OPEN ACCESS

ISSN: 2996-4954

Avinash A, et al. (2022)100027
J Contemp Edu Theo Artific Intel, 2022; 1(1):100103
DOI: 10.71010/2996-4954/JCETAI/103

Research Article

Journal of Contemporary Education Theory & Artificial Intelligence

Leveraging Artificial Intelligence and Big Data Analytics for Early and Accurate Identification of Heart Diseases from Electronic Health Records

Avinash Attipalli[1*], Varun Bitkuri[2], Raghuvaran Kendyala[3], Jagan Kurma[4], Jaya Vardhani Mamidala[5], Sunil Jacob Enokkaren[6]

[1]University of Bridgeport, Department of Computer Science, Attipalli.avinash@gmail.com
[2]Stratford University, Software Engineer, Varunbittu452@gmail.com
[3]University of Illinois at Springfield, Department of Computer Science, raghukend@gmail.com
[4]Christian Brothers University, Computer Information Systems, jagankurmark@gmail.com
[5]University of Central Missouri, Department of Computer Science, mvardhini29@gmail.com
[6]ADP, Solution Architect, sunil.jacob.enokkaren@gmail.com

***Corresponding author:** Avinash Attipalli

Citation: Avinash A, Varun B, Raghuvaran K, Jagan K, Jaya Vardhani M, et al. (2022) Leveraging Artificial Intelligence and Big Data Analytics for Early and Accurate Identification of Heart Diseases from Electronic Health Records. J Contemp Edu Theo Artific Intel: JCETAI-103.

Received Date: 07 December, 2022; **Accepted Date:** 14 December, 2022; **Published Date:** 20 December, 2022

Abstract

Improved treatment results and a lower chance of imminent patient death are made possible by early identification of cardiac disease. This study uses the TabNet architecture and the UCI to introduce a revolutionary deep learning method for reliably predicting heart disease. Cleveland Heart Disease Dataset. The data preprocessing is performed in an organized manner, involving the treatment of missing values, removal of duplicates, feature selection based on correlation, and standardization. The TabNet model aims to leverage sequential attention to learn interpretable features and train efficiently. According to experiment analysis, the provided model performs better, with an accuracy of 99.67%, recall of 99.97%, precision of 99.98%, and F1-score of 99.96%. The artificial neural network (ANN), logistic regression (LR), and support vector machine (SVM) accuracies are 90.40%, 86.80%, and 89.93%, respectively, when the model's performance is compared to that of traditional classifiers. Results indicate that TabNet is a viable solution that could be used in detecting early heart disease in a clinical context in an automated way and still achieve high scores compared to classical models, which do not provide any insight into the interpretation of a managed model.

Keywords: *heart disease prediction, Heart disease UCI Cleveland dataset, Machine learning, TabNet, Deep learning, artificial intelligence.*

1. Introduction

The timely diagnosis and treatment of illnesses is an essential element of contemporary medicine because it contributes to a much better prognosis to patients and reduces the lifetime cost of medical care. The early detection enables clinicians to adopt preventive or remedial measures in time before conditions reach dangerous levels [1]. This is most necessary while addressing long-term conditions like cardiovascular diseases (CVDs), which continue to be the world's leading cause of death [2]. The large numbers of mortality and morbidity rates attributed to CVD have necessitated the high priority of early detection of CVD in the healthcare systems and public health policies across the globe.

One important subset of cardiovascular illness is heart disease CVD, which most of the times develop without providing any warning before leading to life-threatening issues like heart attack or stroke [3], Angina (chest pains) and heart attack (or myocardial infarction) develops due to the impairment or blockage of blood circulatory arteries. Signs of heart disease include angina, shortness of breath, irregular heartbeats, chest pressure and abnormalities of the heart [4]. The prevention of heart diseases and alleviating the pressure on medical systems is an important step that requires early diagnosis. Nonetheless, the common diagnostic tools suffer a failure since they involve symptomatic assessment and analyzing the results manually. This discrepancy highlights the need to use data-driven approaches. Digital health information, such as Electronic Health Records (EHRs), is necessary to identify the early and milder symptoms of cardiac disease.

Electronic Medical Records EHRs are a comprehensive computerized repository for patient health data, including test results, medication histories, diagnostic notes, and vital signs [5]. Properly analyzed, such records may give excellent information about cardiac health and the development of the disease [6]. Nevertheless, the structural has a lot of data in EHR, which poses a challenge to conventional data analysis methods. This is where the adoption of big data analytics will become truly essential, providing the scalability, efficiency, and relevance of multidimensional data and healthcare data interpretation.

Healthcare systems are taking advantage of big data technologies to process and analyze the huge amounts of EHR data to identify any significant trends and risk factors of heart diseases [7]. These systems will be able to combine structured data with unstructured data, such radiology reports and clinical notes, like test results, medications, and vital signs, to gain a better understanding of that patient's health [8]. Besides making it possible to identify the high-risk individuals using predictive modeling, big data analytics can be used to track individuals in

real-time, providing early warnings and preventing the occurrence of incidents and providing timely interventions [9]. Moreover, it assists in identifying the health trends on the population level, and thus, allows the public health authorities to plan the resources efficiently and introduce the prevention efforts to the targeted population [10]. Big data can help clinicians transition to a proactive model of care by delivering personalized healthcare services through a predictive risk score model.

The ML and AI methods have been demonstrated to yield stellar results in the automation of cardiac disease recognition. ML models can analyze old EHR data [11], predict the risk of disease effectively and make clinical decisions, which are more precise than manual approaches [12]. Non-linear relationships between clinical data might be found even in such advanced algorithms as RF, SVM, and network-based TabNet and CNN.[13]. Therefore, by accurately predicting sickness in advance, AI integration with big data and EHRs has the power to completely transform the cardiac care industry.

A. Motivation and Contribution

The purpose behind this work lies in identifying a method of detecting heart diseases early and accurately as manual assessment of clinical information takes time and results in the risk of errors due to the human factor. This paper will contribute to the automation and improved performance of the diagnosis process by applying the latest DL models such as TabNet and allowing medical specialists to detect patients with high risks to achieve better clinical outcomes. In order to automatically categorise heart disease based on clinical needs, this project will apply the DL technique to the UCI Cleveland dataset. These are the primary benefits of heart disease prediction:

- The UCI Cleveland Heart Disease dataset, which has been recognized as one of the standard datasets, will also be used in this study because it is the most varied range of the clinical and demographic characteristics of relevance to the diagnosis of heart disease, thus forming an effective benchmark to assess the predictive models in the healthcare sector.
- An effective pre-processing pipeline is applied to provide the data quality, the model readiness, such as missing value, duplicated records standardization, feature selection to maintain dimension reduction and improve the learning efficiency.
- The suggested TabNet model combines attention-based feature selection and DL to provide reliable and explainable predictions, which is contrasted to the standard approaches such as SVM, RF , and LR.
- A variety of measures are used to analyse the model's performance and provide a comprehensive evaluation of the model, including measures like the confusion matrix, ROC curve, F1-score, precision, accuracy, and recall.

B. Novelty with Justification

The novelty of the work is a combination of advances in the preprocessing of data, feature selection, and DL and considers the TabNet model to automatically predict heart disease. This would be achieved through such approaches as imputation of missing values, duplicate record handle, standardization, feature selection that would ensure that the algorithm would have high-quality input besides being very powerful to learn. The nature of the TabNet architecture related to pre-processing of the data, where attention-based feature selection and DL are used, is able to meaningfully teach the model the underlying pattern of the clinical characteristics. The fact that it has better performance as compared to traditional models, including SVM, RF, and LR, means that it can manage organized medical data and help create precise and understandable early heart disease detection systems.

2. Literature Review

The discussed literature review in this section reports the use of AI and big data techniques to harvest and analyze vast volumes of data in an extremely precise and early, predict and detect heart disease under the application of an effective ML approach.

Du et al. (2020) developed a big data and ML-based high-precision coronary heart disease (CHD) model. Results Equipped with an AUC of 0.943, In the test set, the ensemble model XGBoost performed well in predicting the beginning of CHD three years later. The comparative study revealed that ML algorithms greatly outperformed standard risk scales or parametric frameworks, and Non-linear models (KNN AUC 0.908, RF AUC 0.938) can yield more accurate predictions than linear models (AUC LR 0.865) for the same datasets [14].

Krittanawong et al. (2020) are intended to assess and compile the overall prediction precision of ML-driven systems in cardiac situations. The main finding was a summary of the prognostic possibilities of ML algorithms between heart failure, stroke, coronary artery disease, and cardiac arrhythmias. In case of coronary artery disease prediction with pooled area under the curve amounting to 0.88 (95% CI 0.84091) was realized in using boosting algorithms, whereas AUC of a custom-built algorithm amounting to 0.93 (95% CI 0.85097) was obtained. The pooled AUC of CNN algorithms was 0.90 (95% CI 0.830.95), the pooled AUC of boosting algorithms was 0.91 (95% CI 0.810.96), and the pooled AUC of SVM algorithms was 0.92 (95% CI 0.810.97) in the result of stroke prediction [15].

Fitriyani et al. (2020) provide a CDSS with an efficient heart disease prognostic model (HDPM). Density-Based Spatial Clustering of Applications with Noise (DBSCAN) detects and eliminates outliers, XGBoost predicts cardiac illness, and the hybrid SMOTE-ENN distributes training data evenly. The results of other models, including NB, LR, MLP, SVM, DT, and RF, as well as those of earlier research, were also contrasted with the model's output. Cleveland and Statlog, two publicly available datasets, were used to build the model. The proposed model outperformed the other models and the results of the prior study, achieving an accuracy of 95.90 percent for the Statlog dataset and 98.40 percent for the Cleveland dataset [16].

Sivakumar et al. (2020) In order to explore mental disease risk factors and create a model for forecasting mental illness, present the concept of comorbidities, drug usage, and dietary supplements in individuals with heart disease. In particular, the research's data should be regarded as the medical records of 68,647 heart disease patients, which include details on their comorbidity, usage of dietary supplements, use of antibiotics, and mental health. The depression and mental diseases were linked with gender differences, age (below 61 years) and medicine intake i.e. clarithromycin, azithromycin, vitamin B6 and coenzyme Q10. It is worth noting that a combination of various state-of-the-art ML methods and their properly trained parameters is a good choice when it comes to predictive modelling, which will eventually lead to the following: AUC depression 78.01% accuracy, 72.65% specificity, 79.13% sensitivity, and 86.26%. With 82.93% accuracy, 82.86%

sensitivity, and 83.35% specificity recorded, the anxiety AUC was 88.45%. 92.73% AUC, 85.14% specificity, 87.70% sensitivity, and 87.59% accuracy for schizophrenia. 77.76% specificity, 91.59% AUC, 86.63% accuracy, and 95.50% sensitivity for the disease [17].

Atallah and Al-Mousa (2019) The study trains the model using real data from both healthy and ill patients, which should improve the accuracy and consistency of the doctor's diagnosis. The model generates more accurate results by identifying the patient using the majority vote of many models rather than just one ML model. Ultimately, this method yielded a 90% accuracy rate using the hard voting ensemble model [3].

Javeed et al. (2019) The suggested diagnostic method predicts heart failure using an RF model and selects characteristics using the random search algorithm (RSA). The grid search technique is used to improve the suggested diagnostic system. The Cleveland dataset, an online heart failure database, is used in the studies with just seven characteristics, The suggested approach beats the traditional random forest model by 3.3%, demonstrating its efficacy and simplicity of usage. Furthermore, the suggested approach outperforms five additional cutting-edge ML models. Furthermore, the suggested approach improved the training accuracy while achieving a 93.33% classification accuracy.[18].

Table I presents a comparative summary of recent studies employing the methods, datasets, benefits, drawbacks, and suggested future research paths of ML and DL approaches for heart disease prediction.

Table 1: Summary of the related work for heart disease prediction using machine /deep learning techniques.

Author	Dataset	Methodology	Advantages	Limitations	Future Work
Du et al. (2020)	Population dataset (split into training and testing sets)	ML models including XGBoost, KNN, RF, Logistic Regression; AUC-based evaluation	XGBoost achieved AUC of 0.943; ML models outperformed traditional risk scales	Linear models underperformed; dataset specific to 3-year CHD onset	Explore explainability of models and apply in clinical settings
Krittanawong et al. (2020)	Meta-analysis of 103 studies (3.37M individuals)	Various ML algorithms (boosting, SVM, CNN, custom-built) across multiple cardiovascular conditions	Comprehensive analysis; pooled AUC up to 0.93 for custom algorithms	Heterogeneity in study designs; possible publication bias	Standardize ML model evaluation for cardiovascular diseases
Fitriyani et al. (2020)	Statlog & Cleveland datasets	DBSCAN (outlier removal), SMOTE-ENN (balancing), XGBoost classifier	Achieved high accuracies (95.90%, 98.40%); handles imbalance and noise well	Public datasets may limit real-world generalizability	Apply to real-time CDSS environments and real-world data
Sivakumar et al. (2020)	68,647 heart disease patients	ML models for predicting mental illnesses using comorbidities and supplement/drug intake	Strong AUC scores (up to 92.73%); multi-disease prediction	May lack external validation; domain-specific features	Expand to other mental illness types; personalize treatment recommendations
Atallah et.al. (2019)	Real-life patient data	Hard voting ensemble of several ML models	Improved accuracy (90%) using ensemble learning	No model interpretability discussed; no details on feature importance	Include interpretability tools like SHAP; explore deep learning ensembles
Javeed et al. (2019)	Cleveland dataset	RSA for feature selection, RF for classification, grid search for optimization	High accuracy (93.33%) with only 7 features; lower complexity	Focused only on Cleveland dataset; lacks multi-center validation	Validate on more diverse datasets; apply to real-time monitoring

3. Methodology

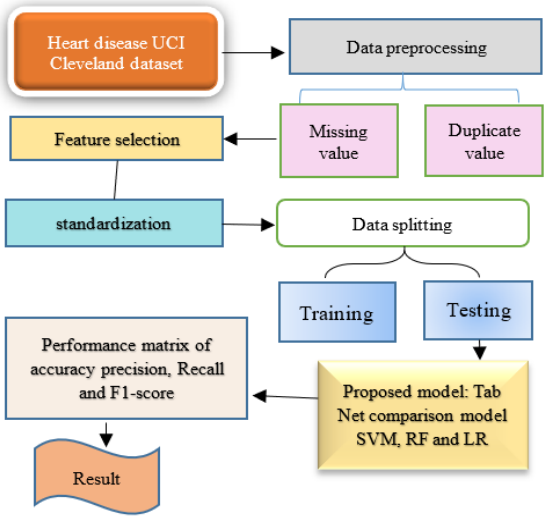

Fig 1: Flowchart for Heart Disease Prediction Using Machine Learning Models.

The UCI Cleveland dataset on heart illness serves as the starting point for the suggested approach, which is shown in Figure 1. It goes through a thorough data preparation stage. To maintain data quality, this entails removing duplicate entries and managing missing information by imputation or removal. Following preprocessing, relevant features are selected by analyzing statistical importance and correlation to eliminate redundant or less impactful variables. After the dataset has been improved, it is divided into input characteristics and target labels. Standardization is applied using normalization techniques to scale numerical features, enhancing model convergence. To assess model generalisation, Subsets of the dataset are then separated for testing and training. For predictive modeling, a novel TabNet-based architecture is proposed and evaluated. The same dataset is also used to train and assess similar models, such as LR, SVM, and ANN, to determine their performance. Standard measures such as F1-score, recall, accuracy, and precision are used to evaluate performance. This systematic pipeline ensures reliable, interpretable, and accurate heart disease prediction.

Each phase is described in the sections that follow, along with the approach and proposed flowchart.

a. Data Collection

The UCI ML Repository offers heart disease a well-liked resource for developing heart disease prediction algorithms is the Cleveland dataset. The Cleveland Clinic Foundation, along with V.A. Medical Centre, provides thirteen. These consist of the following: age, sex, kind of chest pain (cp), max heart rate (thalach), old peak ST depression, slope, number of major arteries (ca), and other clinical and demographic information. The target variable, condition, shows whether heart disease is present (1) or not (0). The dataset is often utilized for training and assessing ML models in cardiovascular risk prediction research, despite its tiny size. Figure 2 displays the important feature pair plot with colour for diagnosis.

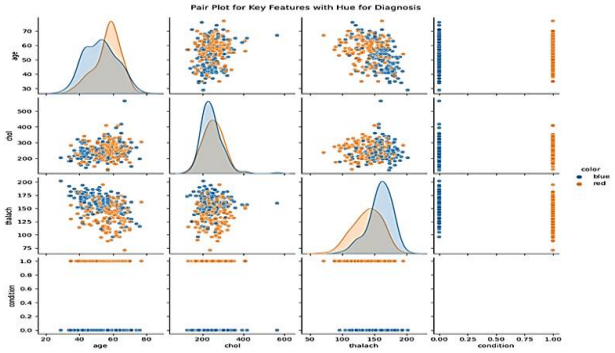

Fig 2: Pair Plot for Key Feature with Hue for Diagnosis.

Figure 2 displays the pair plot visualization representations. The connection between the identification of heart disease as well as crucial factors including age, cholesterol, and maximum heart rate (thalach). Patients with heart illness are identified by their color (red) and those without (blue). Diagonal plots show the kernel density estimation (KDE) for each feature, while scatter plots illustrate feature correlations, enabling pattern recognition and assisting in identifying features most relevant to heart disease prediction.

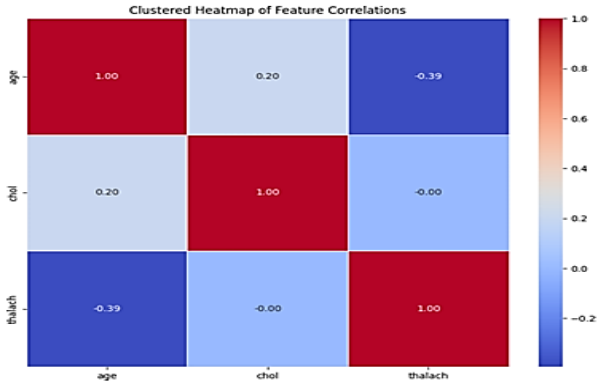

Fig 3: Clustered Heatmap of Feature Correlation.

Figure 3 shows the clustered heatmap that shows the correlation between key features: age, maximal heart rate (thalach), and cholesterol (Chol). Age and thalach have a moderately negative association (-0.39), but age and Chol have a weakly positive connection (0.20). The negligible correlation between Chol and thalach (−0.00) indicates low multicollinearity, suggesting that each feature contributes independently to heart disease prediction.

b. Data Preprocessing

Data preprocessing involves turning unprocessed information into an understandable and practical format prior to utilizing it in data analysis or ML models. It involves several techniques to missing values and duplicate records, transforming, and structure The data is used to enhance the models' functionality and accuracy. Key steps in data preprocessing include:

- **Missing value:** Missing values were checked across all features. Records with significant missing data were removed, while those with minimal gaps were imputed using appropriate techniques mean or mode ensuring the dataset remained complete and consistent.
- **Duplicate value:** Duplicate rows were identified using exact-match checks across all columns. These redundant records were dropped to prevent bias, overfitting, or data leakage during model training.

c. Feature Selection

To increase interpretability and reduce overfitting, a prediction model is constructed by selecting the most pertinent characteristics from a dataset, and improving model performance is known as feature selection. Figure 4 displays the violin plot distribution of "Chol" (cholesterol) and "thalach" (maximum heart rate obtained) across two diagnostic groups to help with feature selection. For 'Chol', the distributions for both diagnosis groups largely overlap.

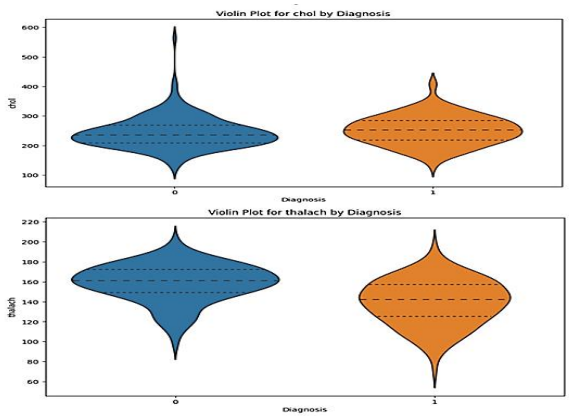

Fig 4: Violin Plot for Key Feature by Diagnosis.

With very similar medians and ranges, suggesting it may not be a strong feature for differentiating between the diagnoses. 'thalach' exhibits distinct distributions between the groups, with noticeably separated medians and less overlap in their interquartile ranges, indicating that 'thalach' is a more promising feature for distinguishing between the two diagnostics.

d. Standardization

Scaling the risk variables and allocating numbers that demonstrate the variation between standard deviations from There are two methods to standardize data: the mean value. Standardization helps ensure that the data is of a similar quality and highlights the importance of particular attributes in the learning process. Model evaluation and validation of unreported data are made feasible by dividing the dataset into testing and training sets. Equation (1) defined as:

$$standardization\ of\ X = (X - mean\ of\ X)/ (standard\ deviation\ of\ X) \quad (1)$$

In order to improve the performance of ML classifiers, the risk factor value is rescaled to have a standard deviation (σ) of 1 and a mean (μ) of 0.

e. Data Splitting

This resampled dataset is split using the train-test split technique into a training 95% and a testing 5% group. To ensure equitable representation for each group, the training and testing datasets are separated by data categorization.

f. Proposed Tabnet Model for Heart Disease Prediction

A decoder architecture is incorporated, each decision step's decoder consists of FC layers and a feature transformer, which merges the outputs to recreate features. Other feature columns can be used to forecast missing feature columns. Assume that r is the proportion of pretraining characteristics that will be arbitrarily disregarded during rebuilding, and that S∈ {0,1} $B\times d$S∈ {0,1} B×d is a binary mask. Consequently, the variable r in Figure (5) denotes the masking ratio inside the binary mask SS and its design. Equation (2) defined as:

$$L\_rec = \sum\_(i = 1)\^B \sum\_j\^d |(((x)\hat{}\_(i,j) - x\_(i,j))s\_(i,j))/(\sqrt{(\sum\_(i = 1)\^B [\![(x\_(i,j)]\!]) - 1/B\sum\_(i = 1)\^B x\_(i,j)\ )\^2\ )| \quad (2)$$

Where the original input is indicated by $\hat{x}\_(i,j)$ and the reconstructed output is represented by $x\_(i,j)$. In order for the model to concentrate on the known characteristics, the term in the encoder is initialised as [0]=(1−S)P[0]=(1−S). In Equation (2), the decoder's final FC layer is the result of multiplying SS by the unknown output characteristics.

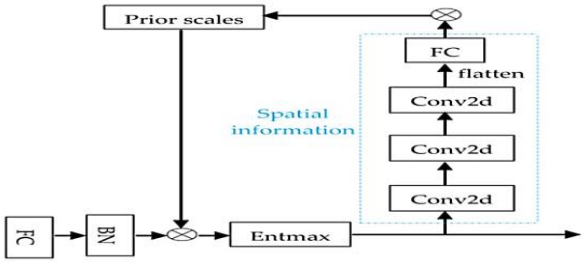

Fig 5: Architecture for the Tabnet Model.

CNN uses 2D kernels to convolve the input data after the kernel product is calculated plus the input data total. The kernel covers the whole geographical region using the provided data. An activation function is added to the convolved features to induce nonlinearity. With the l feature map, the value for the k, layer following activation $A_(k,l)\^(u,v)$at spatial point (u,v) It may be written as Equation (3).

$$A\_(k,l)\^(u,v) = \Psi(e\_(k,l) + \sum\_(\delta = 1)\^(Om - 1) \sum\_(\theta = -\tau)\^\tau \sum\_\varphi\^\varphi f\_(k,l,\delta)\^(\beta,\theta)\ \times A\_(k - 1,l)\^(\mu + \beta, v + \beta)\ ) \quad (3)$$

Where the bias parameter is l and the function of activation is represented by Ψ. With the depth of the kernel fk,l at the k layer for the l Feature map, om−1 indicates the number of feature mappings in the $(m-1$Th layer. With weight parameters fk,l, The kernel's width is represented by 2τ+1 and its height by 2Φ+1.

g. Performance Matrix

A few performance metrics might be employed to evaluate the efficacy of the proposed methodology. A number of factors are used to evaluate a system's efficacy in DL. Several performance indicators, such as F1 score, recall, accuracy, and precision. The values of the confusion matrix, which are shown in Figure 6 as TP, TN, FP, and FN, must be understood before they can be identified.

		Predicted	
		has heart disease (Positive)	no heart disease (Negative)
Actual	has heart disease (Positive)	TP	FN
	no heart disease (Negative)	FP	TN

Fig 6: Confusion Matrix N Heart Disease Prediction.

- **True positive (TP):** The data supports the favorable prediction made by the model.
- **True Negative (TN):** The data will be negative, based on the model.
- **False Positive (FP):** The data is TN, despite according to the model, it would be favorable.
- **False Negative (FN):** The data is TP, even though it was predicted by the model to be negative.

1) Accuracy

The proportion of accurate forecasts to all predictions, as provided by Equation (4):

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \times 100 \quad (4)$$

2) *Precision*

This measure of performance assesses Among the noteworthy cases involving the recovered instances is precision. The following is the precision Equation (5).

$$Precision = TP/(TP + FR) \times 100 \quad (5)$$

3) Recall

Recall is a small percentage of relevant instances that are recovered out of the entire number of examples that are pertinent. The recall Equation (6) may be found below:

$$Recall = TP/(TP + FN) \times 100 \quad (6)$$

4) F1 Score

The F-measure is obtained by multiplying the total of recall and accuracy by the precision, and then multiplying the result by the recall. The F-Measure Equation (7) is shown below:

$$F1 - \text{score} = (2 \times recall \times precision)/(recall + precision) \quad (7)$$

5) ROC Curve

The effectiveness of classification algorithms is assessed using ROC curves. The graph shows the FPR at different threshold values in relation to the TPR, or recall.

4. Results and Discussion

This section displays the experimental findings for DL-based heart disease prediction utilizing the Cleveland heart disease datasets from UCI. The performance of the model is evaluated for clinical classification tasks using a number of crucial measures, such as F1-score, recall, accuracy, and precision. The recommended TabNet architecture comparing ML models for heart disease prediction analysis has high computational requirements. Consequently, the computer platform was chosen to have an NVIDIA RTX 3070 GPU with 8 GB of VRAM and 32 GB of RAM. In addition to Google Colab, Jupiter Notebook, and Python, this platform includes the required Python libraries, such as scikit-learn, Keras, pandas, NumPy, seaborn, TensorFlow, and matplotlib. The next sections present the results of proposed methods for predicting heart disease using the TabNet model.

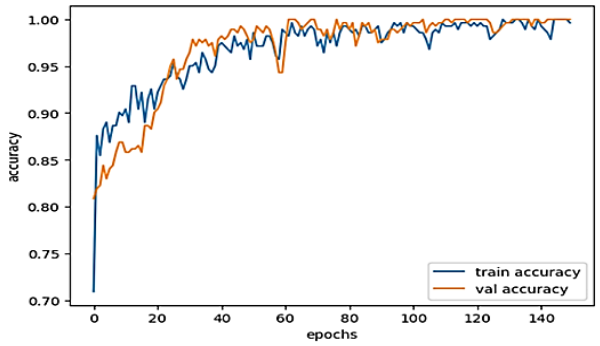

Fig 7: Accuracy Graph of Tabnet Model.

Figure 7 illustrates a rapid improvement in accuracy from approximately 75% to over 99% within the first 50 epochs, with both training (blue) and validation (orange) accuracies converging and stabilizing around 99.5-100%. The overlapping curves throughout the training process, confirm TabNet's effectiveness for reliable heart disease prediction.

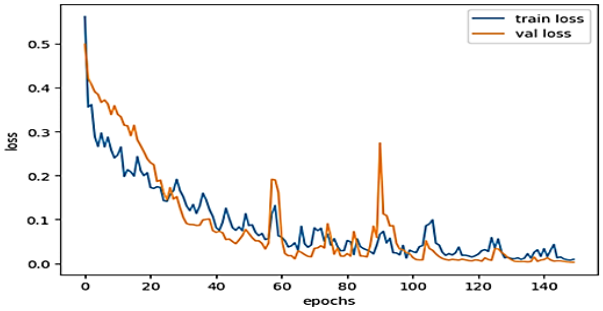

Fig 8: Loss Graph of Tabnet Model.

Figure 8 shows both demonstrate effective model training with both training (blue) and validation (orange) losses rapidly decreasing from approximately 0.5 to near-zero values over 150 epochs. The close alignment between loss curves indicates excellent generalization without overfitting, validating TabNet's efficiency for accurate heart disease prediction.

Table 2: TabNet model Performance on Heart disease UCI Cleveland dataset.

Measure	TabNet
Accuracy	99.67%
Precision	99.98%
Recall	99.97%
F1-score	99.96%

Table II: TabNet model performance assessment on the UCI Cleveland dataset on heart disease. With every evaluation measure above 99.6%, the TabNet model demonstrated remarkable performance on the Heart Disease UCI Cleveland dataset. The model's accuracy, precision, recall, and F1-score were 99.67%, 99.98%, and 99.96%, in that order. These results indicate that the TabNet architecture effectively captured the underlying patterns in the heart disease dataset, achieving near-perfect classification performance with minimal FP and FN. The consistently high values across all metrics suggest robust model generalization and reliable predictive capability for heart disease diagnosis applications.

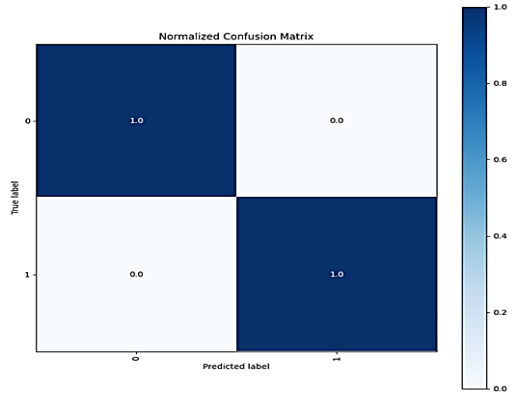

Fig 9: Confusion Matrix of Tabnet Model.

The TabNet model's normalized confusion matrix for classifying heart illness using the UCI Cleveland dataset is displayed in Figure 9. With diagonal values of 1.0 for both classes (0 and 1) and off-diagonal elements of 0.0, the normalized confusion matrix shows that in terms of classification, the TabNet model is faultless; it correctly differentiates between people with and without heart disease without producing any FP or FN.

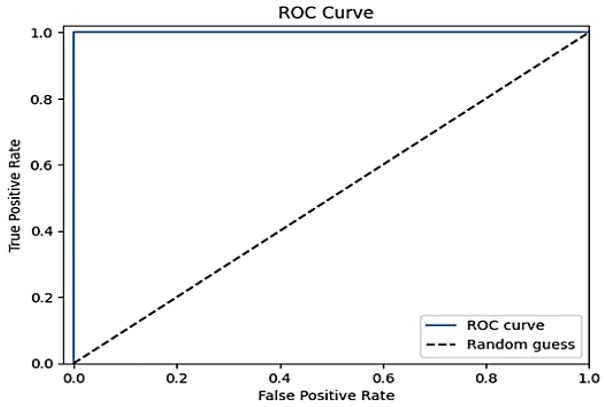

Fig 10: ROC Curve of Tabnet Model.

The TabNet model's ROC curve, seen in Figure 10, demonstrates exceptional classification performance. The curve has a low FPR and a high TPR, and it is concealed in the upper-left corner. This implies that the model forecasts cardiac illness with nearly perfect class discrimination using electronic medical data.

h. Comparative Discussion

This section uses the UCI Cleveland heart disease dataset to compare image classification methods for heart disease prediction. Table III presents the performance measures of four ML models in predicting heart diseases, indicating that in terms of performance measures, the TabNet model performs better than the existing ANN, SVM, and LR ML models. TabNet outperformed the other three models, achieving a 99.96 F1-score, 99.97 recall, 99.98 overall quality, and 99.67 accuracy. However, the SVM model's performance was more modest, with an accuracy of 89.93%. Among the compared models, the LR model was accurate as well, at 86.80 percent, but still, it was not that close to the TabNet model. The results clearly demonstrate that, in comparison to these conventional ML approaches, the TabNet architecture provides scheme much improves the precision and dependability of predicting heart disease.

Table 3: Comparison between TabNet and Existing models for heart disease prediction.

Measure	Accuracy
Proposed TabNet	99.67%
ANN[19]	90.40%
SVM[20]	89.93%
LR[21]	86.80%

The heart illness prediction record achieved a 99.67% accuracy grade, demonstrating the excellent performance of the suggested TabNet architecture. Based on attention-based feature selection and the DL paradigm, it effectively depicts complex correlations between clinical and demographic data, enabling precise and comprehensible projections. The interpretability embedded in it assists clinicians in identifying important parameters being made during the decision-making process, therefore, it is extremely useful in health assurance. However, the model is likely to overfit small or imbalanced data, and its population to different people is to be verified. Altogether, TabNet is a potentially dependable and viable approach to making clinical practice predictions of heart disease.

5. Conclusion and future work

Data-driven algorithms are used to forecast who is most likely to acquire heart disease. Accurate prediction can help with early diagnosis, intervention, and improved patient outcomes in the case of heart disease. Both DL and tabnet are the foundations of this prediction model. The accuracy and longevity of the model were increased by using a comprehensive data pre-processing pipeline. The TabNet model that was suggested was compared to the classical ML approaches such as LR and SVM. A maximum value of accuracy was 99.67, which means that TabNet proved to be remarkably better than ANN, SVM, and LR (90.40%, 89.93%, and 86.80%, respectively). Besides excellent predictive capability, TabNet has an attention mechanism, which helps to make it clearer and allows you to understand what factors mostly contribute to predicting heart disease. This is a key skill, especially in clinical practice where it would be important to understand why one is predicting. Interpretability supported by TabNet also allows being open and welcoming towards AI-based solutions in healthcare. Given its reliability and comprehensibility, all of the findings point to the potential application of TabNet in the early identification of cardiac conditions. The second crucial step in the future is to verify the model so that it may be used to larger and more diverse datasets. Additionally, an attempt will be made to integrate the model into clinical decision support systems that operate in real time, and investigate the possibility of using explainable AI approaches to increase the model's transparency and foster user confidence.

References

1. F. I. Alarsan and M. Younes, "Analysis and classification of heart diseases using heartbeat features and machine learning algorithms," *J. Big Data*, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0244-x.
2. A. O. Badheka *et al.*, "ST-T Wave Abnormality in Lead aVR and Reclassification of Cardiovascular Risk (from the National Health and Nutrition Examination Survey-III)," *Am. J. Cardiol.*, vol. 112, no. 6, pp. 805–810, Sep. 2013, doi: 10.1016/j.amjcard.2013.04.058.

3. R. Atallah and A. Al-Mousa, "Heart Disease Detection Using Machine Learning Majority Voting Ensemble Method," in *2019 2nd International Conference on New Trends in Computing Sciences (ICTCS)*, IEEE, Oct. 2019, pp. 1–6. doi: 10.1109/ICTCS.2019.8923053.
4. M. Woodward, "Cardiovascular disease and the female disadvantage," 2019. doi: 10.3390/ijerph16071165.
5. D. N. Marckini, B. P. Samuel, J. L. Parker, and S. C. Cook, "Electronic health record associated stress: A survey study of adult congenital heart disease specialists," *Congenit. Heart Dis.*, vol. 14, no. 3, pp. 356–361, May 2019, doi: 10.1111/chd 12745.
6. J. Maiga, G. G. Hungilo, and Pranowo, "Comparison of Machine Learning Models in Prediction of Cardiovascular Disease Using Health Record Data," in *Proceedings - 1st International Conference on Informatics, Multimedia, Cyber and Information System, ICIMCIS 2019*, 2019. doi: 10.1109/ICIMCIS48181.2019.8985205.
7. K.Gayathri, N. U. Maheswari, and G.Mariammal, "A Critique on Heart Diseases Predictive Analytics using Big Data Algorithms," *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 9S2, pp. 794–798, Aug. 2019, doi: 10.35940/ijitee.I1164.0789S219.
8. A. Ismail, S. Abdlerazek, and I. M. El-Henawy, "Big Data Analytics In Heart Diseases Prediction," *J. Theor. Appl. Inf. Technol.*, vol. 98, p. 11, 2020.
9. A. Balasubramanian, "Intelligent Health Monitoring: Leveraging Machine Learning and Wearables for Chronic Disease Management and Prevention," *Int. J. Innov. Res. Eng. Multidiscip. Phys. Sci.*, vol. 7, no. 6, pp. 1–13, 2019, doi: 10.5281/zenodo.14535443.
10. O. H. Salman, A. A. Zaidan, B. B. Zaidan, Naserkalid, and M. Hashim, "Novel Methodology for Triage and Prioritizing Using 'Big Data' Patients with Chronic Heart Diseases Through Telemedicine Environmental," *Int. J. Inf. Technol. Decis. Mak.*, vol. 16, no. 05, pp. 1211–1245, Sep. 2017, doi: 10.1142/S0219622017500225.
11. P. Mathur, S. Srivastava, X. Xu, and J. L. Mehta, "Artificial Intelligence, Machine Learning, and Cardiovascular Disease," *Clin. Med. Insights Cardiol.*, vol. 14, Jan. 2020, doi: 10.1177/1179546820927404.
12. V. Gupta, Dr.Pallavi, and M. Goel, "Heart Disease Prediction Using ML," *Int. J. Comput. Sci. Eng.*, vol. 7, no. 6, pp. 17–19, Jun. 2020, doi: 10.14445/23488387/IJCSE-V7I6P105.
13. I. Preethi and K. Dharmarajan, "Diagnosis of chronic disease in a predictive model using machine learning algorithm," in *2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE)*, IEEE, Oct. 2020, pp. 191–196. doi: 10.1109/ICSTCEE49637.2020.9276957.
14. Z. Du *et al.*, "Accurate Prediction of Coronary Heart Disease for Patients With Hypertension From Electronic Health Records With Big Data and Machine-Learning Methods: Model Development and Performance Evaluation," *JMIR Med. Informatics*, vol. 8, no. 7, p. e17257, Jul. 2020, doi: 10.2196/17257.
15. C. Krittanawong *et al.*, "Machine learning prediction in cardiovascular diseases: a meta-analysis," *Sci. Rep.*, vol. 10, no. 1, p. 16057, Sep. 2020, doi: 10.1038/s41598-020-72685-1.
16. N. L. Fitriyani, M. Syafrudin, G. Alfian, and J. Rhee, "HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System," *IEEE Access*, vol. 8, pp. 133034–133050, 2020, doi: 10.1109/ACCESS.2020.3010511.
17. J. Sivakumar, S. Ahmed, L. Begdache, S. Jain, and D. Won, "Prediction of Mental Illness in Heart Disease Patients: Association of Comorbidities, Dietary Supplements, and Antibiotics as Risk Factors," *J. Pers. Med.*, vol. 10, no. 4, p. 214, Nov. 2020, doi: 10.3390/jpm10040214.
18. A. Javeed, S. Zhou, L. Yongjian, I. Qasim, A. Noor, and R. Nour, "An Intelligent Learning System Based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection," *IEEE Access*, vol. 7, pp. 180235–180243, 2019, doi: 10.1109/ACCESS.2019.2952107.
19. A. K. Gárate-Escamila, A. H. El Hassani, and E. Andrès, "Classification models for heart disease prediction using feature selection and PCA," *Informatics Med. Unlocked*, vol. 19, 2020, doi: 10.1016/j.imu.2020.100330.
20. H. M. Le, T. D. Tran, and L. Van Tran, "Automatic Heart Disease Prediction Using Feature Selection And Data Mining Technique," *J. Comput. Sci. Cybern.*, vol. 34, no. 1, pp. 33–48, Aug. 2018, doi: 10.15625/1813-9663/34/1/12665.
21. D. Khanna, R. Sahu, V. Baths, and B. Deshpande, "Comparative Study of Classification Techniques (SVM, Logistic Regression and Neural Networks) to Predict the Prevalence of Heart Disease," *Int. J. Mach. Learn. Comput.*, vol. 5, no. 5, pp. 414–419, Oct. 2015, doi: 10.7763/IJMLC.2015.V5.544.
22. Polu, A. R., Vattikonda, N., Buddula, D. V. K. R., Narra, B., Patchipulusu, H., & Gupta, A. (2021). Integrating AI-Based Sentiment Analysis With Social Media Data For Enhanced Marketing Insights. *Available at SSRN 5266555*.
23. Katari, A., & Kalla, D. (2021). Cost Optimization in Cloud-Based Financial Data Lakes: Techniques and Case Studies. *ESP Journal of Engineering & Technology Advancements (ESP-JETA)*, *1*(1), 150-157.
24. Polu, A. R., Vattikonda, N., Gupta, A., Patchipulusu, H., Buddula, D. V. K. R., & Narra, B. (2021). Enhancing Marketing Analytics in Online Retailing through Machine Learning Classification Techniques. *Available at SSRN 5297803*.
25. Polu, A. R., Buddula, D. V. K. R., Narra, B., Gupta, A., Vattikonda, N., & Patchipulusu, H. (2021). Evolution of AI in Software Development and Cybersecurity: Unifying Automation, Innovation, and Protection in the Digital Age. *Available at SSRN 5266517*.
26. Gupta, K., Varun, G. A. D., Polu, S. D. E., & Sachs, G. Enhancing Marketing Analytics in Online Retailing through Machine Learning Classification Techniques.
27. Kakani, A. B., Nandiraju, S. K. K., Chundru, S. K., Vangala, S. R., Polam, R. M., & Kamarthapu, B. (2021). Big Data and Predictive Analytics for Customer Retention: Exploring the Role of Machine Learning in E-Commerce. *International Journal of Emerging Trends in Computer Science and Information Technology*, *2*(2), 26-34.
28. Polam, R. M., Kamarthapu, B., Kakani, A. B., Nandiraju, S. K. K., Chundru, S. K., & Vangala, S. R. (2021). Data Security in Cloud Computing: Encryption, Zero Trust, and Homomorphic Encryption. *International Journal of Emerging Trends in Computer Science and Information Technology*, *2*(3), 70-80.

29. Polam, R. M., Kamarthapu, B., Kakani, A. B., Nandiraju, S. K. K., Chundru, S. K., & Vangala, S. R. (2021). Big Text Data Analysis for Sentiment Classification in Product Reviews Using Advanced Large Language Models. *International Journal of AI, BigData, Computational and Management Studies*, *2*(2), 55-65.
30. Kamarthapu, B., Kakani, A. B., Nandiraju, S. K. K., Chundru, S. K., Vangala, S. R., & Polam, R. M. (2021). Advanced Machine Learning Models for Detecting and Classifying Financial Fraud in Big Data-Driven. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *2*(3), 39-46.
31. Vangala, S. R., Polam, R. M., Kamarthapu, B., Kakani, A. B., Nandiraju, S. K. K., & Chundru, S. K. (2021). Smart Healthcare: Machine Learning-Based Classification of Epileptic Seizure Disease Using EEG Signal Analysis. *International Journal of Emerging Research in Engineering and Technology*, *2*(3), 61-70.
32. Gangineni, V. N., Pabbineedi, S., Penmetsa, M., Bhumireddy, J. R., Chalasani, R., & Tyagadurgam, M. S. V. (2021). Strengthening Cybersecurity Governance: The Impact of Firewalls on Risk Management. *International Journal of AI, BigData, Computational and Management Studies*, *2*(4), 60-68.
33. Pabbineedi, S., Penmetsa, M., Bhumireddy, J. R., Chalasani, R., Tyagadurgam, M. S. V., & Gangineni, V. N. (2021). An Advanced Machine Learning Models Design for Fraud Identification in Healthcare Insurance. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *2*(1), 26-34
34. Tyagadurgam, M. S. V., Gangineni, V. N., Pabbineedi, S., Penmetsa, M., Bhumireddy, J. R., & Chalasani, R. (2021). Enhancing IoT (Internet of Things) Security Through Intelligent Intrusion Detection Using ML Models. *International Journal of Emerging Research in Engineering and Technology*, *2*(1), 27-36
35. Penmetsa, M., Bhumireddy, J. R., Chalasani, R., Tyagadurgam, M. S. V., Gangineni, V. N., & Pabbineedi, S. (2021). Next-Generation Cybersecurity: The Role of AI and Quantum Computing in Threat Detection. *International Journal of Emerging Trends in Computer Science and Information Technology*, *2*(4), 54-61.
36. Krutthika H. K. & A.R. Aswatha. (2021). Implementation and analysis of congestion prevention and fault tolerance in network on chip. Journal of Tianjin University Science and Technology, 54(11), 213–231. https://doi.org/10.5281/zenodo.5746712
37. Krutthika H. K. & A.R. Aswatha. (2020). FPGA-based design and architecture of network-on-chip router for efficient data propagation. IIOAB Journal, 11(S2), 7–25.
38. Krutthika H. K. & A.R. Aswatha (2020). Design of efficient FSM-based 3D network-on-chip architecture. International Journal of Engineering Trends and Technology, 68(10), 67–73. https://doi.org/10.14445/22315381/IJETT-V68I10P212
39. Krutthika H. K. & Rajashekhara R. (2019). Network-on-chip: A survey on router design and algorithms. International Journal of Recent Technology and Engineering, 7(6), 1687–1691. https://doi.org/10.35940/ijrte.F2131.037619 (53 citations) (Now it is 17)
40. S. Ajay, et al., & Krutthika H. K. (2018). Source hotspot management in a mesh network-on-chip. 22nd International Symposium on VLSI Design and Test (VDAT-2018). https://doi.org/10.1007/978-981-13-5950-7_51