Otherwise, the value is set to 0 to denote that the patient does not have heart disease. Data are preprocessed by converting medical records into diagnostic values. For example, data preprocessing of 297 patient records resulted in records showing that 137 records had a value of 1, indicating the presence of heart disease. The remaining 160 records had a value of 0, indicating no heart disease. Figure 1 shows the distribution of target labels.

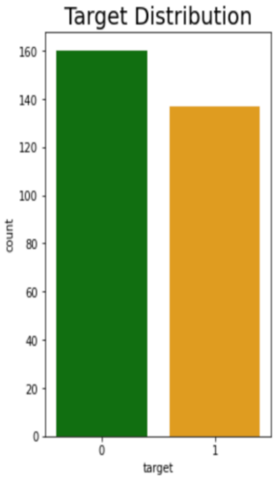

Figure 1. Target distribution.

Age: Age-related changes can increase the risk of heart disease. According to the National Institutes of Health, people over the age of 65 are much more likely to have a heart attack or develop coronary artery disease or heart failure. Data distribution based on age is indicated in figure 2.

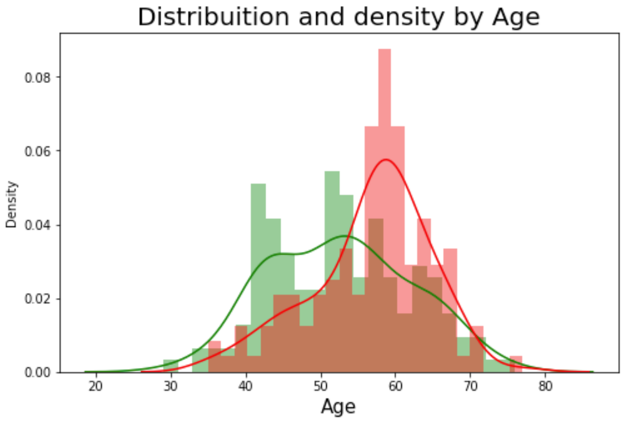

Figure 2. Density by age.

Binary variables: Sex, fasting blood sugar, and exercised-induced angina are binary variables. According to the Centers for Disease Control, high blood sugar content can damage the blood vessels that control the heart. Patients with diabetes are also more likely to have several conditions that increase the risk of heart disease. For example, angina is chest pain caused by exercise, stress, or other factors that make the heart work harder. This is a prevalent symptom of heart disease caused by the coronary arteries clogged with cholesterol. Figure 2 shows the distribution of binary variables.

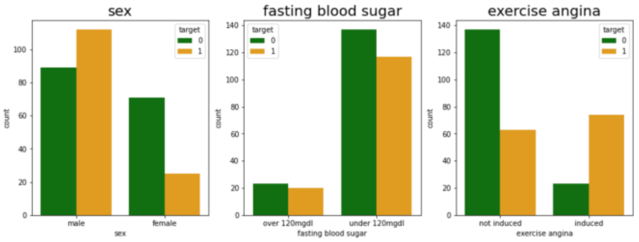

Figure 3. Binary variables distribution.

Chest Pain Type: Chest pain and discomfort are the most common symptoms of heart disease. Chest pain can occur when an artery is narrowed by excess plaque buildup. A narrowed artery can block blood flow to the heart muscles, which can cause chest pain. The diagram shows the data distribution for chest pain types. For example, chest pain distribution is shown in figure 4.

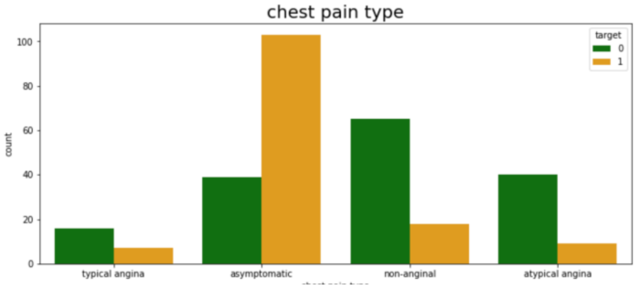

Figure 4. Chest pain distribution.

Cholesterol: Cholesterol helps the body grow new cells, protect nerves, and produce hormones. Typically, the liver makes all the cholesterol the body needs. However, cholesterol also enters the body from animal foods such as milk, eggs, and meat. Too much cholesterol in the body is a risk factor for heart disease. As a result, arteries narrow, slowing or blocking blood flow to the heart muscle. A heart attack occurs when a blockage completely prevents blood supply to part of the heart. The density distribution for cholesterol is shown in figure 5.

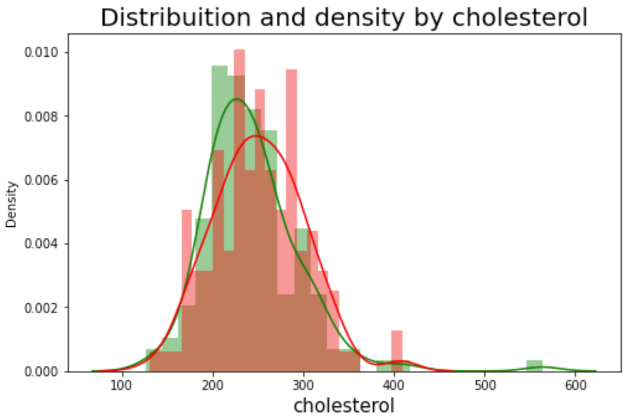

Figure 5. Density by cholesterol.

Resting ECG: A resting ECG is a standard test that measures the heart's electrical function. An ECG can be used as routine testing to check for heart disease before signs or symptoms appear. For example, resting 12-lead ECG can detect abnormalities such as arrhythmia, evidence of coronary artery disease, left ventricular hypertrophy, and bundle branch block. The resting ECG distribution is shown in figure 6.

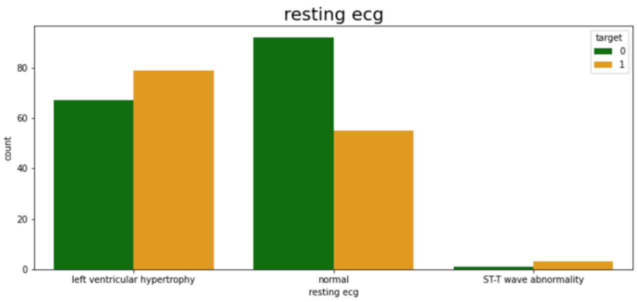

Figure 6. Resting ECG distribution.

ST Slope: On the ECG, the ST segment connects the QRS complex and T wave and lasts 0.005 to 0.150 seconds. The regular ST segment is slightly concave upwards. Therefore, a flat, downsloping, or sunken ST segment may indicate coronary artery disease. The ST slope distribution is shown in figure 7.

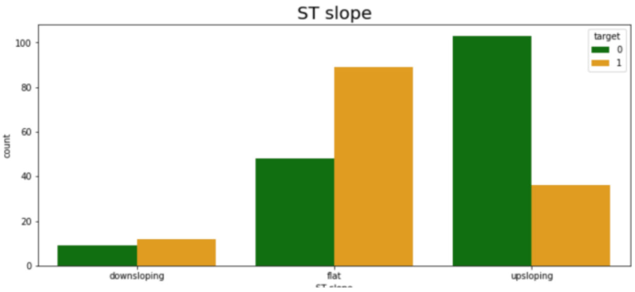

Figure 7. ST Slope distribution.

Ca: Fluoroscopy shows how blood flows through the coronary arteries and assesses whether a route is blocked. Data distribution of the number of colored arteries is shown in figure 8.

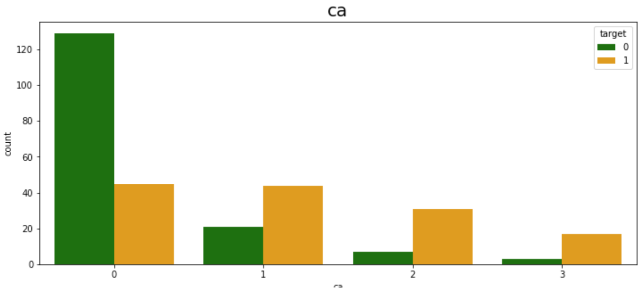

Figure 8. Ca distribution.

We evaluated the relationships between numerical features and targets to identify the components with the most significant connections to the target variables. The number of colored arteries showed the highest positive correlation, followed by old peak values, cholesterol values, resting blood pressure, and age. Maximum heart rate showed the highest negative correlation. Correlation details are shown in figure 9.

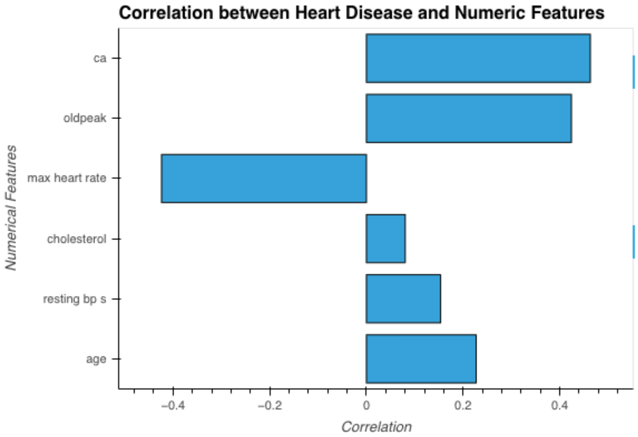

Figure 9. Correlation with Numeric features.

2.2 Base Models:

Classification models attempt to conclude observed values. Given one or more inputs, a classification model attempts to predict one or more outcome values. We used standard classification models such as logistic regression, random forest, XGBoost, and gradient boost as base models.

1. ***Logistic Regression*:** Logistic regression harnesses the power of regression for classification and has worked very well for decades, making it one of the most popular models. One of the main reasons for the model's success is that it can be explained by quantitatively calculating the contributions of individual predictors.

2. ***Random Forest*:** A decision tree builds a model in the form of a tree structure. It divides the dataset into smaller subsets and creates the associated decision trees step by step. A random forest consists of a set of individual decision trees that act as an ensemble. Each tree in the forest returns class prediction, and the class that gets the most votes become the model's prediction.

3. ***XGBoost***: Boosting is a sequential technique that works on the ensemble principle. This technique combines a set of weak learners to improve prediction accuracy. Extreme Gradient Boosting belongs to the family of boosting algorithms and uses the gradient boosting framework at its core. This is an optimized distributed gradient boosting library.

4. ***Gradient Boost***: Gradient boosting also belongs to the family of boosting algorithms and uses the gradient boosting framework at its core. With gradient boosting, each predictor improves the previous predictor by reducing its error. An important feature is that instead of fitting a predictor to the data at each iteration, the new predictor is fitted to the residual errors of the previous predictor.

2.3 TabNet:

TabNet[10] is a novel deep neural network for structured and tabular data. Traditional decision tree-based architectures learn well from tabular datasets. TabNet uses traditional DNN building blocks to return decision trees like output. TabNet uses a single deep learning architecture for feature selection and inference, known as soft function selection. We can use sequential attention to choose which features to infer at each decision step. This allows for interpretability and more efficient learning, as learning power is used for the most salient features. TabNet inputs raw tabular data without preprocessing and is trained using gradient descent-based optimization, allowing flexible integration into end-to-end learning. TabNet uses sequential attention to select features that conclude each decision step. This provides for interpretability and better learning by leveraging the ability to learn the most salient features. TabNet allows two types of interpretability, local interpretability, which visualizes the importance of features and their combinations, and global interpretability, which quantifies the contribution of each feature to the trained model.

Success Metrics

Accuracy: Accuracy is used to measure how well a binary classification test identifies or excludes conditions. In other words, accuracy is the ratio of correct predictions out of the total number of cases tested.

Area Under Curve: AUC-ROC curves are performance measures for classification problems at various threshold settings. ROC is the probability curve, and AUC represents the degree or measure of separability. Indicates how well the model can distinguish between classes. For example, the higher the AUC, the better the model can distinguish between diseased and disease-free patients.

Precision: Precision is the ratio of the true positives to the total of true positives and false positives. Precision describes the classifier's ability not to flag negative samples as positive.

Recall: Recall is the ratio of the true positives to the total of true positives and false negatives. Recall indicates the classifier's ability to find all positive samples.

Model Training: We randomly split the training data by 30% as validation and test data. The base models were trained with default hyperparameters. Tabnet has been trained up to 1000 epochs. Early stopping was achieved at 126 epochs. ROC and accuracy were used to determine early stop criteria. A list of categorical feature indices was supplied to the Tabnet classifier. Adam was used as the Pytorch optimizer function with an initial learning rate of 0.01. The 'sparsemax' masking function was used for feature selection.

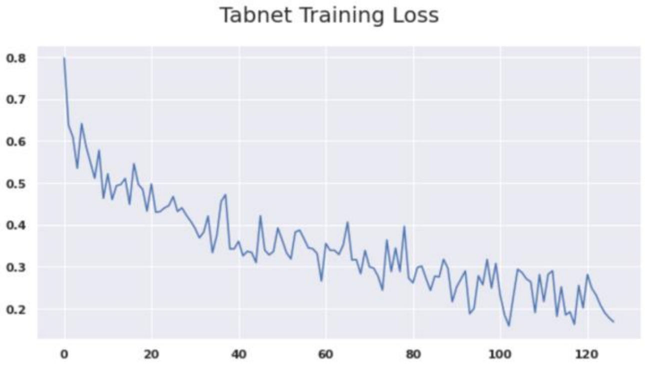

Figure 10. Tabnet training loss.

The Tabnet training loss is plotted in figure 10.

3. RESULTS

To evaluate the performance of models, we used unseen test data to predict the heart disease labels. The test set contains 36 patients randomly sampled from the full dataset with no patient overlap with the train set. Results are shown in table 2.

Results - Base Models

Model	*auroc*	*accuracy*	*precision*	*recall*
Logistic Regression	0.90	91.7%	1.00	0.80
Random Forest	0.88	88.9%	0.92	0.80
XGBoost	0.84	86.1%	0.92	0.73
Gradient Boost	0.79	80.6%	0.83	0.67

Table 2. Base models performance.

Logistic regression achieved the best performance among the base models based on the overall test metrics. This model achieved a ROC score of 0.90 and an accuracy of 91.7%. The other models had ROC values between 0.79 and 0.88 and accuracies between 81% and 86%. The confusion matrix shown in Figure 11 shows the sensitivity and specificity of different models. Logistic regression correctly predicted disease identifiers in 33 cases from a test set of 36 cases.

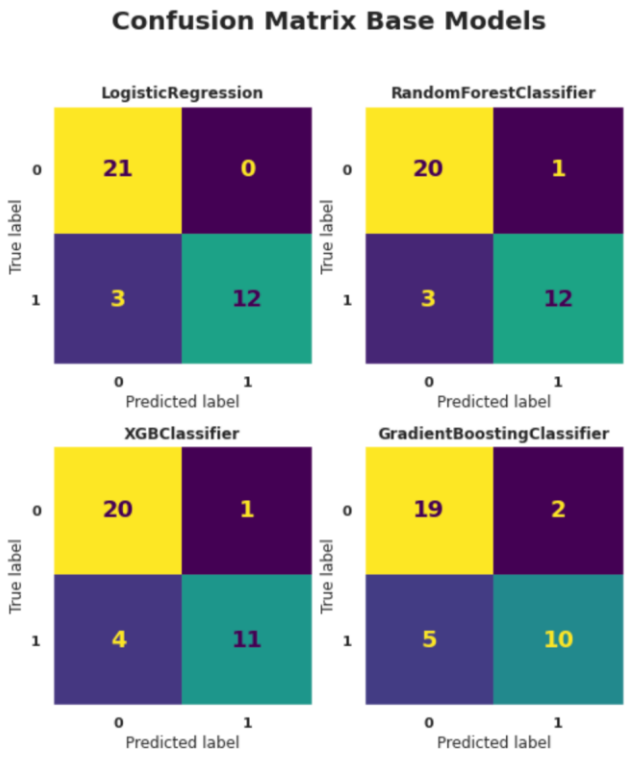

Figure 11. Confusion Matrix for base models.

<u>Results – Tabnet Classifier</u>

Model	*auroc*	*accuracy*	*precision*	*recall*
Tabnet	0.94	94.4%	0.93	0.93

Table 3. Tabnet performance.

The Tabnet classifier achieved a ROC of 0.94 and an accuracy of 94%, much better than the base models. Sensitivity and specificity exceed 0.93.

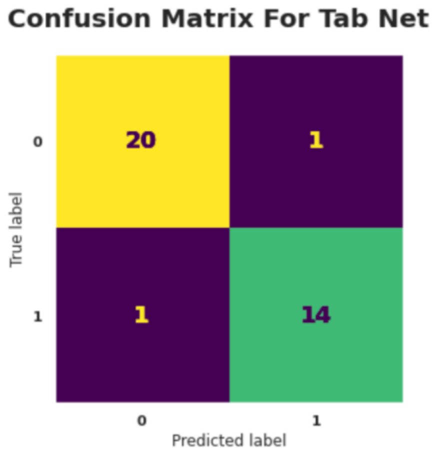

Figure 12. Confusion Matrix for Tabnet.

Out of the 36 cases in the test set, the Tabnet model correctly identified labels for 34 cases. In addition, 14 true positives and 20 true negative cases were labeled correctly. Results for the Tabnet model are shown in Table 3. The confusion matrix is shown in Figure 12.

4. DISCUSSION

The Tabnet model outperformed the other base models. Accuracy for various models is shown in Figure 13.

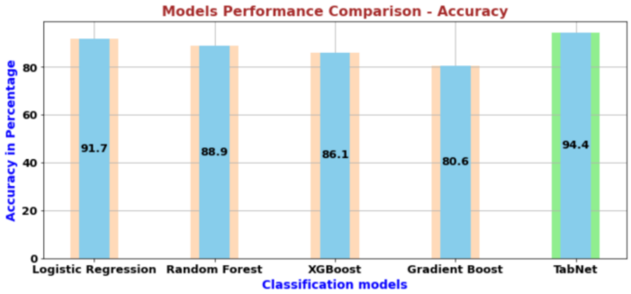

Figure 13. Performance comparison - Accuracy

A comparison of ROC scores is shown in Figure 14.

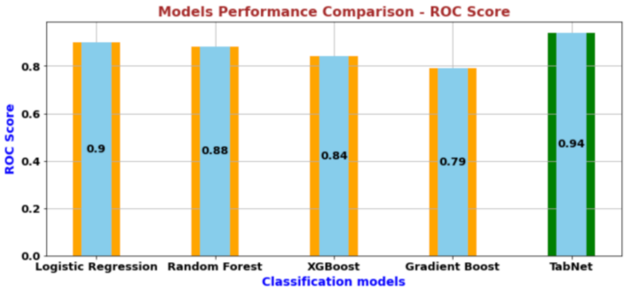

Figure 14. Performance comparison - ROC

Feature importance refers to the technique of assigning scores to input features based on their usefulness in predicting the target variable. TabNet provides access to feature rankings in terms of overall importance. Feature importance is crucial in predictive modeling, giving insight into data and models. Feature importance for the Tabnet classifier is shown in figure 15. Features ca, thal, and oldpeak contributed the most to predicting target labels.

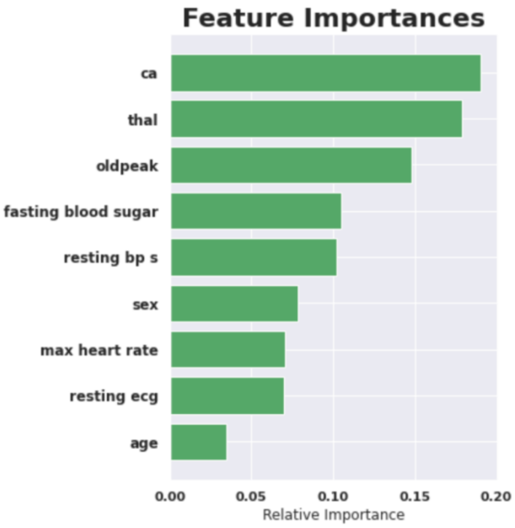

Figure 15. Feature importance

In addition to predicted values, TabNet also provides a feature importance output mask that indicates whether a feature is selected at a particular decision step in the model. The mask can be used to retrieve feature importance. The prediction output returns the aggregated mask value, as shown in figure 16.

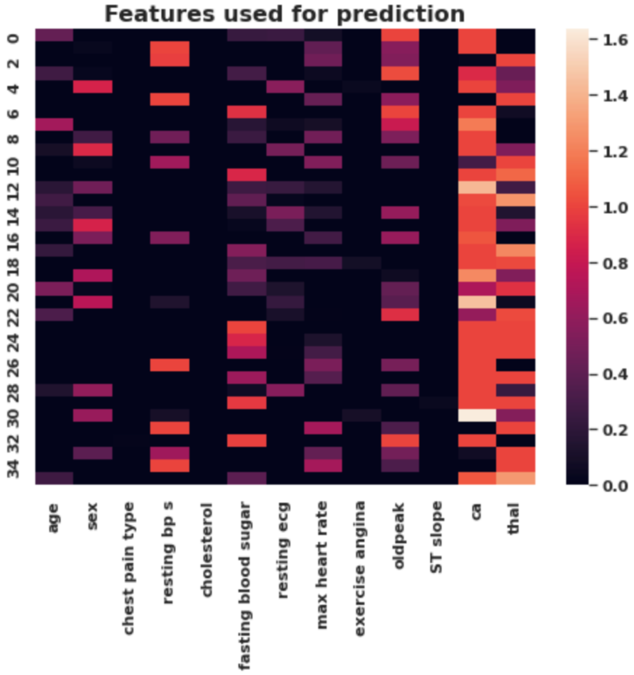

Figure 16. Feature importance masks for Tabnet

This is most useful for explaining the model. The higher the mask value for a particular sample, the more critical the corresponding feature is. Brighter colors have higher values. Each row represents a mask for each input.

5. CONCLUSIONS

Identifying and processing raw cardiac health data can save lives in the long term and help detect heart disease abnormalities early. This work used machine learning techniques to process raw data and provide new and novel differentiation for heart disease. Predicting heart disease is a challenge and of great importance in the medical field. However, mortality can be significantly reduced if the condition is detected early and preventive measures are taken as soon as possible. Therefore, extending this work to focus the investigation on larger datasets would be highly desirable. The Tabnet deep learning proposed was reasonably accurate in predicting heart disease and achieved an accuracy of over 94%. The future course of this research can be done with various combinations, from machine learning techniques to better predictive techniques. Furthermore, novel feature selection methods can be developed to gain broader recognition of essential features and enhance cardiac disease's predictive power.

CODE LOCATION

https://github.com/aravindsp/Tabular_Neural_Network-/tree/main

Notebook :
https://github.com/aravindsp/Tabular_Neural_Network-/blob/main/cardiac-disease-prediction-with-tabnet-deep-learn.ipynb

CONFLICTS OF INTEREST

The author declares no conflicts of interest regarding the publication of this paper.

REFERENCES

[1] R. Hajar, “Risk Factors for Coronary Artery Disease: Historical Perspectives.,” *Heart Views*, vol. 18, no. 3, pp. 109–114, DOI: 10.4103/HEARTVIEWS.HEARTVIEWS_106_17.

[2] H. Jindal, S. Agrawal, R. Khera, R. Jain, and P. Nagrath, “Heart disease prediction using machine learning algorithms,” in *IOP Conference Series: Materials Science and Engineering*, 2021, vol. 1022, no. 1. doi: 10.1088/1757-899X/1022/1/012072.

[3] S. Mohan, C. Thirumalai, and G. Srivastava, “Effective heart disease prediction using hybrid machine learning techniques,” *IEEE Access*, vol. 7, 2019, DOI: 10.1109/ACCESS.2019.2923707.

[4] M. S. Amin, Y. K. Chiam, and K. D. Varathan, “Identification of significant features and data mining techniques in predicting heart disease,” *Telematics and Informatics*, vol. 36, 2019, DOI: 10.1016/j.tele.2018.11.007.

[5] S. Bashir, Z. S. Khan, F. Hassan Khan, A. Anjum, and K. Bashir, “Improving Heart Disease Prediction Using Feature Selection Approaches,” in *Proceedings of 2019 16th International Bhurban Conference on Applied Sciences and Technology, IBCAST 2019*, 2019. DOI: 10.1109/IBCAST.2019.8667106.

[6] H. M. Balaha, A. O. Shaban, E. M. El-Gendy, and M. M. Saafan, “A multi-variate heart disease optimization and recognition framework,” *Neural Comput Appl*, vol. 34, no. 18, pp. 15907–15944, Sep. 2022, DOI: 10.1007/s00521-022-07241-1.

[7] A. Lahsasna, R. N. Ainon, R. Zainuddin, and A. Bulgiba, “Design of a Fuzzy-based Decision Support System for Coronary Heart Disease Diagnosis,” *J Med Syst*, vol. 36, no. 5, pp. 3293–3306, Oct. 2012, DOI: 10.1007/s10916-012-9821-7.

[8] A. S. Pillai, “Multi-Label Chest X-Ray Classification via Deep Learning,” *Journal of Intelligent Learning Systems and Applications*, vol. 14, no. 04, pp. 43–56, 2022, doi: 10.4236/jilsa.2022.144004.

[9] A. Dutta, T. Batabyal, M. Basu, and S. T. Acton, “An Efficient Convolutional Neural Network for Coronary Heart Disease Prediction,” Sep. 2019.

[10] S. O. Arik and T. Pfister, “TabNet: Attentive Interpretable Tabular Learning,” Aug. 2019.