Enhancing Heart Disease Prediction Performance with a Soft Voting Ensemble and Additional Techniques

A DISSERTATION

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE AWARD OF THE DEGREE

of

MASTER OF TECHNOLOGY

IN

INFORMATION SYSTEMS

Submitted by

KAUSTAV SEN

2K21/ISY/12

Under the supervision of

DR. BINDU VERMA

Assistant Professor

Department of Information Technology

DEPARTMENT OF INFORMATION TECHNOLOGY

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi college of engineering)

Bawana road, delhi-110042

MAY, 2023

DEPARTMENT OF INFORMATION TECHNOLOGY

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi

CANDIDATE'S DECLARATION

I, Kaustav Sen, with the roll number **2K21/ISY/12**, **pursuing M.Tech in Information Systems**, hereby solemnly affirm that the project dissertation titled **"Enhancing Heart Disease Prediction Performance with a Soft Voting Ensemble and Additional Techniques,"** presented to the **Department of Information Technology, Delhi Technological University, Delhi**, is a product of my original research work. It has not been derived or copied from any other source without appropriate acknowledgement. This work has not been previously submitted, in whole or in part, for the attainment of any Degree, Diploma, Associateship, Fellowship, or similar titles or recognitions.

Place: Delhi NAME: **KAUSTAV SEN**

Date:

DEPARTMENT OF INFORMATION TECHNOLOGY

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi

CERTIFICATE

I hereby affirm that the Project Dissertation, titled **"Enhancing Heart Disease Prediction Performance with a Soft Voting Ensemble and Additional Techniques"**, submitted by **Kaustav Sen, Roll No 2K21/ISY/12** from the Information Technology Department, Delhi Technological University, Delhi, is a testament of the project work undertaken by the student under my guidance. It fulfills the partial requirements for the award of the Master of Technology degree. To the best of my understanding, this work, either in whole or in part, has not been submitted for any Degree or Diploma at this University or any other institution.

Place: Delhi

Date:

Dr. Bindu Verma

Assistant Professor

DEPARTMENT OF INFORMATION TECHNOLOGY

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi

ACKNOWLEDGEMENT

I extend my profound gratitude to **Dr. Bindu Verma**, Assistant Professor, IT Department at Delhi Technological University, who has been my guiding light throughout this project, offering her invaluable insights, feedback, and recommendations.

I recognize that this project's success would not have been achievable without the collective effort of everyone who, either directly or indirectly, contributed to its progression. My heartfelt appreciation goes to my parents, to **Prof. Dinesh Kumar Vishwakarma**, HOD (Department of Information and Technology) and the faculty members of the Department of Information Technology at Delhi Technological University for their unwavering support, cooperation, and encouragement, which were instrumental in the successful completion of this report. It is my sincere hope that this project achieves its intended purpose to the maximum extent possible.

Kaustav Sen

ABSTRACT

Heart disease is a leading cause of mortality, affecting a significant number of people worldwide. The pressing demand for diagnostic methods that deliver both superior effectiveness and accuracy is evident. Machine learning techniques, including deep tabular learning models, have been extensively applied to tabular healthcare data, demonstrating promising results in prediction and analysis. However, traditional machine learning models may suffer from limitations in accuracy, precision, and recall values. We propose a Soft voting meta-classifier comprising Catboost, Light-Gradient Boosting Machine, Gaussian Naive Bayes, Random Forest, and XGBoost to address these issues. Additionally, we explored deep tabular learning models TabNet and TabPFN. Our study was conducted on a fused dataset from UCI heart disease and Statlog sources. The proposed soft voting ensemble outperformed the individual models and achieved an accuracy of 91.85% and an AUC score of 0.9344, showcasing its potential for effective heart disease prediction.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

List of Symbols, Abbreviations and Nomenclature

SVM – Support Vector Machine

PCA – Principal Component Analysis

MLP – Multi Layer Perceptron

LR – Logistic Regression

LGBM- Light Gradient Boosting Machine

ROC - Receiver operating characteristic

AUC - Area under the ROC Curve

CHAPTER 1

INTRODUCTION

Heart disease is a group of illnesses that impact heart function; it is a subset of cardiovascular diseases (CVD). According to WHO, CVD deaths accounted for 32% of all deaths in 2019, with 17.9 million deaths [1]. Heart disease is expected to increase due to population expansion and ageing since the number of older adults is expected to double in many parts of the world [2]. Hazardous lifestyle choices, such as drug or alcohol misuse, high cholesterol levels, obesity, increased triglyceride levels, hypertension, and other health problems, enhance an individual's heart disease [3]. It is critical to create heart disease prediction methods that can help medical practitioners to diagnose the disease. Patients use of smart bands and other forms of intelligent health wearable connected to the Internet of Things (IoT) has become an integral part of modern medical care [4], [5]. Researchers and medical professionals all over the world have machine learning to increase the accuracy of predicting heart disease [6], [7], [8] . Deep learning and machine learning are commonly employed to detect health-related concerns [9]. Deep learning models have been created using medical data using pictures such as x-rays, chest scans [10]. A range of machine learning models, including CatBoost (CB), Decision Tree, Logistic Regression, Gaussian Naive Bayes classifier (GNB), and more, can be utilized to diagnose, categorize, or predict heart conditions.

The paper proposed a voting ensemble that uses soft voting, and the five algorithms used are selected due to their performance as individual models in predicting heart disease. Gaussian Naive Bayes, XGBoost, Catboost, LGBM and Random forests are used to create the proposed model. An AUC of 0.9351 is found for the proposed soft voting ensemble model, and its accuracy is 0.9185. A combined Heart Dataset set consisting of the UCI (University of California Irvine) Heart disease dataset and the UCI Statlog dataset is used for the experiment. The reason for developing a soft voting ensemble rather than individual models is that individual modes have higher False Negative and False Positive values. When predicting heart disease, a lower false negative value is preferred. Our proposed model minimized the false negative values compared to conventional machine learning models.

The dissertation is organized this way: Literature work is discussed in Chapter 2. Chapter 3 describes the dataset description, feature importance, pre-processing, and proposed model. In Chapter 4 the detailed result analysis of the proposed soft voting ensemble and TabNet is discussed. Finally, in Chapter 5 the conclusion of the work is done.

CHAPTER 2

RELATED WORKS

Heart disease prediction has received much attention, with many studies focusing on the UCI Cleveland dataset. On the combined UCI heart disease dataset, the UCI Irvine Statlog dataset, and other heart-related datasets like the Framingham heart study dataset, many researchers have studied the results of standard machine learning classifiers and ensemble techniques.

2.1 Traditional Methods Review

In the work of Garate-Escamila et al. [11], PCA and Chi-Squared is used and experiments are conducted on Cleveland and Hungarian dataset from the UCI heart disease datasets. Compared to the results of ML models on raw datasets, using chi-square for feature selection and PCA to form new features significantly improves performance. In their review paper, Ramalingam et al. [12] highlight the strong performance of Random Forest, ensemble-based methods, and SVM in predicting heart diseaseOn the Statlog heart disease dataset, Raza [12] discovered that a majority voting ensemble model consisting of logistic regression, nave Bayes, and multi-layer perceptron achieved 88% accuracy for Statlog dataset.

Kim et al. [13] compared three oversampling techniques for the imbalanced dataset. Compared to Adasyn and the Borderline SMOTE oversampling technique, they found SMOTE to be the most promising. Ishaq et al. [14] used 299 samples from the UCI heart failure clinical records dataset. With an extra tree classifier model, they achieved 92.62% accuracy and 0.93 F1- score using random forest ranking for feature selection and SMOTE for oversampling. Kavitha et al. [15] used the UCI Cleveland dataset to create a combination model of random forest and decision tree with an accuracy of 88%. DBSCAN was used by Fitriyani et al. [16] to eliminate outliers in a heart disease prediction model. They used the SMOTE-ENN oversampling technique and XGBoost to balance the data to predict heart disease. The Statlog dataset had a 95.90% accuracy rate, while the UCI Cleveland dataset had a 98.40% accuracy rate. Wang et al. [17] developed a stacking-based model that achieved 95.43% accuracy and 95.84% recall and 94.44% specificity for coronary heart disease detection. Using data from a hospital in Beijing, China, Chang et al. [18] created a hybrid XGBSVM model to predict cardiovascular disease in hypertensive patients. While other machine learning models struggled on their dataset, XGBSVM performed admirably. Liu et al. [19] utilized a stacking-based model with five base learners (Extra Tree Classifier, Logistic Regression, MLP, CatBoost, and Random Forest) for heart disease prediction on a dataset of 918 samples. The final estimator was Logistic Regression. This approach yielded an accuracy rate of 89.86%, demonstrationg the effectiveness of stacking-based ensemble models in this domain.

In their paper, Farman Ali et al. [20] proposed an intelligent healthcare system for heart disease using feature fusion and ensemble deep learning techniques. Extracted features from a sensor and the electronic medical record are combined using feature fusion. Then the information gain technique eliminates redundant features, and conditional probability is used to assign feature weight for each class. The data is then trained on an ensemble deep-learning model. Amarbayasgalan et al. [21] use PCA to split the training dataset into highly biased and highly regular subgroups for coronary heart disease prediction. Variational auto-encoders are used to enrich the highly biased group. Separate deep neural networks are used to train the two groups. Mienye et al. [22] for heart disease prediction, used an enhanced stacked sparse auto-encoder network (SSAE) which consisted of multiple sparse encoders and a Softmax classifier for feature learning. The algorithms parameters were optimized using the particle swarm optimization (PSO) based technique. Ghosh et al. [23] conducted a study using a heart disease dataset comprising 918 observations. In their comparison of five classification models, they employed 10-fold cross-validation. The Random Forest model emerged as the most accurate predictor, achieving a mean accuracy of 86.93%.

2.2 Deep Learning Methods Review

Researchers have started using tabular deep learning models in recent times for classification and regression tasks. The use of TabNet and other deep tabular deep learning models is minimal compared to traditional models used for heart disease prediction. Some of the tabular deep learning techniques is explored here.

In a study focused on predicting heart disease using the UCI Cleveland dataset (303 samples), Pillai [24] evaluated various models on unseen test data. TabNet outperformed the other models, achieving a 94.4% accuracy and a 0.94 AUC score. This performance was superior to the best baseline model, logistic regression, which had a 91.7% accuracy and a 0.90 ROC score. Moshawrab et al. [25] employed TabNet and TabTransformer for heart disease prediction on the SHAREE-DB dataset, consisting of 139 samples. TabNet achieved a 76% accuracy and a 0.7650 F1-score, while TabTransformer obtained a 90.38% accuracy and a 0.9545 F1-score, surpassing the SVM model. Wang et al. [26] utilized an improved TabNet model for hyperspectral estimation of soil copper concentration. In their paper[27] , Prabowo et al. introduced the MLP-LSTM, a dual-input deep learning model processing both time-series and tabular data for student GPA prediction. Despite outperforming other models in metrics like MSE, MAE, and R^2 score, it necessitates a smooth target GPA distribution. The study also identified persistent long-range dependencies issues despite LSTM use, suggesting transformers as a potential solution for future work. Borisov et.al [28] introduced the GReaT method that utilizes a generative LLM with self-attention to create synthetic tabular datasets, supporting both discrete and numerical features. The procedure involves transforming feature vectors into text, randomizing feature order, and using the text for transformer finetuning. In experiments, classifiers trained on GReaT synthetic data outperformed those trained on other synthetic data. Despite similarities, GReaT does not copy the training data, and while synthetic data can be distinguished from original data, it is more challenging to differentiate than with other synthetic data generation methods. Teresa et al. [29] proposed a novel deep learning approach that combines a Sparse Autoencoder (SAE) for feature augmentation and a Convolutional Neural Network (CNN) for classification tasks. Applied to a dataset of 918 patient records, this method achieved an accuracy of 90.088%, outperforming traditional classifiers and other state-of-the-art methods.

2.3 Research Gap

The arena of heart disease prediction has seen extensive application of various machine learning and deep learning models, ranging from standalone models, ensemble-based approaches, oversampling techniques, and even hybrid models. Although these techniques have demonstrated effective results, there remains an exploration gap in the application of ensemble techniques, specifically soft voting mechanisms, which could potentially enhance model robustness and generalizability. Stand-alone models, while efficient, may lack the diversity of predictions that an ensemble of models can provide. Oversampling techniques address the issue of class imbalance but do not necessarily enhance model performance. Stacking based approach makes final decision based on the final estimator limits the model's ability to capture the breadth of insights a diverse ensemble of models can offer.

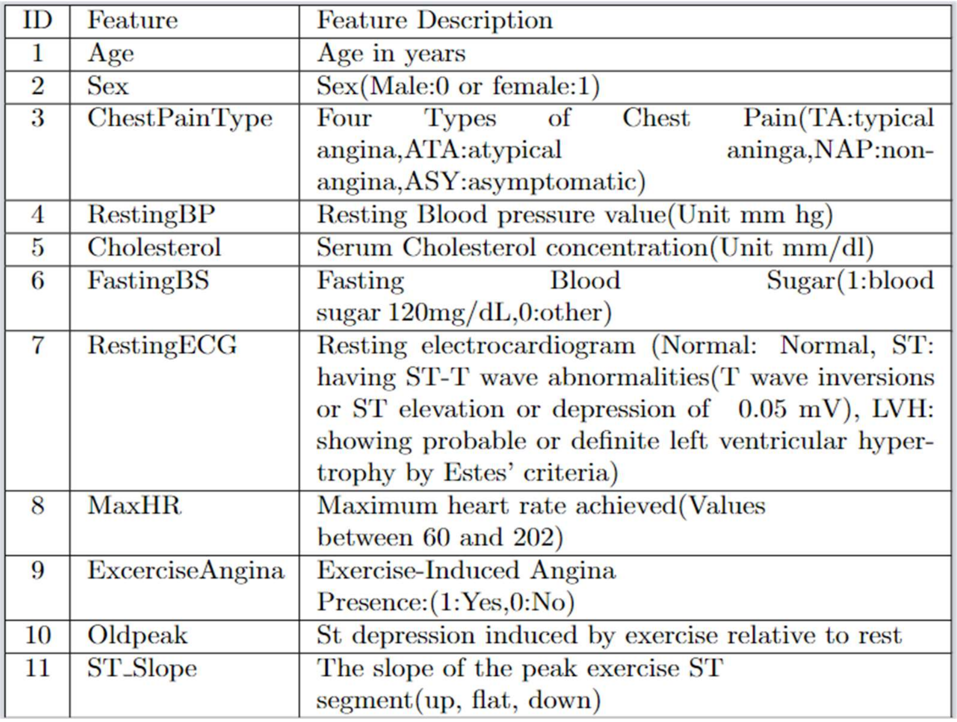
In my research, I attempt to address this gap by proposing a soft voting ensemble model, incorporating the capabilities of multiple classifiers including Random Forest, CatBoost, XGBoost, Light Gradient Boosting Machine, and Gaussian Naive Bayes. This ensemble model, although not outperforming all existing models, shows competitive performance, with an accuracy of 91.85%. This research, therefore, contributes a novel perspective by demonstrating the potential of a soft voting ensemble approach, inviting further investigations and improvements in this direction. It underscores the importance of diversifying prediction methodologies in heart disease prediction, ultimately driving the field towards enhanced prediction models.

CHAPTER 3

METHODOLOGY

3.1 Dataset Description

UCI's heart disease dataset and Statlog dataset were used for this research, and the features employed here were derived from both of those sources. The number of samples from each of the datasets that are used is Cleveland (303), Hungarian (294), Switzerland (123), Long Beach, VA (200), and Statlog (Heart) Data Set (270). Out of which 272 are duplicated rows, therefore the total number of samples is 918. There are 11 features and one binary output variable named “heart disease". Figure 3.1 provides a comprehensive depiction of the characteristics in detail. The output variable has 508 samples with heart disease and 410 samples that do not have heart disease. Figure 3.2 shows the heat map of features and the binary output.



ID	Feature	Feature Description
1	Age	Age in years
2	Sex	Sex(Male:0 or female:1)
3	ChestPainType	Four Types of Chest Pain(TA:typical angina,ATA:atypical aninga,NAP:non-angina,ASY:asymptomatic)
4	RestingBP	Resting Blood pressure value(Unit mm hg)
5	Cholesterol	Serum Cholesterol concentration(Unit mm/dl)
6	FastingBS	Fasting Blood Sugar(1:blood sugar 120mg/dL,0:other)
7	RestingECG	Resting electrocardiogram (Normal: Normal, ST: having ST-T wave abnormalities(T wave inversions or ST elevation or depression of 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes’ criteria)
8	MaxHR	Maximum heart rate achieved(Values between 60 and 202)
9	ExcerciseAngina	Exercise-Induced Angina Presence:(1:Yes,0:No)
10	Oldpeak	St depression induced by exercise relative to rest
11	ST_Slope	The slope of the peak exercise ST segment(up, flat, down)

Figure 3.1: Dataset Features Description

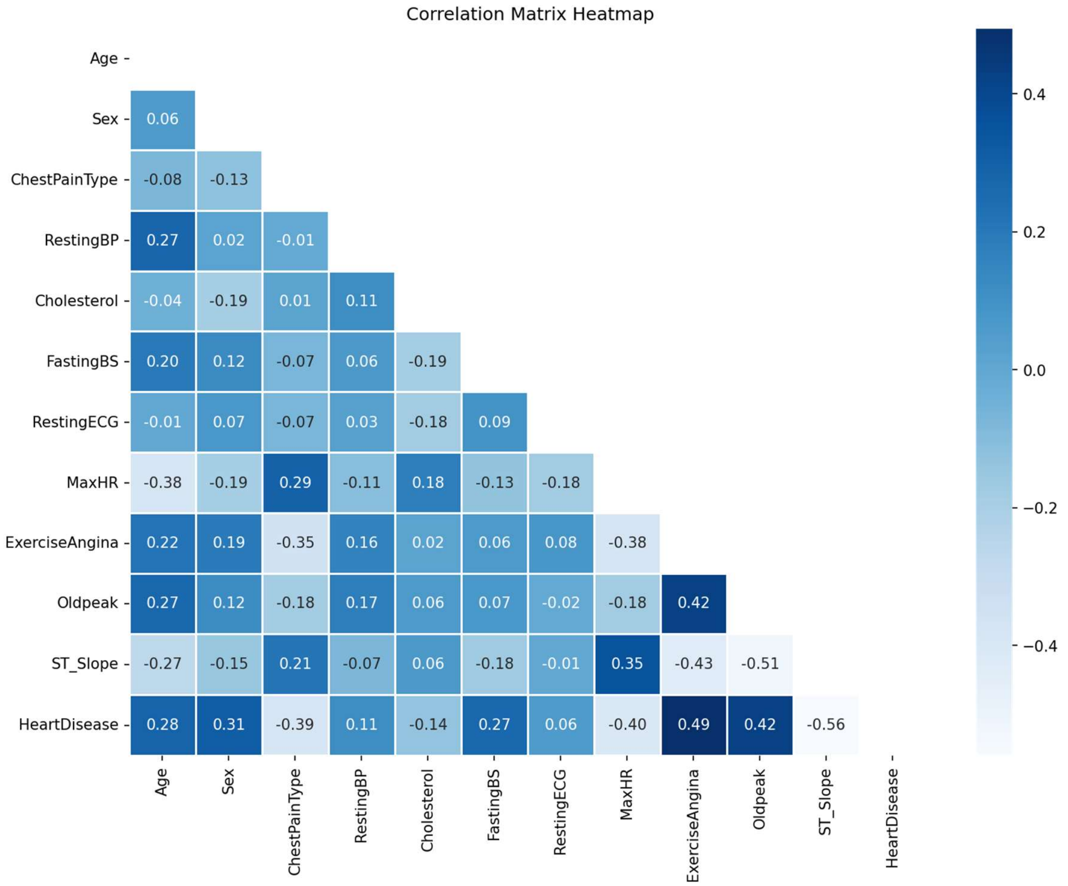


Figure 3.2: Heatmap of the Features

3.2 Feature Importance According To XGBoost

Understanding the key attributes of the dataset is vital in accurately predicting heart disease. We utilized the XGBoost algorithm, well-known for its feature ranking capabilities to visualize the importance of the features as show in Figure 3.3. As per XGBoost's analysis, ST Slope emerged as the most influential feature, followed in importance by ChestPainType, ExerciseAngina. On the other hand, the algorithm identified Sex, Oldpeak, FastingBS, Cholesterol, RestingECG, MaxHr, Age, RestingBP as the least significant features.

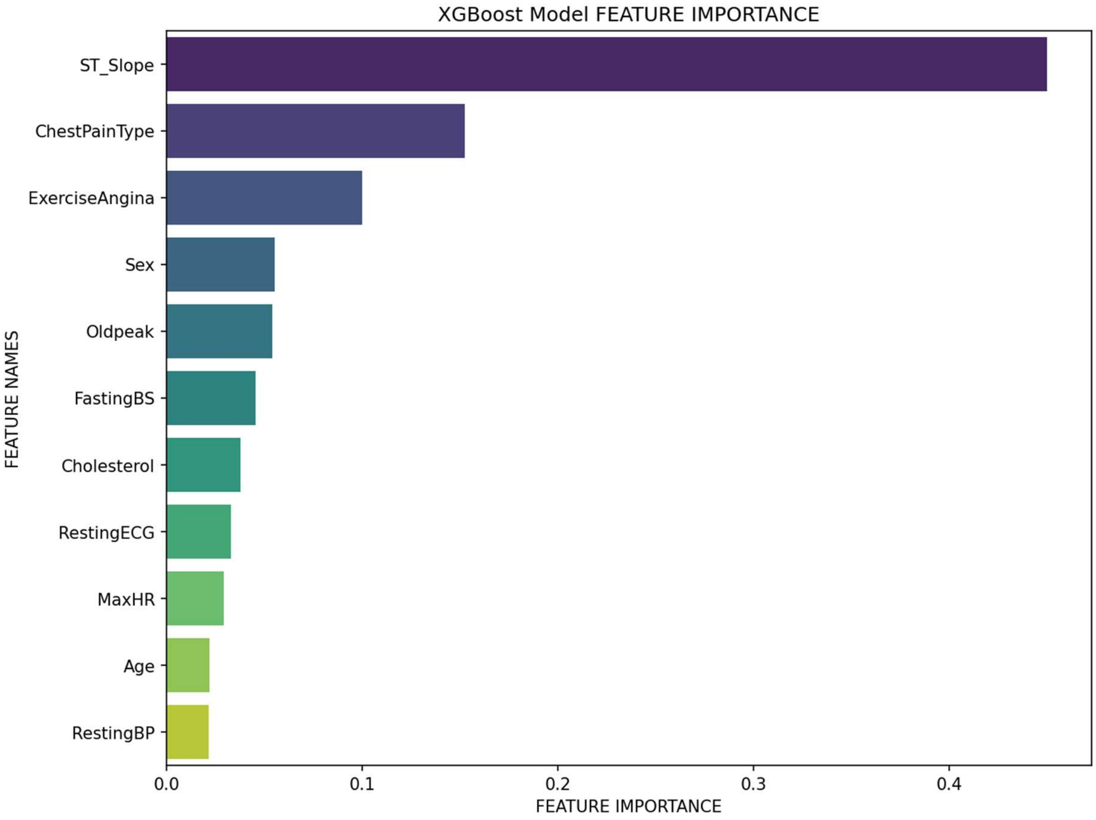


Figure 3.3: Feature Importance Plot According to XGBoost

3.3 Preprocessing Steps For Soft Voting

Data pre-processing is a crucial step before feeding data to the model, as it helps transform raw data into a valuable and efficient format for analysis. The dataset contained five categorical features, with the remaining features being numerical. Label Encoding was employed to convert the absolute values of those categorical features into numerical data. The numerical data was then scaled to a standard range, for the models that are sensitive to feature scaling like SVM and Logistic Regression.

Notably, tree-based models, being naturally immune to feature scaling, were not scaled. These models are inherently invariant to feature scaling. Thus, applying scaling to the data would have no significant impact on the performance of these tree-based models.

3.4 Preprocessing Steps For TabNet

Data pre-processing is required before feeding data to the model to transform data into a valuable and efficient format. The dataset contained five categorical features, with the remaining features being numerical. The dataset was partitioned into training (80%) and testing (20%) subsets using stratified sampling to ensure the same class distribution in both subsets.

The TabNet was prepared for this dataset in the following way:

- Numerical features were standardized using the StandardScaler from scikit-learn. The scaler was adjusted to the training set and was then implemented to both the training and testing sets.
- For each categorical feature, a separate LabelEncoder was used to transform the feature into integer-encoded values. ensuring a suitable representation of these variables in both the training and testing sets.
- A custom PyTorch module was developed to create embeddings for categorical features, initializing an embedding layer for each unique category. The module's forward method combines individual embeddings into a single tensor. The model, initialized with the number of unique categories per feature, generates embeddings for training and testing sets after converting the encoded categorical features into PyTorch tensors with a long data type.
- The standardized numerical features and the computed categorical feature embeddings were horizontally stacked to create the final preprocessed training and testing sets, which are then used as input for TabNet .

The reason for using embedding layer for categorical features is that the embedding layer can capture more complex relationships between categorical features and target variable. Representing the categorical features in a continuous space, the model can better generalize to unseen data.

3.5 Methods Used For Soft Voting Ensemble

XGBoost: It is another ensemble machine learning algorithm that uses decision trees as weak learners based on a gradient boosting framework. It performs very well on structured data. XGBoost employs a specific type of decision tree known as CART (Classification and Regression Trees) in its operation. The leaf nodes contain real-value scores of each instance belong to a class label. When the tree reaches its max depth, decisions are made by converting this score into categories using the threshold. The general formula of XGboost can be formed as:

$$\widehat{\mathbf{y}}_{\mathbf{i}} = \sum_{\mathbf{t=1}}^{\mathbf{T}} \mathbf{g}_{\mathbf{t}}(\mathbf{x}_{\mathbf{i}}) \, , \mathbf{g}_{\mathbf{t}} \in \mathbf{G} \tag{3.1}$$

g is the functional space of G T is the number of Trees, and G is the set of possible CARTS.

Light Gradient Boosting Machine: It is a fast open-source gradient boosting derived framework. It's tree grows vertically compared to other algorithms whose level of use grows horizontally. Generally, it is used for larger datasets. LGBM removes instances with low gradients and keeps instances with more significant gradients.

Given the Supervised training set $\boldsymbol{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{m}$. LGBM attempts to find estimates

$\boldsymbol{g}(\boldsymbol{a})$ to a specific function $\boldsymbol{g}^*(\boldsymbol{a})$ that restricts the normal estimation of a given loss function $\boldsymbol{F}(\boldsymbol{c}, \boldsymbol{g}(\boldsymbol{a}))$ as shown in [30]:

$$\boldsymbol{g} = \boldsymbol{argmin}_{\boldsymbol{g}} \boldsymbol{Q}_{\boldsymbol{c},\boldsymbol{D}} \boldsymbol{F}(\boldsymbol{c}, \boldsymbol{g}(\boldsymbol{a})) \tag{3.2}$$

Random Forest: A random forest comprises numerous decision trees, each utilizing a data subset from the original dataset with replacement. After training the individual decision trees, a random forest picks the class that gets the maximum prediction vote from each tree. It helps random forests to overcome the problem of over-fitting. It uses the bagging concept of ensemble learning, which combines many weak classifiers to provide a solution. Mathematically it can be expressed as [6]:

For a given Data $\boldsymbol{A} = \{\boldsymbol{a_1}, \boldsymbol{a_2}, \boldsymbol{a_3}, \ldots \ldots \boldsymbol{a_n}\}$ with responses $\boldsymbol{O} = \{\boldsymbol{a_1}, \boldsymbol{a_2}, \boldsymbol{a_3}, \ldots \ldots \boldsymbol{a_n}\}$ the bagging is repeated from $\boldsymbol{b = 1 \quad to \quad B}$. By averaging the predictions $\sum_{b=1}^{B} \boldsymbol{fb(a')}$ for every individual trees on a' the unseen samples a' is made:

$$\boldsymbol{K = \frac{1}{B} \sum_{b=1}^{B} fb(a')} \tag{3.3}$$

Gaussian Naive Bayes: Gaussian Naive Bayes is a classification method that applies the Bayes theorem with an assumption of conditional independence among features given the class label.Gaussian Naive Bayes is a continuous data model that models the likelihood of features given a class label using a Gaussian (normal) distribution. Despite its simplicity and strict independence assumptions, the method frequently performs unexpectedly well in practice, offering good classification accuracy and scalability. One disadvantage is its sensitivity to irrelevant traits, which might have a negative impact on its performance. Gaussian Naive Bayes is represented mathematically as follows:

$$\mathbf{P(z|x_1, x_2, \ldots, x_n) = \frac{P(z) \prod_{i=1}^{n} P(x_i|z)}{P(x_1, x_2, \ldots, x_n)}} \tag{3.4}$$

where $\mathbf{P(z)}$ is the prior probability of class $\boldsymbol{z}$, $\mathbf{P(x_i|z)}$ is the likelihood of feature $\mathbf{x_i}$ given class $\boldsymbol{z}$ (modeled using Gaussian distribution), and $\mathbf{P(x_1, x_2, \ldots, x_n)}$ is the evidence.

CatBoost: It is one of the algorithms based on gradient decision trees. It does permutation-based boosting compared to the classic boosting algorithm. Gradient boosting is often overfitted on small datasets. CatBoost is used to solve the overfitting problem on small datasets. Its general equation can be formed as [31]:

$$\boldsymbol{G_p(x) = G(p-1)(x) + \gamma_p K_p(x)} \tag{3.5}$$

$$\boldsymbol{\gamma_p = argmin \textstyle\sum_{c=1}^{c} Y\left(b_c, G(p-1)(x_c) + \gamma_p K_p(a_c)\right)} \tag{3.6}$$

$\boldsymbol{G_p(x)}$ represents the final outcome, the loss function is denoted by **Y**, and $\boldsymbol{K_p(x)}$ is the pseudo-residuals where the count of iterations is $\boldsymbol{p}$, and $\boldsymbol{\gamma_p}$ is the multiplier in Equation 5 and 6.

3.6 Soft Voting Ensemble

A voting ensemble employs two voting methods: majority voting and soft voting. Majority voting establishes the final classification using the mode of all base classifier predictions. Soft voting computes the average of the predicted probabilities for class labels from all individual models in the ensemble. It then assigns the final class label based on the highest average probability. Our proposed soft voting ensemble is a meta-model composed of CatBoost, LGBM, XGBoost, Random Forest, and Gaussian Naive Bayes models. We used equal weights for each classifier because all of the classifiers used in the soft voting ensemble have similar performances. Let's denote the probabilities predicted by each model as $\mathbf{P_{1,0}, P_{1,1}, P_{2,0}, P_{2,1} \dots . P_{n,0}, P_{n,1}}$. For binary classification, the probability for each class label (0 or 1) is computed by each model. The class label is determined by the soft voting ensemble by averaging probabilities and selecting the one with the highest average.The flowchart of Soft Voting Ensemble for our work is shown in Figure 3.4.

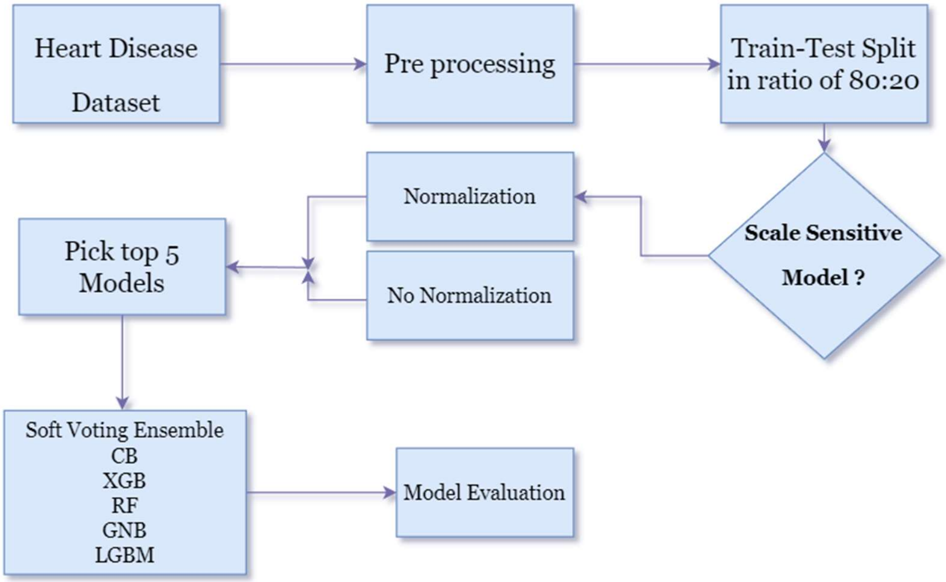


Figure 3.4 Flowchart of Proposed Soft Voting Ensemble

Let $P_{i,0}$ $P_{i,1}$ represent the predicted probabilities of class 0 and class 1, respectively, for the i-th model. The soft voting ensemble probabilities for each class can be calculated as:

$$\boldsymbol{P_{ensemble,0} = \frac{1}{n}\sum_{i=1}^{n} P_{i,0}} \tag{3.7}$$

$$\boldsymbol{P_{ensemble,1} = \frac{1}{n}\sum_{i=1}^{n} P_{i,1}} \tag{3.8}$$

After calculating the ensemble probabilities, the class label with the highest ensemble probability is chosen as the final prediction:

$$\boldsymbol{\hat{y}_{ensemble} = \begin{cases} 0 \;\; if\ P_{ensemble,0} > P_{ensemble,1} \\ 1 \;\; if\ P_{ensemble,1} \geq P_{ensemble,0} \end{cases}} \tag{3.9}$$

3.7 TabNet

TabNet is a deep learning framework that was originally introduced in the paper "TabNet: Attentive Interpretable Tabular Learning" by Sercan O. Arik and Tomas Pfister in 2019 [32] and is designed particularly for handling tabular data. It offers an effective method to capture complex patterns in tabular data. The model is particularly well-suited for classification and regression tasks, as it employs an encoder architecture with sequential multi-step processing, feature selection, and attention mechanisms to create interpretable and sparse representations of the input data. Additionally, TabNet can be adapted for self-supervised learning tasks, such as predicting missing feature columns, by incorporating a decoder architecture to reconstruct tabular features from TabNet-encoded representations.The detailed architecture of TabNet encoder and Attentive Transformer is shown in Figure 3.5 and 3.6 respectively [32].

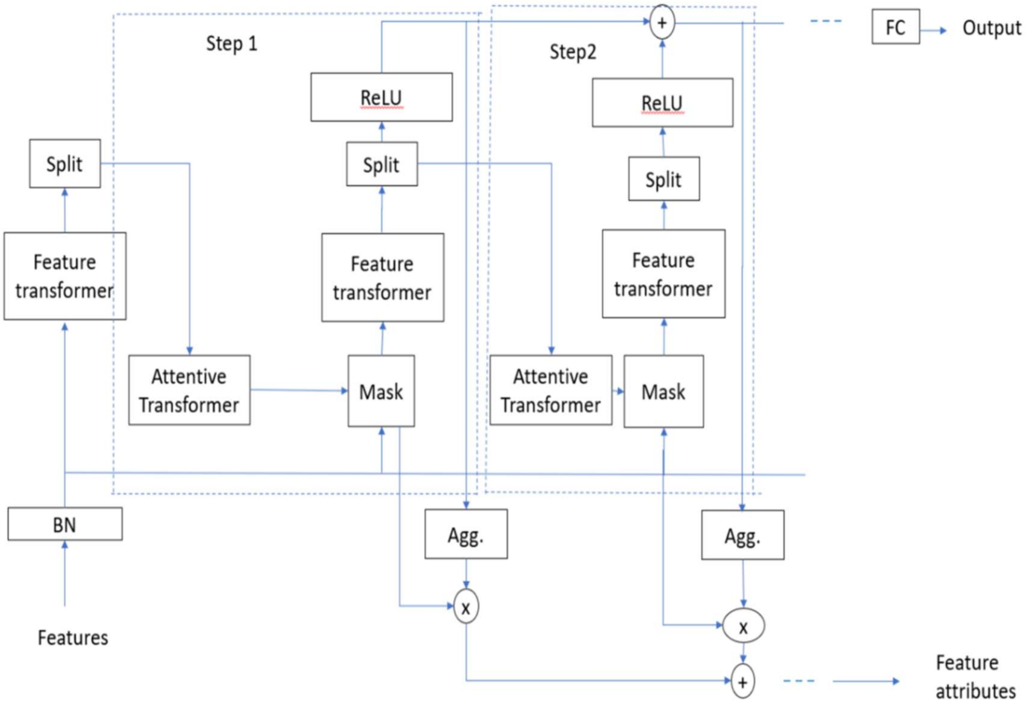


Figure 3.5: TabNet Encoder Architecture

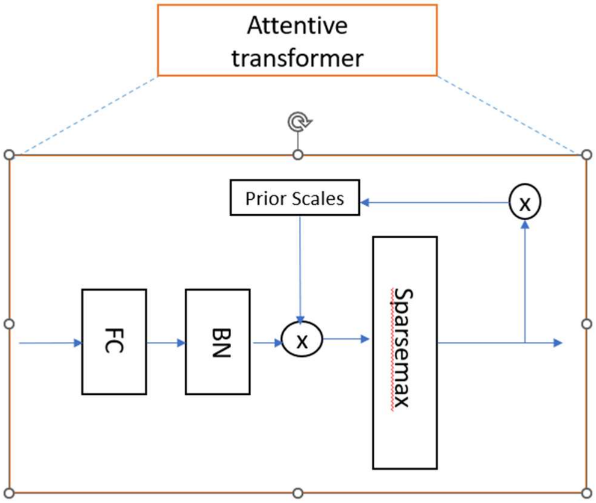


Figure 3.6: Attentive Transformer of TabNet

During preprocessing, TabNet applies batch normalization to raw numerical features, while categorical features using mapped using trainable embeddings. The resulting X-dimensional input features are then passed to each decision step. The encoder processes these features through N decision steps, which involve attentive transformers, feature selection, and feature processing. At each decision step, the attentive transformer computes a mask (G[i]) for soft selection of important features based on information from the previous step (a[i-1]). This mask is applied to the input features (G[i] * f), resulting in a sparse selection of the most relevant features. These features are then processed through a feature transformer, generating an output that is split into a decision step output (d[i]) and information for the subsequent step (a[i]).

The overall decision embedding (dout) is constructed by aggregating the outputs of all decision steps, using the ReLU activation function. In the final stage of prediction, a linear mapping is applied to the aggregated decision embedding (Wfinal * dout) to produce the output. For binary classification tasks, softmax is applied during training to calculate the probability of each class, and argmax is used during inference to determine the most probable class.TabNet's encoder architecture, interpretability, and ability to handle a diverse range of feature types make it an invaluable tool for binary classification tasks in various application domains.The overall Flowchart for TabNet model for the work is shown in Figure 3.7. The attention mechanism of the model enables it to concentrate on the most significant features at each decision-making step, providing improved performance and interpretability compared to traditional deep learning models.

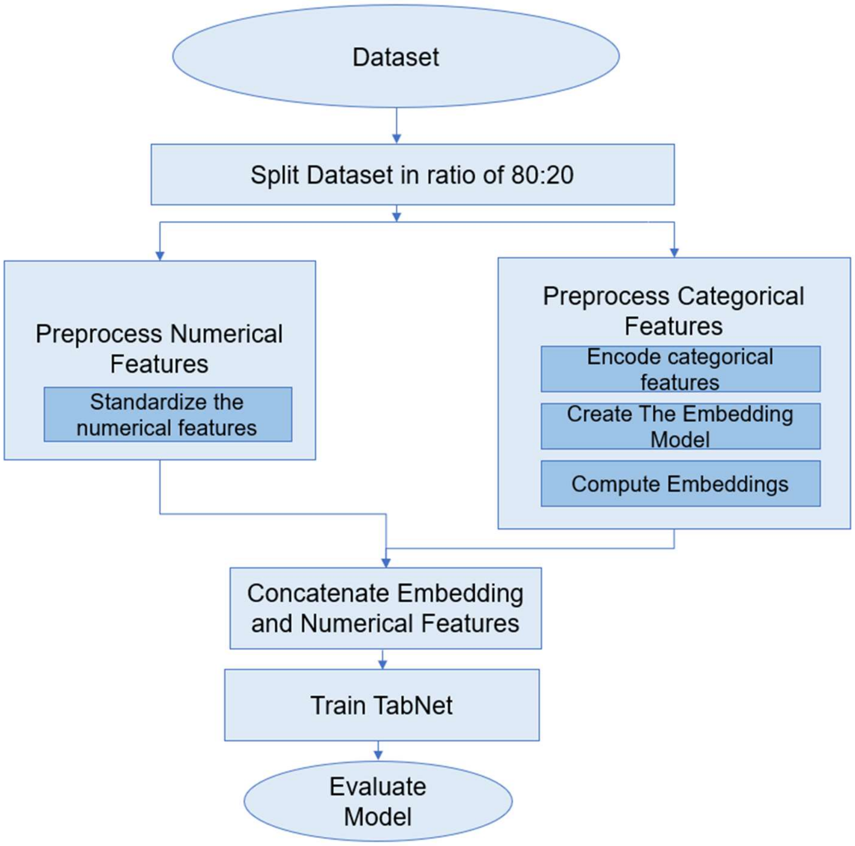
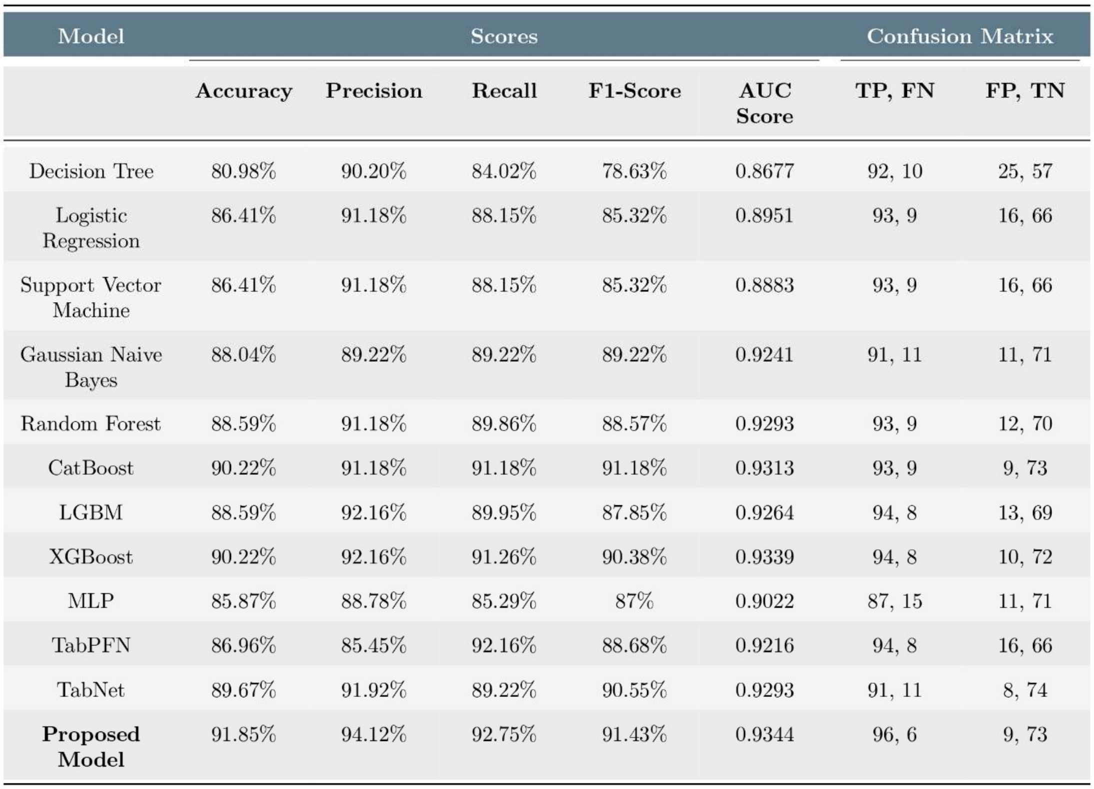

Dataset
Split Dataset in ratio of 80:20
Preprocess Numerical Features
Standardize the numerical features
Preprocess Categorical Features
Encode categorical features
Create The Embedding Model
Compute Embeddings
Concatenate Embedding and Numerical Features
Train TabNet
Evaluate Model

Figure 3.7: FlowChart For TabNet Model

CHAPTER 4

EXPERIMENTAL RESULTS

For creating our proposed soft voting ensemble out of all the models, we picked the top-performing models: Catboost, GNB, LGBM, Random Forest, and XGBoost. Hyperparameter tuning was done for all of these models for was done using GridSearchCV technique from scikit-learn. Our proposed model achieved 91.85% accuracy, 91.43% f1-score, 94.12% precision, 92.75% recall, and the AUC score was 0.9344. In terms of accuracy, the ensemble model, with a score of 0.9185, outperforms the best individual classifier, XGBClassifier (0.9022), as well as other models. The detailed results of all the models are shown in Figure 4.1. Figure 4.2 presents the Receiver Operating Characteristic (ROC) curve for the proposed Soft Voting Ensemble model.



Model	Scores					Confusion Matrix	
	Accuracy	Precision	Recall	F1-Score	AUC Score	TP, FN	FP, TN
Decision Tree	80.98%	90.20%	84.02%	78.63%	0.8677	92, 10	25, 57
Logistic Regression	86.41%	91.18%	88.15%	85.32%	0.8951	93, 9	16, 66
Support Vector Machine	86.41%	91.18%	88.15%	85.32%	0.8883	93, 9	16, 66
Gaussian Naive Bayes	88.04%	89.22%	89.22%	89.22%	0.9241	91, 11	11, 71
Random Forest	88.59%	91.18%	89.86%	88.57%	0.9293	93, 9	12, 70
CatBoost	90.22%	91.18%	91.18%	91.18%	0.9313	93, 9	9, 73
LGBM	88.59%	92.16%	89.95%	87.85%	0.9264	94, 8	13, 69
XGBoost	90.22%	92.16%	91.26%	90.38%	0.9339	94, 8	10, 72
MLP	85.87%	88.78%	85.29%	87%	0.9022	87, 15	11, 71
TabPFN	86.96%	85.45%	92.16%	88.68%	0.9216	94, 8	16, 66
TabNet	89.67%	91.92%	89.22%	90.55%	0.9293	91, 11	8, 74
Proposed Model	91.85%	94.12%	92.75%	91.43%	0.9344	96, 6	9, 73

Figure 4.1: Results of Proposed Model and other Models

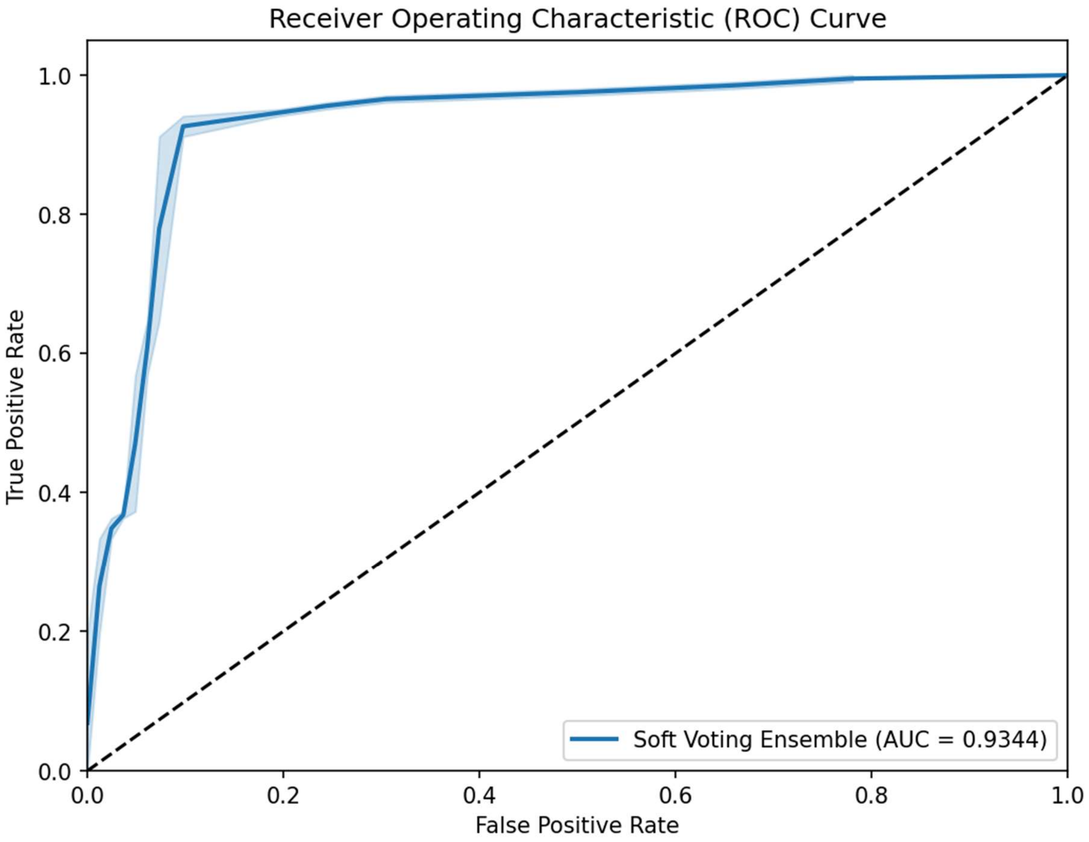
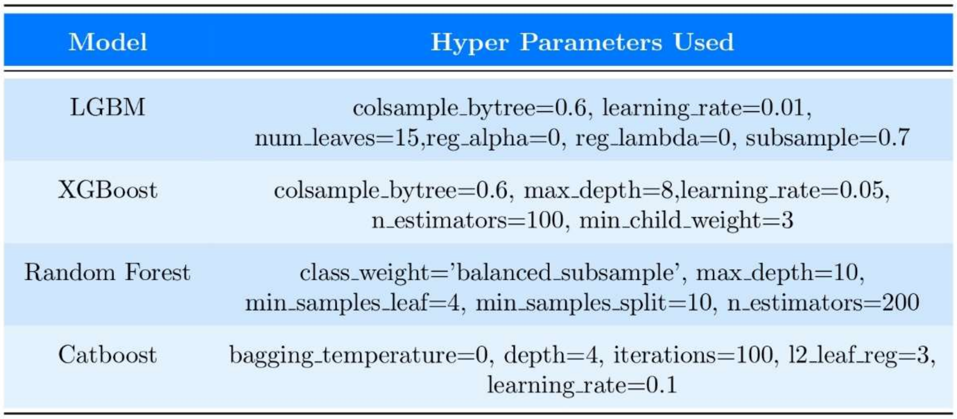


Figure 4.2: Roc Plot of Proposed Model



Model	Hyper Parameters Used
LGBM	colsample_bytree=0.6, learning_rate=0.01, num_leaves=15,reg_alpha=0, reg_lambda=0, subsample=0.7
XGBoost	colsample_bytree=0.6, max_depth=8,learning_rate=0.05, n_estimators=100, min_child_weight=3
Random Forest	class_weight='balanced_subsample', max_depth=10, min_samples_leaf=4, min_samples_split=10, n_estimators=200
Catboost	bagging_temperature=0, depth=4, iterations=100, l2_leaf_reg=3, learning_rate=0.1

Figure 4.3: Hyperparameters Used for the Soft Voting Model

The ensemble model demonstrates superior performance across various metrics compared to individual classifiers and other models. With a precision of 94.12%, it effectively identifies true positive cases and reduces false positives. Its recall of 92.75% indicates a higher effectiveness in detecting positive cases, minimizing false negatives. The ensemble model achieves the highest F1-score of 0.9143, reflecting a balanced performance between precision and recall, ensuring neither overly conservative nor aggressive predictions. Finally, its AUC of 0.9344, slightly higher than the best individual classifier, XGBClassifier (0.9339), highlights its enhanced discrimination ability, making it more effective in distinguishing between positive and negative instances. The Hyperparameters used for Models in Soft Voting ensemble are shown in Figure 4.3.

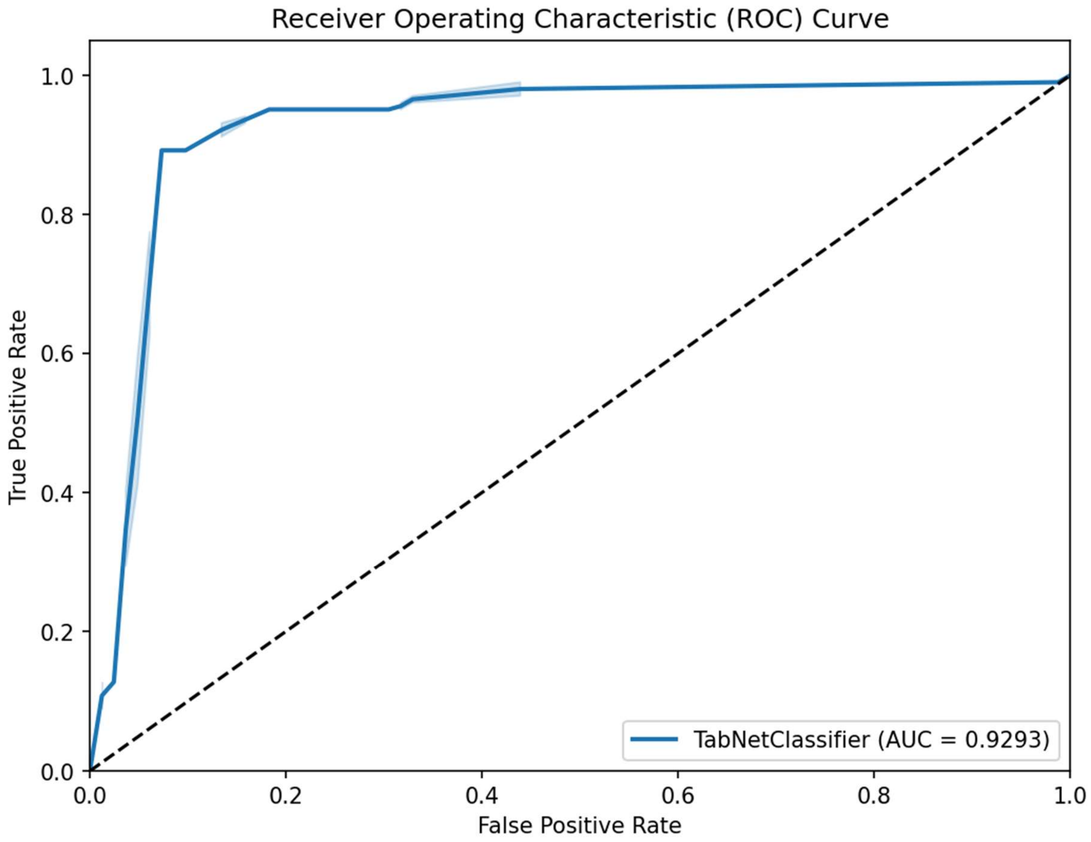


Figure 4.4: ROC Plot of TabNet

The results show that TabNet outperforms LGBM, a popular gradient boosting model known for its strong predictive capabilities. TabNet achieves a higher accuracy of 89.67%, F1-score of 90.55%, and AUC score of 0.9293, compared to LGBM's accuracy of 88.59%, F1-score of 87.85%, and AUC score of 0.9264. This demonstrates the potential of TabNet as a powerful alternative to traditional gradient boosting models when dealing with tabular data The Receiver operating characteristic curve of TabNet is shown in Figure 4.4 . TabPFN, on the other hand, is another deep tabular learning technique which is pretrained Transformer to solve small classification problems [33]. In our experiments, TabPFN achieves an accuracy of 86.96%, F1-score of 88.68%, and AUC score of 0.9216, which are competitive results compared to other models in the study. Overall, the superior performance of TabNet and TabPFN highlights the benefits of leveraging deep learning techniques specifically tailored for tabular data, leading to more accurate and efficient models for various predictive tasks.

The proposed ensemble had the lowest false positives and false negatives values compared to the other models used in the experiment, as shown in Figure 4.1. Table 4.1 provides a comparison of our proposed approach with existing state-of-the-art methods outlined in the literature. Compared to studies using the same dataset [19,23,29], my work has shown significant advancements. Liu et al. [19] reached 89.86% accuracy with a stacking-based model, while Ghosh et al.'s [23] best model, Random Forest, yielded 86.93% accuracy. Teresa et al. [29] combined a Sparse Autoencoder and a Convolutional Neural Network for 90.088% accuracy. The results discussion presented here highlights the key performance metrics of the ensemble model in comparison to the individual classifiers and other models. These performance enhancements over base classifiers can be attributed to the soft voting mechanism, which combines the complementary strengths of the individual classifiers, ultimately leading to a more accurate prediction of heart disease.

Table 4.1: Contrasting the outcomes with those achieved in the state-of-the-art studies

Study	**Model Used**	**Accuracy (%)**
Liu et al. (2022)	Stacking Based [19]	89.86
Teresa García-OrdásOrd et al. (2023)	CNN-SAE [29]	90.09
Ghosh et.al (2022)	Random Forest [23]	86.93
Proposed Work	Soft Voting Ensemble	91.85

CHAPTER 5

CONCLUSION AND FUTURE WORK

In conclusion, our soft voting ensemble model, which combines gradient-based boosting models (LightGBM, CatBoost, and XGBoost), Random Forest, and Gaussian Naive Bayes, has demonstrated superior performance in predicting heart disease. The improved metrics, including area under the curve (AUC), F1-score, and others, indicate the potential of this ensemble approach for reliable disease prediction.

The strength of our Soft Voting ensemble model lies in its ability to leverage the unique capabilities of each individual model. Gradient-based boosting models, such as LightGBM, CatBoost, and XGBoost, are well-suited to handling large datasets and can effectively model complex relationships in the data. Random Forest provides a robust approach, reducing overfitting by averaging multiple decision trees. Lastly, Gaussian Naive Bayes offers a simple yet effective probabilistic approach, complementing the other methods in the ensemble. The soft voting technique employed in our model contributes to its success by combining the probability estimates from each individual model. This approach allows the ensemble to capture a more nuanced representation of the underlying data, resulting in improved performance.

The results also highlight the promise of using novel deep learning architectures like TabNet and meta-learned algorithms such as TabPFN for tabular data problems. TabNet performed better than light gradient boosting machine further modification of the algorithm can lead to even better results in future.

For future research, several areas of exploration could further enhance the model's performance and utility:

- Feature engineering: Investigating additional features and feature transformations may lead to improved predictive power. Domain-specific knowledge could be leveraged to create new, meaningful features or to identify potential interactions among existing ones.
- Model diversity: Incorporating additional models, such as deep tabular learning techniques like TabTransformer, FTTransofrmer, Modification of TabNet and others, could be explored.

- Real-world validation: Finally, validating the model using real-world data from various sources, such as electronic health records or biobanks, will be crucial for assessing its generalizability and applicability in clinical settings.

By exploring these avenues, our soft voting ensemble model could be further refined and expanded, potentially offering an even more powerful and practical tool for predicting heart disease.

REFERENCES

[1] “Cardiovascular diseases.” https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1 (accessed Mar. 17, 2023).

[2] “Cardiovascular Disease Burden, Deaths Are Rising Around the World - American College of Cardiology.” https://www.acc.org/about-acc/press-releases/2020/12/09/18/30/cvd-burden-and-deaths-rising-around-the-world (accessed Jun. 11, 2022).

[3] “Heart disease - Symptoms and causes - Mayo Clinic.” https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118 (accessed Mar. 17, 2023).

[4] M. A. Khan and F. Algarni, “A Healthcare Monitoring System for the Diagnosis of Heart Disease in the IoMT Cloud Environment Using MSSO-ANFIS,” *IEEE Access*, vol. 8, pp. 122259–122269, 2020, doi: 10.1109/ACCESS.2020.3006424.

[5] Y.-L. Zheng *et al.*, “Unobtrusive Sensing and Wearable Devices for Health Informatics,” *IEEE Trans Biomed Eng*, vol. 61, no. 5, pp. 1538–1554, Jun. 2014, doi: 10.1109/TBME.2014.2309951.

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

[7] P. Ghosh *et al.*, “Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques,” *IEEE Access*, vol. 9, pp. 19304–19326, 2021, doi: 10.1109/ACCESS.2021.3053759.

[8] A. Gani, A. V Gribok, S. Rajaraman, W. K. Ward, and J. Reifman, “Predicting Subcutaneous Glucose Concentration in Humans: Data-Driven Glucose Modeling,” *IEEE Trans Biomed Eng*, vol. 56, no. 2, pp. 246–254, Jun. 2009, doi: 10.1109/TBME.2008.2005937.

[9] A. K. Sahoo, C. Pradhan, and H. Das, “Performance evaluation of different machine learning methods and deep-learning based convolutional neural network for health decision making,” *Studies in Computational Intelligence*, vol. SCI 871, pp. 201–212, 2020, doi: 10.1007/978-3-030-33820-6_8/COVER/.

[10] Y. Xu *et al.*, “Deep Learning Predicts Lung Cancer Treatment Response from Serial Medical Imaging,” *Clinical Cancer Research*, vol. 25, no. 11, pp. 3266–3275, Jun. 2019, doi: 10.1158/1078-0432.CCR-18-2495.

[11] A. K. Gárate-Escamila, A. H. El Hassani, and E. Andrès, “Classification models for heart disease prediction using feature selection and PCA,” *Inform Med Unlocked*, vol. 19, p. 100330, 2020.

[12] K. Raza, “Improving the prediction accuracy of heart disease with ensemble learning and majority voting rule,” *U-Healthcare Monitoring Systems*. Elsevier, pp. 179–196, 2019.

[13] Y.-T. Kim, D.-K. Kim, H. Kim, and D.-J. Kim, “A comparison of oversampling methods for constructing a prognostic model in the patient with heart failure,” in *2020 International Conference on Information and Communication Technology Convergence (ICTC)*, IEEE, 2020, pp. 379–383.

[14] A. Ishaq *et al.*, “Improving the prediction of heart failure patients’ survival using SMOTE and effective data mining techniques,” *IEEE access*, vol. 9, pp. 39707–39716, 2021.

[15] M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai, and R. S. Suraj, "Heart disease prediction using hybrid machine learning model," in *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, IEEE, 2021, pp. 1329–1333.

[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. Wang *et al.*, "A Stacking-Based Model for Non-Invasive Detection of Coronary Heart Disease," *IEEE Access*, vol. 8, pp. 37124–37133, 2020, doi: 10.1109/ACCESS.2020.2975377.

[18] W. Chang, Y. Liu, X. Wu, Y. Xiao, S. Zhou, and W. Cao, "A New Hybrid XGBSVM Model: Application for Hypertensive Heart Disease," *IEEE Access*, vol. 7, pp. 175248–175258, 2019, doi: 10.1109/ACCESS.2019.2957367.

[19] J. Liu, X. Dong, H. Zhao, and Y. Tian, "Predictive Classifier for Cardiovascular Disease Based on Stacking Model Fusion," *Processes 2022, Vol. 10, Page 749*, vol. 10, no. 4, p. 749, Apr. 2022, doi: 10.3390/PR10040749.

[20] F. Ali *et al.*, "A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion," *undefined*, vol. 63, pp. 208–222, Nov. 2020, doi: 10.1016/J.INFFUS.2020.06.008.

[21] T. Amarbayasgalan, V.-H. Pham, N. Theera-Umpon, Y. Piao, and K. H. Ryu, "An Efficient Prediction Method for Coronary Heart Disease Risk Based on Two Deep Neural Networks Trained on Well-Ordered Training Datasets," *IEEE Access*, vol. 9, pp. 135210–135223, 2021, doi: 10.1109/ACCESS.2021.3116974.

[22] I. D. Mienye, Y. Sun, J. M. Corchado, and R. Mehmood, "Improved Heart Disease Prediction Using Particle Swarm Optimization Based Stacked Sparse Autoencoder," *Electronics 2021, Vol. 10, Page 2347*, vol. 10, no. 19, p. 2347, Sep. 2021, doi: 10.3390/ELECTRONICS10192347.

[23] A. Ghosh and S. Jana, "A Study on Heart Disease Prediction using Different Classification Models based on Cross Validation Method," *International Journal of Engineering Research & Technology*, vol. 11, no. 6, Jun. 2022, doi: 10.17577/IJERTV11IS060029.

[24] S. Pillai, "Cardiac Disease Prediction with Tabular Neural Network; Cardiac Disease Prediction with Tabular Neural Network," *Article in International Journal of Engineering and Technical Research*, 2022, doi: 10.5281/zenodo.7750620.

[25] M. Moshawrab, M. Adda, A. Bouzouane, H. Ibrahim, and A. Raad, "Machine Learning Models to Predict Cardiovascular Events from Heart Rate Variability Data," in *2022 3rd International Conference on Human-Centric Smart Environments for Health and Well-being (IHSH)*, IEEE, Oct. 2022, pp. 82–87. doi: 10.1109/IHSH57076.2022.10092060.

[26] Y. Wang *et al.*, "Hyperspectral Estimation of Soil Copper Concentration Based on Improved TabNet Model in the Eastern Junggar Coalfield," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–20, 2022, doi: 10.1109/TGRS.2022.3190310.

[27] H. Prabowo, A. A. Hidayat, T. W. Cenggoro, R. Rahutomo, K. Purwandari, and B. Pardamean, "Aggregating Time Series and Tabular Data in Deep Learning Model for University Students' GPA Prediction," *IEEE Access*, vol. 9, pp. 87370–87377, 2021, doi: 10.1109/ACCESS.2021.3088152.

[28] V. Borisov, K. Seßler, T. Leemann, M. Pawelczyk, and G. Kasneci, “LANGUAGE MODELS ARE REALISTIC TABULAR DATA GENERATORS”, Accessed: May 16, 2023. [Online]. Available: https://github.com/kathrinse/be_great

[29] M. Teresa García-OrdásOrd *et al.*, “Multimedia Tools and Applications Heart disease risk prediction using deep learning techniques with feature augmentation”, doi: 10.1007/s11042-023-14817-z.

[30] X. Sun, M. Liu, and Z. Sima, “A novel cryptocurrency price trend forecasting model based on LightGBM,” *Financ Res Lett*, vol. 32, p. 101084, Jan. 2020, doi: 10.1016/j.frl.2018.12.032.

[31] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, “CatBoost: unbiased boosting with categorical features,” *Adv Neural Inf Process Syst*, vol. 2018-December, pp. 6638–6648, Jun. 2017, doi: 10.48550/arxiv.1706.09516.

[32] S. Arık and T. Pfister, “TabNet: Attentive Interpretable Tabular Learning,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 8, pp. 6679–6687, May 2021, doi: 10.1609/AAAI.V35I8.16826.

[33] N. Hollmann, S. Müller, K. Eggensperger, and F. Hutter, “TABPFN: A TRANSFORMER THAT SOLVES SMALL TABULAR CLASSIFICATION PROBLEMS IN A SECOND”, Accessed: May 14, 2023. [Online]. Available: https://github.com/automl/TabPFN.

PAPER NAME

KaunstavContentPlagCheck.pdf

AUTHOR

M.tech kaustav

WORD COUNT 5930 Words	CHARACTER COUNT 32735 Characters
PAGE COUNT 29 Pages	FILE SIZE 2.2MB
SUBMISSION DATE May 23, 2023 4:39 PM GMT+5:30	REPORT DATE May 23, 2023 4:40 PM GMT+5:30

● 8% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 4% Internet database
- Crossref database
- 5% Submitted Works database
- 4% Publications database
- Crossref Posted Content database

● Excluded from Similarity Report

- Bibliographic material
- Small Matches (Less then 8 words)

Gmail Kaustav Sen <kaustavsen11@gmail.com>

Acceptance Notification - IEEE 4th INCET 2023

1 message

Microsoft CMT <email@msr-cmt.org> Mon, Apr 24, 2023 at 3:59 PM
Reply-To: Deepak Gupta <deepak_gupta@gibds.org>
To: Kaustav Sen <kaustavsen11@gmail.com>

Dear Kaustav Sen

Paper ID / Submission ID : 1369

Title : Heart Disease Prediction Using a Soft Voting Ensemble of Gradient Boosting Models, RandomForest, and Gaussian Naive Bayes

Greeting from 4th INCET 2023.

We are pleased to inform you that your paper has been accepted for the Oral Presentation as a full paper for the- “IEEE 2023 4th INTERNATIONAL CONFERENCE OF EMERGING TECHNOLOGIES (INCET), Belagavi, Karnataka, India with following reviewers’ comment.

All accepted and presented papers will be submitted to IEEE Xplore for the further publication.

Note:
All of Accepted and Presented Papers of INCET series has been Published by IEEE Xplore and indexed by Scopus and other Reputed Indexing partners of IEEE. - http://www.incet.org/history-incet/

You should finish the registration before deadline, or you will be deemed to withdraw your paper:

Complete the Registration Process (The last date of payment Registration is
27 APRIL 2023)

Payment Links

For Indian Authors: https://rzp.io/l/MRk2pUd

For Foreign Authors: https://in.explara.com/e/ieee-incet-2023-ippbj8ay27gam0t

Further steps like IEEE PDF xpress and E copyright will be given later once registration is over after the deadline.

Note :

1. Any changes with the Author name, Affiliation and content of paper will not be allowed after acceptance.
2.This is Hybrid Conference, both online and physical presentation mode is available,

The reviews are below.

======= Review 1 =======

*** Relevance and timeliness: Rate the importance and timeliness of the topic addressed in the paper within its area of research.

INCET 2023

Technically Co-Sponsored by ....

Click-Submit paper

History INCET

Home › History INCET

All accepted and Presented Papers of

1st INCET 2020 has been Published by IEEE Xplore and indexed by Scopus

2nd INCET 2020 has been Published by IEEE Xplore and indexed by Scopus

All papers of the INCET 2020 & 2021 are Indexed by SCOPUS

CLICK HERE – LINK OF WHO – SOURCE SCOPUS INDEXING

CLICK HERE- SOURCE SCOPUS INDEXING

Search

Search