DOI: 10.21917/ijsc.2021.0354

EARLY ONSET DETECTION OF DIABETES USING FEATURE SELECTION AND BOOSTING TECHNIQUES

Shruti Srivatsan[1] and T. Santhanam[2]
[1]*Department of Computer Science and Engineering, Sri Venkateswara College of Engineering, India*
[2]*Department of MCA, DG Vaishnav College, India*

Abstract

Diabetes is one of the most common diseases present in human beings. It is well known that diabetes is a metabolic disease with no permanent cure but on early detection longevity can be increased. This research work focuses on predicting the early onset of diabetes. The diabetic dataset from UCI Machine Learning Repository is used. The necessary preprocessing techniques have been carried out to make the data more robust and suitable for further processing. This research work proposes two feature selection and ensemble boosting techniques resulting in four combinations (models) to predict the presence of diabetes in persons. Also, a novelty is introduced in further reducing the number of features selected by the feature selection techniques. The reduction in the number of features will reduce the memory and time complexity of the model. Among the models proposed, Light Gradient Boosting (LightGBM) with Recursive Feature Elimination (RFE) as feature selector has produced better performance. Further, LightGBM with least features gave satisfactory results.

Keywords:

Data Mining, Boosting, Medical Mining, Diabetes, Feature Selection

1. INTRODUCTION

Various changes in lifestyle have led to the rise of different lifestyle diseases like diabetes, obesity and hypertension. A lifestyle disease refers to a person's illness which is developed due to their daily habits and unhealthy eating patterns. They are chronic diseases which can be cured only on early diagnosis since they affect longevity.

Lifestyle diseases in general act as comorbidities for coronavirus, a deadly infection caused by severe acute respiratory syndrome coronavirus 2 (SARS-COV-2). They have shown detrimental effects on the host body, worsening the prognosis and leading to increased mortality, especially in patients with Black Fungus [1]. Certain evidence suggest an increase in Parkinson's disease among diabetics who have a higher Body Mass Index (BMI) [2]. When diabetics suffer with acute myocardial infarction, they have about 50% higher risk of encountering heart failure [3].

According to a World Health Organisation (WHO) report, about 5.8 million deaths occur in India every year due to non-communicable diseases like cancer and diabetes. About 25% of the population remains at risk of premature deaths and 55% of the mortality rate in the working population is caused due to several metabolic risk factors like high cholesterol and sugar levels [4].

This research study focuses on using Data Mining (DM) in real-time applications to predict the risk of a person becoming diabetic. This article comprises of five sections. Following the introduction, the second section presents the review of literature. Section three deals with a brief description of the dataset and section four describes the proposed models used. In the final section, the results and conclusions are discussed.

1.1 DIABETES MELLITUS

Diabetes is a metabolic disorder caused due to a high blood sugar level over a prolonged period of time. It is caused due to insufficient insulin production in a person's body [5].

From 2.8% prevalence worldwide in 2000, it is estimated that about 4.4% will be diabetic in 2030. About 20% of people aged 65 years and above are said to be diabetic. In India, about 9% of the population is affected by this disease.

There exists two types of diabetes namely Type-I and Type-II. In Type-I diabetes, there is a deficiency of insulin production in the body and requires daily administration of insulin. It is usually associated with gene mutations. Diabetic Ketoacidosis is an acute life-threatening complication that arises due to Type-I diabetes. Due to insufficient insulin, the glucose is unable to enter the cells and is filtered by kidney through urine. When the body cells require sugar for energy, they break down fat and muscle deposit. This results in chemical imbalance in the body, which is quite dangerous.

Type-II diabetes is more prevalent among the diabetic population, caused due to excessive body weight and physical inactivity. It usually occurs in persons with high blood sugar levels, leading to insulin resistance in their body. When there is an excessive rise in blood sugar, the body tries to remove the excess sugar through urine and this causes excessive urination and thirst. When this condition deteriorates, the person might encounter seizures, further leading to their death.

Some of the health complications caused by diabetes are stroke, vision loss, kidney failure and microvascular disorders [6]. Hence, early detection of diabetes is highly preferred. Majority of the people suffering from diabetes are undiagnosed because of its long-term asymptomatic phase. This can be detected only by examining the symptoms which are less prevalent and also the common ones, which could be found in different phases from disease initiation up to diagnosis.

1.2 DATA MINING

Data Mining (DM) refers to the process of extracting data, analyzing it from many dimensions or perspectives and producing a summary of the information in a useful form that identifies relationships within the data. This information is very useful for decision making.

Knowledge Discovery in Databases (KDD) is considered as a programmed, exploratory analysis and modeling of vast data repositories. The model is used to extract the knowledge, analyze and predict the necessary data [7]. Nowadays, it has become a

buzzword and all leading companies are trying to implement Data Mining to develop better business strategies.

The different steps involved in performing Data Mining over a given dataset are outlined below:

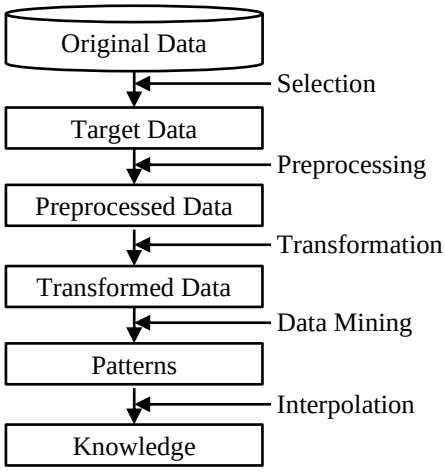

Fig.1. KDD Process

1. *Selection*: It is a collection of in-depth insights of data which are obtained along with the necessary requirements.
2. *Target Dataset Creation*: Based on the insights gained, a suitable dataset is chosen or created.
3. *Data Cleaning/Preprocessing*: Researchers and data scientists spend majority of their time in preprocessing, since it is a vital step in Data Mining. Preprocessing techniques such as missing value replacement (imputation), removing outliers, scaling, balancing of the dataset, encoding of data and selecting best features are carried out to enhance the overall performance of the model.
4. *Dimensionality reduction*: It reduces the effective number of dimensions for better interpretation of input features and visualisation. This will essentially reduce the memory and time requirements.
5. *Data Mining Algorithm*: A suitable algorithm is chosen based on the application domain.
6. *Data Mining*: Using the selected algorithm, relevant patterns are inference from the specific representation.
7. *Interpretation of Mined Data*: On examining the mined data, interpretation is done based on the problem.
8. *Consolidation*: The discovered knowledge is consolidated and its impact is measured for generating reports [8].

1.3 DATA MINING APPLICATIONS

Data Mining can be applied to any form of data that exist as *images, text, maps* and many more. It has varied applications across different domains and some of its applications are illustrated below:

- **Text Mining**: In [9], K-means algorithm has been used iterating over different values of k. Techniques such as word-level analysis and information retrieval have been used for intelligent food production and human nutrition [10]. Online Customer Reviews (OCR) have been analysed using Latent Dirichlet Allocation (LDA) to predict airline recommendations with an accuracy of 79.9% [11].
- **Web Mining**: In Web Mining, information has been automatically discovered and extracted from web documents. There has been a usage of resource finding and generalization to find general patterns on websites [12]. Online Social Networks (OSN) have been analysed to terminate websites spreading harmful content in the event of a terrorist attack [13].
- **Process Mining**: Oracle Data Mining (ODM) has been applied to banking data using K-means clustering algorithm. The data has been segregated into respective clusters followed by its analysis [14]. Fuzzy systems have been used to compute the trust rate in customer's behavior. The customer bank Internet transaction has been analysed and steps have been taken to increase the security of their internet banking [15].
- **Medical Data Mining**: Medical Mining is used to study a patient's vital signs to understand his illness and predict the future by analysing them. Medical dataset consists of mostly images in the form of scans. Image processing has been carried out to detect anomalies in these images [16]. Daily activity data has been collected from wearable sensors like smart watches to monitor a person's heart rate [17].

2. REVIEW OF LITERATURE

Data Mining has a lot of use cases in the medical domain which has been used for diagnosis and prognosis of diseases including stroke, cancer, cataract and diabetes.

Chaurasia et al. [18] have used Sequential Minimal Optimization (SMO) and Bloom Filter Trees, obtaining around 96% accuracy in cancer detection. Colon cancer and breast cancer have been detected by using Support Vector Machine (SVM). Information gain has been used to remove irrelevant features and the selected features were further reduced by Grey Wolf Optimization (GWO) [19].

Stroke occurrence has been predicted by Sheetal Singh et al. using decision trees for feature selection and back propagation neural networks to get an accuracy of 97% [20]. Haemorrhagic stroke has been analysed using J48, Jrip and multilayer perceptron. The mortality rate was predicted using J48 model, since it yielded better performance metrics [21].

Decision trees and Generalized Linear Model (GLM) were among the six machine learning techniques used by Shahbaz M. et al. with 88% accuracy in [22]. When Bagging, Boosting and Stacking were applied to the OASIS dataset, Random Forest (Bagging algorithm) gave an accuracy of 90% for Alzheimer disease prediction [23].

C5.0 algorithm has been used by Nair et al [24] to build a model which predicted the occurrence of Cataract. In [25], for the diagnosis of eye diseases, J48 pruned classification was used and it gave 98.5% accuracy.

The spread of coronavirus has been studied in different continents using Boosting algorithms. This helped in understanding the prevalent growth rate of the disease [26]. In [27], around 99.85% accuracy was obtained applying decision tree algorithm on the dataset from South Korea.

Srivatsan and Santhanam [28] have obtained an accuracy of 79% using ID3 Decision Tree algorithm [28]. Intra Uterine Growth Restriction was predicted using Naïve Bayes resulting in good performance metrics including an accuracy of 84% and recall of 86.7% [29].

Perveen et al. [30] have used boosting algorithms and J48 decision tree to improve performance analysis of Diabetes prediction in their article titled "Performance Analysis of Data Mining Classification Techniques to Predict Diabetes". Adaboost gave better results than other ensemble classifiers used. Das H. et al. [31] performed various classification and clustering algorithms for getting the required accuracy. Ahmed Tariq [32] has used WEKA tool on J48 decision tree classifier to obtain an accuracy of about 70 %. In [33], when a comparative analysis was done for predicting diabetes, K Nearest Neighbors (KNN) gave the best performance. Yang et al. [34] have compared algorithms like SVM and LDA with ensemble classifiers, with the latter giving significantly higher metrics. In [35], an accuracy of 93% is obtained on using LightGBM with RFE.

Random Forest yielded significant results in predicting diabetes in female patients [36]. Risk of type-2 diabetes was predicted using weighted feature selection. Out of the different models used, Adaboost performed the best [37]. Decision Tree performed well when the risk factors correlation were analysed for the Ulster Community and Hospitals Trust (UCHT) dataset [38]. In [39], when a comparison was drawn between algorithms like Deep Neural Network, Random Forest and SVM, DNN yielded a very high accuracy. Islam et al. [40] have used Logistic Regression and Random Forest with cross validation and percentage split method. Among the algorithms applied, Random Forest has given the highest accuracy of 97.4%.

3. DATASET DESCRIPTION

UC Irvine Machine Learning Repository is a well-known repository used by researchers across different domains including Life Sciences, Engineering and Business. Currently at UCI, there are 588 datasets of which 138 are under Life Sciences domain. The dataset used in this study is one of them [41]. The dimension of the dataset is (520,17) with 16 input features and 1 output feature.

A brief description of the attributes or features is given below:

1. *Age:* It refers to the age of the person having a greater risk of diabetes ranging from 16-90.
2. *Gender:* It denotes the gender of the patient. (Male/Female)
3. *Polyuria*: It determines whether the patient has Polyuria or excessive urination. (Yes/No)
4. *Polydipsia*: It represents whether the patient has Polydipsia or excess thirst. (Yes/No)
5. *Sudden weight loss*: It indicates whether the patient has sudden weight loss or not. (Yes/No)
6. *Weakness*: It determines whether the patient has weakness or not. (Yes/No)
7. *Polyphagia*: It obtains information on whether the patient eats excessively. (Yes/No)
8. *Genital thrush*: It determines whether the patient has genital thrush or vaginal infection. (Yes/No)
9. *Visual blurring:* It shows whether the patient has encountered blurred vision. (Yes/No)
10. *Itching*: It determines whether the patient has any kind of itching. (Yes/No)
11. *Irritability:* It finds whether the patient has irritation. (Yes/No)
12. *Delayed healing*: It tells whether the patient has delay in healing of wounds. (Yes/No)
13. *Partial paresis*: It determines whether the patient has weak muscles. (Yes/No)
14. *Muscle stiffness*: It suggests whether the patient muscles feel tight. (Yes/No)
15. *Alopecia*: It is used to find whether the patient has severe hair loss. (Yes/No)
16. *Obesity*: It finds whether the patient is obese or not. (Yes/No)
17. *Class*: Class is an output feature which indicates whether the patient is at risk of diabetes with Negative and Positive as its values.

4. PROPOSED MODEL

The objectives of this research study are given below:

- The first objective of this research study is to use around 50% of the input features and build a suitable ensemble model to get reasonable accuracy, precision and recall.
- The second objective is to decrease the memory usage and execution time.
- The final objective is to bring down the false negative rate to the extent possible, otherwise it may lead to further complications which might be fatal. In healthcare domain, recall is given more importance than precision.

The snapshot of the dataset is shown in the Fig.2:

	Age	Gender	Polyuria	Polydipsia	Sudden Weight Loss	Weakness	Polyphagia	Genital Thrush
0	40	Male	No	Yes	No	Yes	No	No
1	58	Male	No	No	No	Yes	No	No
2	41	Male	Yes	No	No	Yes	Yes	No
3	45	Male	No	No	Yes	Yes	Yes	Yes
4	60	Male	Yes	Yes	Yes	Yes	Yes	No

Fig.2(a). Columns 1-8 of the diabetes dataset

	Visual Blurring	Itching	Irritability	Delayed Healing	Partial Paresis	Muscle Stiffness	Alopecia	Obseity	Class
0	No	Yes	No	Yes	No	Yes	Yes	Yes	+ve
1	Yes	No	No	No	Yes	No	Yes	No	+ve
2	No	Yes	No	Yes	No	Yes	Yes	No	+ve
3	No	Yes	No	Yes	No	No	No	No	+ve
4	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	+ve

Fig.2(b). Columns 9-17 of the diabetes dataset

From the Fig.2(a) and Fig.2(b), it is clear that Age is the only continuous feature whose range is from 16 to 90. The rest of the

features are categorical with only two possible values. An analysis has been done between Gender and Age in Fig.3.

Gender	Class	Age				
		Mean	Median	Max	Min	Count
Female	-ve	46.315789	50	65	28	19
	+ve	46.860465	47	90	25	172
Male	-ve	46.364641	45	72	26	181
	+ve	51.150685	53	85	16	146

Fig.3. Analysis between Gender and Age

In Fig.3, out of 191 female samples in the dataset, 90% are positive (diabetic) and the remaining 10% are negative (non-diabetic). Similarly, out of 327 male samples, about 45% of them are diabetic and the rest are non-diabetic. One can infer from the Fig.3 that females are more susceptible to diabetes at an earlier age than males. The maximum age of a male is 85 and that of a female is 90.

The dataset is then split into 80:20 where 80% split is used for training the model and the remaining 20%, which is unseen by the model is used to validate it.

The preprocessing steps that have been carried out are explained below:

If missing values are present in the dataset and are not handled properly, it will result in poor learning by the model and undesirable outcomes. Therefore, it is necessary to deal with them properly using suitable imputation techniques. In this dataset, there are no missing values. The next step is to examine outliers.

Outliers are extreme data points which are different from other data points in a continuous feature. There are many techniques available to detect outliers and one of them is boxplot, a visualization tool. It provides information about Quartile-1 (25%), Quartile-2 (50%) or median, Quartile-3 (75%), lower whisker and upper whisker. Any datapoint present outside the whiskers are referred to as outliers [42].

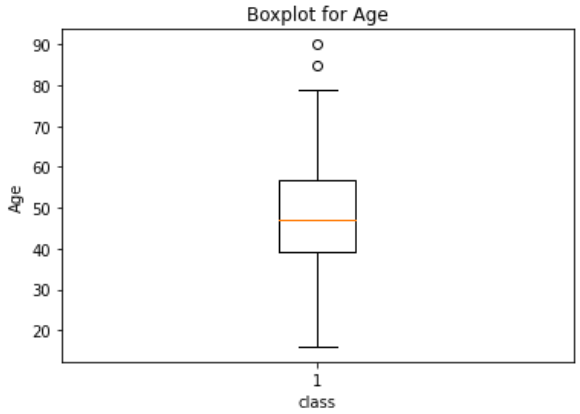

Fig.4. Boxplot of Age

Boxplot is drawn for the feature Age and shown in Fig.4. It is evident from the plot that there are two data points encircled above the upper whisker referred as outliers. On close observation of the two data points we concluded that their values 85 and 90 are within acceptable limits since nowadays people generally live upto 100 years. Hence, they are retained. Data imbalance will be analysed in the following step.

Out of 520 records, 320 records are positive and the remaining are negative which is in the ratio of 60:40 approximately. This suggests that the dataset is not highly imbalanced. The final preprocessing step is to encode the categorical variables in the dataset.

Input encoding techniques such as One-Hot encoding increase the number of columns for each feature which inturn increases the memory space. A novelty is introduced in this study by using the Label encoding, an output encoding technique to encode the input features as well.

In this step, two feature selection techniques namely SelectKBest and Recursive Feature Elimination (RFE) are applied to select the six best input features. Then, the dataset with these six features is applied to two boosting models XGBoost and LightGBM to learn (using the training set) and its performance is validated using the test set.

Each of the two feature selection and boosting techniques used in this study give rise to four possible combinations.

The four models proposed to predict the risk of diabetes are:

- SelectKBest with XGBoost – Model 1
- RFE with XGBoost – Model 2
- SelectKBest with LightGBM – Model 3
- RFE with LightGBM – Model 4

4.1 PROPOSED DESIGN FLOW

The flowchart given in Fig.5 portrays the steps involved in the proposed model.

The methods used in this research study are explained below:

4.1.1 SelectKBest for Feature Selection:

SelectKBest is a type of Univariate selection which is a filter-based technique. In a filter-based feature selection technique, features are selected using a rank ordering method to filter out irrelevant features. The ranks are devised by analysing the statistical scores, which are determined by the correlation of the features with the target variable. It is used to select features which have the strongest relationship with the output variable. This can be used for numerical and categorical data. The K-highest scoring features are retained.

This technique has been used to get meaningful insights from hotel booking data, SelectKBest with chi2-score function is used as filter based feature selector along with Kmeans Clustering [43].

Computation is performed faster than wrapper-based feature selection techniques since there is no involvement of trained models. When there is lesser data, the best subset of features may not be found by filter methods. There are two parameters for this technique namely score function and the value of K. The score function can use ANOVA F-value, mutual information or chi-squared tests for computation. The number of top features to be selected is denoted by the value of K [44]. The method adopted in this study is ANOVA F-value.

SelectKBest Algorithm:

Step 1: Apply parameter score function to (X,Y) where X refers to the predictors and Y refers to the target variable.

Step 2: An array of scores are returned for each feature.

Step 3: The first k features with highest scores are retained [45].

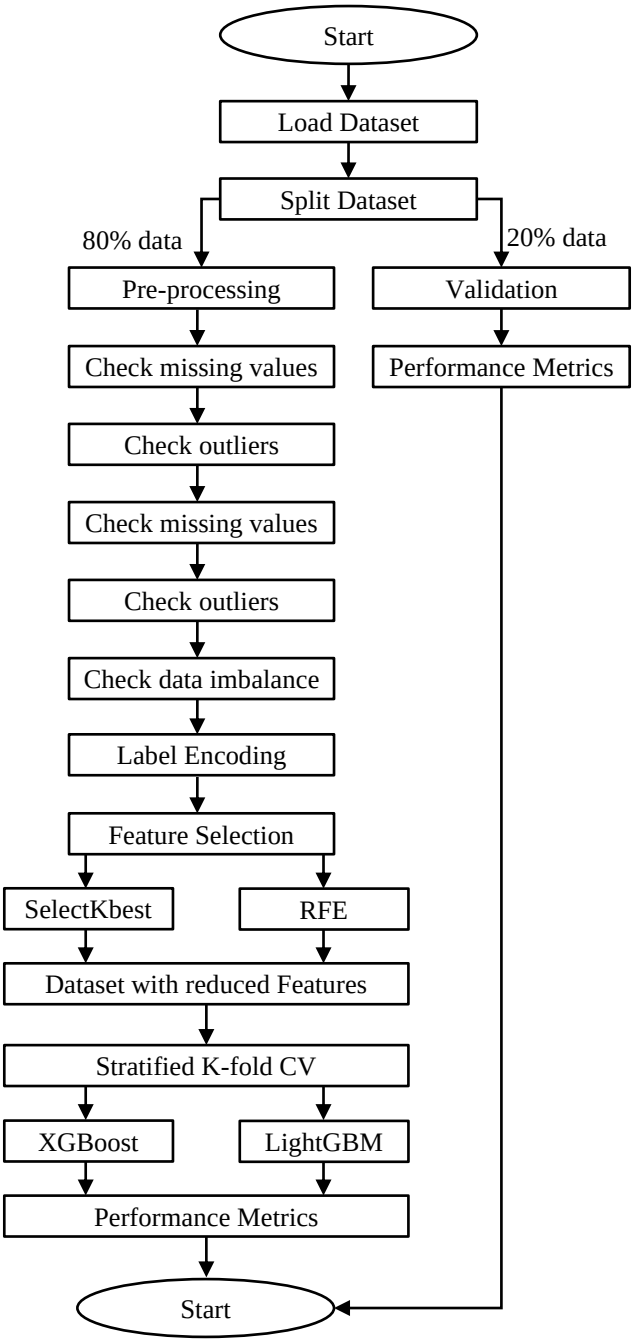

Fig.5. Flow Diagram of the Proposed Model

4.1.2 Recursive Feature Elimination (RFE):

RFE is a wrapper-based feature selection technique. It performs a greedy search for a subset of features starting with all features in the training dataset and successfully removing features until the desired number remains. A machine learning algorithm is fitted in the core of the model for ranking features by importance, discarding the least important features and re-fitting the model. This process is repeated until a specified number of features remain.

Recursive Feature Elimination is applied to problems based on Binary Biogeography Optimization (BBO) along with Support Vector Machine (SVM) to give significant results [46].

In a wrapper-based technique, a machine learning algorithm is chosen and the features best suited for the algorithm are retained. For a classification problem, the predictive accuracy is considered whereas in a clustering problem, the goodness of the cluster is analysed. Though it takes a longer time for computation than filter techniques, wrapper methods always provide the best subset of features since they are exhaustive in nature. The parameters in this technique are the estimator or the model to fit the features and the number of features to be selected [47].

RFE Algorithm:

Step 1: Search for a subset of features among all features.

Step 2: A machine learning model is fitted and the features are ranked based on importance.

Step 3: The least important features are discarded.

Step 4: The model is refitted and the process continues till a certain number of features remain [48].

4.2 BOOSTING ALGORITHMS

A boosting algorithm is an ensemble model which combines many simple models and generates the final output. It builds a strong classifier using weak classifiers. It is said to improve the performance of the models [49].

To predict the GDP growth of Japan, Gradient Boosting and Random Forest models were used, with the former yielding better metrics [50]. For network intrusion detection, Modest Adaboost, Gentle and Real Adaboost algorithms were applied to public datasets. Modest Adaboost was faster than the others but yielded higher error rate [51].

The algorithm was proposed in 1990 as an answer by Robert Schapire, an American Computer Scientist at Princeton University to a question raised by a Harvard professor Valiant. He was awarded with the prestigious Gödel prize in 2003 [52].

A predictive model is built initially using a trained model and subsequently another model is used which rectifies the errors from the previous model. The process continues till the model makes no errors or the maximum number of models are obtained. Different weights are assigned to the previously classified incorrect samples as a penalty [53].

It is an easy-to-interpret algorithm which gives higher performance since weak classifiers are made strong learners. Overfitting is curbed easily. There are mainly three types of Boosting namely AdaBoost, Gradient Boosting and Extreme Gradient Boosting [54].

In this work, Extreme Gradient Boosting (XG Boost) technique has been applied.

4.2.1 XGBoost Algorithm:

Initially started as a research project, Tianqi Chen and Carlos Guestrin developed this algorithm for Distributed Machine Learning Community (DMLC). It became popular among developers since it gave winning solutions for different machine learning challenges [55].

Extreme Gradient Boost is one of the Boosting algorithms which has a faster execution speed and better model performance when compared to other boosting algorithms. It uses an approach where new models are created which predict the errors of prior models after which they are added together to make the final

prediction. A gradient descent algorithm is used to minimize the loss when adding new models. Both classification and regression type problems use this algorithm.

Its applications include determining the durability of concrete, Electrical Resistivity Measurement (ERM) analysis [56]. Real-time accident detection on highways was performed with XGBoost and SHAP (SHapley Additive exPlanation) and an accuracy of 79% was obtained [57].

A few of the salient features of this algorithm include clever penalization of sub-trees constructed, shrinking of leaf nodes proportionally, using cache optimisation and performing out-of-core computation [58]. A hyperparameter is a value which is used to control the learning process of a model. There are many hyperparameters for this algorithm. Some of its vital hyperparameters include learning rate, minimum child weight, regularization term and gamma [59].

Working of XG Boost:

Step 1: Selection of '*k*' features

Step 2: Smaller trees are built with fewer splits

Step 3: Subsequent building of trees happen.

Step 4: Errors present in predecessors are updated in residual errors [60]

4.2.2 Light Gradient Boosting (lightGBM) Algorithm:

Guolin Ke and her team developed this algorithm in 2016 as a part of Microsoft Research to reduce memory usage and increase computation speed [61].

Blood Brain Barrier (BBB) permeability has been predicted using LightGBM, resulting in an accuracy of 90% [62]. LightGBM has performed better than other models used for forecasting the cryptocurrency price [63].

Since the algorithm puts continuous feature values into discrete bins, training happens faster than other boosting algorithms and there is a reduced usage of memory. It supports parallel learning.

To enhance its performance, hyperparameters like number of leaves in a tree, maximum buckets for filling feature values, maximum depth are altered with suitable values [64].

LightGBM is a fast and high-performance gradient boosting framework which is built with decision tree algorithm, used for ranking and classification. It yields a better accuracy because of splitting of tree-leaf with best fit rather than a depth wise approach. It is highly efficient and also uses lesser memory.

LightGBM Algorithm:

Step 1: The best attributes are selected.

Step 2: A decision tree model is built with leaf nodes split using best fit

Step 3: Maximum depth is created to avoid overfitting [65]

4.3 PERFORMANCE METRICS

Our study considers the following classification performance metrics:

4.3.1 Confusion Matrix:

It is a performance measurement evaluation method used for obtaining the correctness of an algorithm. It comprises of a table like format with four values computed using Actual and Predicted values.

Type	Actual Positive	Actual Negative
Predicted Positive	TP (11)	FP (10)
Predicted Negative	FN (01)	TN (00)

Fig.6. Confusion Matrix for Binary Classifier

Description of Confusion Matrix: 00 represents True Negative (TN), 01 denotes False Negative (FN), 10 indicates False Positive (FP), 11 shows True Positive (TP).

- *True Negative*: When the real and predicted labels of a sample are negative.
- *True Positive*: When the real and predicted labels of a sample are positive.
- *False Negative*: When the real value of a sample is positive while its predicted label is negative.
- *False Positive*: When the real value of a sample is negative while its predicted label for the sample is positive [66].

4.3.2 Precision:

Precision refers to the measure of exactness or the ratio of correctly classified classes of diabetes amongst all positive classes.

$$Precision=TP/(TP+FP) \quad (1)$$

4.3.3 Recall:

Recall represents the measure of completeness or the ratio of correctly classified classes with all correctly predicted classes.

$$Recall = TP/(TP+FN) \quad (2)$$

4.3.4 Accuracy:

Accuracy is the most intuitive performance measure referring to the ratio of correctly classified classes among all classes [67].

$$Accuracy = (TP+TN)/(TP+TN+FP+FN) \quad (3)$$

4.3.5 f-Beta-Measure:

The f-Beta measure is a performance metric used to compare two models with varying recall and precision by penalising the extreme values [68].

$$F\text{-}measure=((1+\beta^2)*Recall*Precision))/(\beta^2)*(Recall+Precision) \quad (4)$$

Choosing a suitable Beta value depends on the application domain. When Beta value is 1, equal weightage is given to both precision and recall. When more emphasis is given on precision, a value lesser than 1 is chosen and when recall is given higher importance, a value greater than 1 is chosen [69].

For object detection using an efficient network MINet, a Beta value of 0.3 is chosen [70]. In a Hope Speech detection, the Beta value remains as 1 [71]. Usually in medical datasets, a Beta value greater than 1 is chosen to give priority to minority classes [72]. Since, in medical domain, recall plays a vital role, the value of Beta in this study has been chosen as 2 on a trial and error basis.

4.3.6 Area Under the Curve:

It is used to measure how accurately a classifier is able to distinguish among classes. If its value is greater than 0.8, the model is said to be well-fitted [73].

4.3.7 Kappa-Statistic:

Cohen's Kappa value is used to determine the interrater reliability for a data item. This way, the possibility of a prediction by a classifier overlapping with a random guess is removed. A value above 0.81 signifies a near-perfect agreement.

$$K = (p_0\text{-}p_e)/(1\text{-}p_e) \quad (5)$$

where p_0 is the overall accuracy of the model and p_e is the measure of agreement between the classifier's prediction and the actual class label assuming it to be random [74].

5. EXPERIMENTAL RESULTS

In this work, it has been decided to evaluate the performance of the proposed models with all input features, and with reduced input features obtained using feature selection techniques. Also, it was decided to test the performance of the model with less than 50% of the input features on a trial and error basis.

The outcome of the two feature selection techniques are given below:

- The six best features obtained using SelectKBest are Polyuria, Polydipsia, Gender, sudden weight loss, Polyphagia and partial paresis.
- The six best features obtained using RFE are Polyuria, Polydipsia, Gender, Genital thrush, Itching and Irritability.
- Out of the six features listed above, three features namely: Polyuria, Polydipsia and Gender were found to be common.

Baseline Accuracy is a simple prediction of the model's performance without any rules and can be used as a baseline for achieving more accuracy [75]. ZeroR, a baseline classifier, is used in this study to predict the majority class only based on the target by ignoring all the predictors. The baseline accuracy obtained from ZeroR classifier can be used as a benchmark to analyse the performance of other classification techniques [76]. The Baseline Accuracy of the proposed models has been estimated as 61.53% approximately.

The next step is to train the model using Stratified K Fold cross validation technique. Generally, Stratified K-fold cross validation results in better performance of the model when compared with a single train:test split technique. The training set of 80% is used by the K-Fold cross validation technique with K as 10. The mean accuracy is taken as the training accuracy. The testing accuracy predicted by the model for the real-world data is the final accuracy. This helps to prevent the model from overfitting and high bias. Further, the feature selection techniques also helps to reducing overfitting.

Next, we wish to examine the memory and training time requirements of both the models with all input features. In the case of XGBoost, it is 69.2kB and 0.62s respectively whereas in LightGBM, the memory requirement remains the same but the time taken for training is 0.89s. The models have been implemented in python using suitable libraries. The results obtained for the four different proposed models are described in the following sections:

5.1 PERFORMANCE OF XGBOOST MODEL WITH SELECTKBEST AS FEATURE SELECTOR

The performance metrics of using SelectKBest with XGBoost is shown in Table.1 and Table.2.

Table.1. Results of XGBoost with SelectKBest as feature selector (Proposed Model 1)

Metrics	Value	Metrics	Value
Train Accuracy	89.6%	Train time	0.46s
Test Accuracy	93.26%	Test time	0.15s
Precision	95.2%	AUC Score	0.972
Recall	88.8%	Kappa	0.861
F1-Score	0.901		

Table.2. Confusion Matrix of XGBoost

Type	Actual Positive	Actual Negative
Predicted Positive	40	2
Predicted Negative	5	57

From Table.1, it is clear that the built model does not suffer from overfitting or underfitting seeing the values of training or testing accuracies i.e. it is a generalised model.

The precision and recall values are satisfactory and the overall accuracy is reasonable. F-Measure has a value of 0.901, which is closer to 1 denotes a better performance in real-life classification of data. AUC score of 0.972 signifies that the model is able to classify between two classes 97% of times and it further suggests that the model is well fitted. It is reported in literature that a model with AUC score closer to 1 is generally preferred. Kappa value of 0.861 indicates a near-perfect agreement. The training and testing times are 0.46 and 0.15 seconds respectively. Generally, a model which can learn quickly with lesser data is preferred.

In Table.2, the false negative count is reported as 5, which is slightly on the higher side. This count should be less as far as possible especially in healthcare domain. From Table.1 and Table.2, it can be concluded that overall performance of the model is satisfactory.

5.2 PERFORMANCE OF XGBOOST MODEL WITH RFE AS FEATURE SELECTOR

The performance metrics of using RFE with XGBoost is shown in Table.3 and Table.4.

Table.3. Results of XGBoost with RFE as feature selector (Proposed Model 2)

Metrics	Value	Metrics	Value
Train Accuracy	91.84%	Train time	0.59s
Test Accuracy	94.23%	Test time	0.04s
Precision	90.4%	AUC Score	0.991
Recall	95.0%	Kappa	0.881
F1-Score	0.940		

Table.4. Confusion Matrix of XGBoost

Type	Actual Positive	Actual Negative
Predicted Positive	38	4
Predicted Negative	2	60

It is clear that RFE as a feature selector has performed better than SelectKBest when combined with XGBoost, as it has increased accuracy and recall (an important measure), despite taking more time for execution. Further in Table.4, the false negative count is only 2 instead of 5 (from Proposed Model 1). This suggests that RFE is preferred over SelectKBest algorithm.

Hence, it can be easily concluded that RFE with XGBoost performs better than the SelectKBest with XGBoost. If execution time is not a major constraint, Proposed Model 2 can be used. The overall performance metrics are very encouraging. The second objective of this research study is fulfilled to a greater extent with the Proposed Model 2.

5.3 PERFORMANCE OF LIGHTGBM WITH SELECTKBEST AS FEATURE SELECTOR

The performance metrics of using SelectKBest with LightGBM is shown in Table.5 and Table.6.

Table.5. Results of LightGBM with SelectKBest as feature selector (Proposed Model 2)

Metrics	Value	Metrics	Value
Train Accuracy	89.6%	Train time	0.41s
Test Accuracy	94.2%	Test time	0.11s
Precision	97.6%	AUC Score	0.977
Recall	89.1%	Kappa	0.879
F1-Score	0.906		

Table.6. Confusion Matrix of LightGBM

Type	Actual Positive	Actual Negative
Predicted Positive	41	1
Predicted Negative	5	57

From Table.5, one can say that the built model does not suffer from overfitting or underfitting seeing the values of training or testing accuracies (i.e.) it is a generalised model.

The precision and recall values are encouraging and the overall accuracy is good. When compared with Proposed Model 1, there is a slight variation in the confusion matrix, resulting in better performance. The training and testing times are 0.41 and 0.11 seconds respectively, which are lesser than the ones reported in Proposed Model 1. If execution time is a constraint, Proposed Model 3 can be opted for disease classification.

In Table.6, the false negative count is reported as 5, which is not desirable. From Table.5 and Table.6, it can be observed that the overall performance of Proposed Model 3 is highly satisfactory, when compared with Proposed Model 1.

5.4 PERFORMANCE OF LIGHTGBM WITH RFE AS FEATURE SELECTOR

The performance metrics of using RFE with LightGBM is shown in Table.7 and Table.8.

Table.7. Results of LightGBM with RFE as feature selector (Proposed Model 2)

Metrics	Value	Metrics	Value
Train Accuracy	91.84%	Train time	0.33s
Test Accuracy	94.23%	Test time	0.03s
Precision	90.4%	AUC Score	0.989
Recall	95.0%	Kappa	0.879
F1-Score	0.940		

Table.8. Confusion Matrix of LightGBM

Type	Actual Positive	Actual Negative
Predicted Positive	38	4
Predicted Negative	2	60

In Table.7, the precision and recall values are encouraging and the overall accuracy is highly encouraging. Further, the LightGBM model built using SelectKBest feature selection technique has yielded better precision value than with RFE technique. When compared with Proposed Model 2, the metrics are found to be very similar. The training and testing times are 0.33 and 0.03 seconds respectively, which are lesser than the ones reported in Proposed Model 2. If execution time is a constraint, Proposed Model 4 can be opted in comparison with Proposed Model 2.

From Table.8, the values reported in the confusion matrix are similar to Proposed Model 2 and better than Proposed Models 1 and 3. From Table.7 and Table.8, it can be summarised that the overall performance of the model is sufficiently good, when compared with other Proposed Models.

5.4.1 Feature Reduction:

In order to reduce the features further along with the memory and time requirements, a novelty has been tried with regard to the selection of input features based on the existing input features obtained through the two feature selection methods namely SelectKBest and RFE. Let the features obtained from SelectKBest be Set A and those obtained from RFE be Set B. Then the union and intersection of sets A and B results in two more sets C and D. These two feature sets were used by the best model (Proposed Model 4) and their performances were analysed.

Set A comprises of the six features chosen by SelectKBest technique.

A = {***Polyuria, Polydipsia, Gender, Polyphagia, sudden weight loss, partial paresis***}

Set B consists of the six features chosen by RFE technique.

B = {***Polyuria, Polydipsia, Gender, Genital thrush, Itching, Irritability***}

Set C is *A* union *B.*

C = {***Polyuria, Polydipsia, Gender, Genital thrush, Itching, Irritability, Polyphagia, sudden weight loss, partial paresis***}

Set D is *A* intersection *B.*

D = {***Polyuria, Polydipsia, Gender***}

Since LightGBM gave a better performance than XGBoost, it has been decided to apply LightGBM to build three more models. Two models (Proposed Models 5 and 6) have been built after introducing a novelty with regard to the input features available in sets C and D. Also, another machine learning model has been built with all (16) input features (Proposed Model 7). Finally, the performances of the proposed models 4, 5, 6 and 7 have been compared and the obtained results have been reported in Table.9:

Table.9. Comparison of Proposed Model 4 with other models (Proposed Models 5, 6 and 7)

Metrics	**Input features**			
	16	**9**	**6**	**3**
Memory Usage (kB)	69.2	40.6	28.5	16.2
Training time (s)	0.46	0.43	0.33	0.3
Training Accuracy (%)	96.41	94.25	91.84	88.23
Testing time (s)	0.129	0.115	0.095	0.035
Testing Accuracy (%)	98.07	98.07	94.23	94.23
Precision (%)	100	100	90.4	88.1
Recall (%)	95.4	95.4	95	97.3
AUC Score	1	0.994	0.989	0.954
Kappa Statistics	0.963	0.96	0.879	0.878

In Table.9, the first row results indicate the model built with all the input features. The above values suggest that the model performance is highly satisfactory.

In the second row, the results of using the features in set C are shown. The values in rows 1 and 2 are almost similar, leading to the conclusion that the model with 9 input features is more than adequate than with all features. Since, the features are reduced by around 50 %, there is a saving of 42% in memory and 11% in time requirements.

The third row indicates the results of the best model (LightGBM+RFE) and it is very encouraging.

In the fourth row, the results of using the features in set D are given. From the values, one can conclude the model has produced a reasonably good performance with just 3 features and recall being the best in the whole table. Also, the memory and time requirements are further decreased when compared with the previous row. This gives a conclusion that the proposed model with 3 input features is sufficient to a greater extent to decide whether the subject is positive or negative.

Also, the accuracy of all the models proposed in this study have given a better performance when compared with the accuracy of ZeroR classifier (61%) by over 30%.

6. CONCLUSION

This research work has made use of Recursive Feature Elimination and SelectKBest as feature selection algorithms, then boosting algorithms, XGBoost and LightGBM have been applied for predicting diabetes. This study has proposed 7 models and all the models gave encouraging results. Among the two feature selection algorithms that were tried out, RFE with LightGBM gave better results reported in Table.7 than the other three proposed models. Also, other variants in the input features based on the outcome of the two feature selection techniques were explored. Based on this, the model with 3 input features (Polyuria, Polydipsia, Gender) has given reasonably good performance with regard to accuracy, precision and recall. Further, it reduces the time and memory complexity to a certain extent. Also, the 9 input features suggested in this study gave performance similar to having all input features. This implies data collection in future can be done with only those 9 features.

All the stated objectives have been achieved to a great extent in this study.

The scope of this work is limited to this dataset related to people of Bangladesh. It is known that more the number of records, better will be the learning power of the model and inturn it increases the overall performance.

The following points are suggested for future work:

- The best model proposed (RFE with LightGBM) can be experimented with other diabetes datasets like PIMA to assess its performance.
- Other feature selection techniques can be explored.
- New or modified features using the existing features (feature engineering) can be added.
- Mathematical, statistical and neural network classifiers can be tried for building the model.

Hyperparameter tuning using optimization techniques can be implemented to improve the model performance.

REFERENCES

[1] Yue Zhou, “Obesity and Diabetes as High-Risk Factors for Severe Coronavirus Disease”, *Diabetes/Metabolism Research and Reviews*, Vol. 37, No. 2, pp. 1-13, 2021.

[2] Su Min Jeong, “Body Mass Index, Diabetes, and the Risk of Parkinson’s Disease”, *Movement Disorders*, Vol. 35, No. 2, pp. 236-244, 2020.

[3] V. Ritsinger, “Heart Failure Is a Common Complication after Acute Myocardial Infarction in Patients with Diabetes: A Nationwide Study in the Swedeheart Registry”, *European Journal of Preventive Cardiology*, Vol. 27, No. 17, pp. 1890-1901, 2020.

[4] D.C. Klonoff, “The Increasing Incidence of Diabetes in the 21st Century”, *Journal on Diabetes Science and Technology*, Vol. 3, No. 1, pp. 1-2, 2009.

[5] H.A. Shouip, “Diabetes Mellitus”, Technical Report, Faculty of Pharmacy and Pharmaceutical Industries, Sinai University, pp. 1-13, 2007.

[6] P.H. Reddy, “Can Diabetes Be Controlled by Lifestyle Activities?”, *Current Research in Diabetes and Obesity*, Vol. 1, No. 4, pp. 1-13, 2017.

[7] *U.* Fayyad, “Data Mining and Knowledge Discovery: Making Sense Out of Data”, *IEEE Expert*, Vol. 11, No. 5, pp. 20-25, 1996.

[8] M. Goebel and L. Gruenwald, “A Survey of Data Mining and Knowledge Discovery Software Tools”, *Proceedings of ACM Conference on Explorations Newsletter*, pp. 20-33, 1999.

[9] *Said* Salloum and Al-Emran, "Using Text Mining Techniques for Extracting Information", *Proceedings of International Conference on Intelligent Natural Language Processing: Trends and Applications*, pp. 373-397, 2018.
[10] Pengkun Yang and Hao Feng, "Utilization of Text Mining as a Big Data Analysis Tool for Food Science and Nutrition", *Comprehensive Reviews in Food Science and Food Safety*, Vol. 19, No. 2, pp. 875-894, 2020.
[11] *F.R.* Lucini, "Text Mining Approach to Explore Dimensions of Airline Customer Satisfaction using Online Customer Reviews", *Journal of Air Transport Management*, Vol. 83, pp. 1-13, 2020.
[12] R. Kosala and H. Blockeel, "Web Mining Research A Survey", *Proceedings of ACM Conference on Explorations Newsletter, pp. 1-15, 2000.*
[13] Rinkle Goradia, "Web Mining to Detect Online Spread of Terrorism", *International Journal of Engineering Research and Technology*, Vol. 9, No. 7, pp. 645-648, 2020.
[14] Hilala Jafarova and Rovshan Aliyev, "Applying K-Means Clustering Algorithm using Oracle Data Mining to Banking Data", Springer, 2015.
[15] Hamid Bekamiri, "A New Model to Identify the Reliability and Trust of Internet Banking Users Using Fuzzy Theory and Data-Mining", *Mathematics*, Vol. 9, No. 9, pp. 916-927, 2021.
[16] Luiza Antonie and Alexandru Coman, "Application of Data Mining Techniques for Medical Image Classification", *Proceedings of International Conference on Mining Multimedia and Complex Data*, pp. 94-101, 2001.
[17] M. Muzammal, "A Multi-Sensor Data Fusion Enabled Ensemble Approach for Medical Data from Body Sensor Networks", *Information Fusion*, Vol. 53, pp. 155-164, 2020.
[18] Vikas Chaurasia and Saurabh Pal, "A Novel Approach for Breast Cancer Detection using Data Mining Techniques", *International Journal of Innovative Research in Computer and Communication Engineering*, Vol. 2, No. 1, pp. 1-14, 2014.
[19] Mohammed Loey, "Breast and Colon Cancer Classification from Gene Expression Profiles using Data Mining Techniques", *Symmetry*, Vol. 12, No. 3, pp. 408-423, 2020.
[20] M.S. Singh and P. Choudhary, "Stroke Prediction using Artificial Intelligence", *Proceedings of International Conference on Industrial Automation and Electromechanical Engineering*, pp. 158-161, 2017.
[21] Ema Utami and Suwanto Raharjo, "Mortality Prediction using Data Mining Classification Techniques in Patients with Hemorrhagic Stroke", *Proceedings of International Conference on Cyber and IT Service Management*, pp. 2222-2226, 2020.
[22] M. Shahbaz, S. Ali, and A. Umer, "Classification of Alzheimer's Disease using Machine Learning Techniques", *Proceedings of International Conference on Mining and Multimedia Data*, pp. 296-303, 2019.
[23] Al Hagery, Mohammed Abdullah, Ebtehal Ibrahim Al Fairouz and Norah Ahmed Al Humaidan. "Improvement of Alzheimer Disease Diagnosis Accuracy using Ensemble Methods", *Indonesian Journal of Electrical Engineering and Informatics*, Vol. 8, No. 1, pp. 132-139, 2020.
[24] M.S. Nair and U.K. Pandey, "A Study of Cataract Patient Data using C5", *Proceedings of International Conference on ICT Systems and Sustainability*, pp. 407-414, 2020.
[25] N. Mishra and J.M. Samuel, "Towards Integrating Data Mining with Knowledge-Based System for Diagnosis of Human Eye Diseases: The Case of an African Hospital", *Handbook of Research on Disease Prediction Through Data Analytics and Machine Learning*, pp. 470-485, 2021.
[26] A.S. Albahri, "Role of Biological Data Mining and Machine Learning Techniques in Detecting and Diagnosing the Novel Coronavirus (Covid-19): A Systematic Review", *Journal of Medical Systems*, Vol. 44, pp. 1-11, 2020.
[27] L.J. Muhammad, "Predictive Data Mining Models for Novel Coronavirus (COVID-19) Infected Patients' Recovery", *SN Computer Science*, Vol. 1, No. 4, pp. 1-7, 2020.
[28] Shruti Srivatsan and T. Santhanam, "A Study on Caesarean Section Prediction using ID3 Decision Tree Classifier", *Proceedings of International Virtual Conference on Computational Intelligence and Applications*, pp. 310-317, 2020.
[29] Tessy Badriyah, "Application of Naive Bayes Method for IUGR (Intra Uterine Growth Restriction) Diagnosis on The Pregnancy", *Proceedings of International Conference on Electrical, Communication, and Computer Engineering*, pp. 1-13, 2020.
[30] S. Perveen and M. Shanbhaz, "Performance Analysis of Data Mining Classification Techinques to Direct Diabetes", *Procedia Computer Science*, Vol. 82, pp. 115-121, 2016.
[31] H. Das, B. Naik and H.S. Behera, "Classification of Diabetes Mellitus Disease (DMD): A Data Mining (DM) Approach", *Proceedings of International Conference on* Progress in Computing, Analytics and Networking. Advances in Intelligent Systems and Computing, pp. 1-13, 2018.
[32] Tarig Mohamed, "Developing A Predicted Model for Diabetes Type 2 Treatment Plans by using Data Mining", *Journal of Theoretical and Applied Information Technology*, Vol. 90, No. 2, pp. 181-192, 2016.
[33] K. Shweta, Aishwarya Raj and Girija Attigeri, "Comparative Analysis of Prediction Algorithms for Diabetes", *Proceedings of International Conference on Advances in Computer Communication and Computational Sciences*, pp. 331-337, 2019.
[34] T. Yang, L. Zhang, L. Yi and H. Feng, "Ensemble Learning Models Based on Noninvasive Features for Type 2 Diabetes Screening: Model Development and Validation", *JMIR Medical Informatics*, Vol. 8, No. 6, pp. 1-13, 2020.
[35] Shruti Srivatsan and T. Santhanam, "A Comparison of Feature Selection Techniques and Ensemble Classifiers for Early Diabetes Prediction", *Proceedings of International Conference on AI, Robotics and Automation*, pp. 132-137, 2020.
[36] B. Pranto, S.M. Mehnaz, E.B. Mahid and I.M. Sadman, "Evaluating Machine Learning Methods for Predicting Diabetes among Female Patients in Bangladesh", *Information*, Vol. 11, pp. 374-386, 2020.
[37] Z. Xu and Z. Wang, "A Risk Prediction Model for Type 2 Diabetes Based on Weighted Feature Selection of Random Forest and XGBoost Ensemble Classifier", *Proceedings of International Conference on Advanced Computational Intelligence*, pp. 278-283, 2019.

[38] F. Faisal, S. Asaduzzaman and H. Minhaz, "Predicting Diabetes Mellitus and Analysing Risk-Factors Correlation", *EAI Endorsed Transactions on Pervasive Health and Technology*, Vol. 8, No. 5, pp. 1-7, 2019.

[39] A.A. Fareeha, Qurat-Ul-Ain and Y.E. Muhammad, "Comparative Analysis on Diagnosis of Diabetes Mellitus using Different Approaches - A Survey", *Informatics in Medicine*, Vol. 21, pp. 1-21, 2020.

[40] S. Rahman and H.Y. Bushra, "Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques", *Proceedings of International Conference on Computer Vision and Machine Intelligence in Medical Image Analysis*, pp. 551-559, 2020.

[41] M.M. Faniqul, "Early-Stage Diabetes Risk Prediction Dataset", Available at https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset, Accessed at 2019.

[42] D.F. Williamson, Robert A. Parker and Juliette S. Kendrick, "The Box Plot: A Simple Visual Method to Interpret Data", *Annals of Internal Medicine*, Vol. 110, No. 11, pp. 916-921, 1989.

[43] Mert Akyol, "Clustering Hotels and Analyzing the Importance of Their Features by Machine Learning Techniques", *Bilgisayar Bilimleri Ve Teknolojileri Dergisi*, Vol. 1, No. 1, pp. 16-23, 2016.

[44] S. Visalakshi and V. Radha, "A Literature Review of Feature Selection Techniques and Applications: Review of Feature Selection in Data Mining", *Proceedings of* IEEE International Conference on Computational Intelligence and Computing Research, pp. 1-6, 2014.

[45] Sayank Paul, "Beginner's Guide to Feature Selection in Python", Available at https://www.datacamp.com/community/tutorials/feature-selection-python, Accessed at 2021.

[46] Dheeb Albashish, "Binary Biogeography-Based Optimization based SVM-RFE for Feature Selection", *Applied Soft Computing*, Vol. 101, pp. 1-18, 2021.

[47] Jason Brownlee, "Recursive Feature Elimination (RFE) for Feature Selection in Python", Available at https://machinelearningmastery.com/rfe-feature-selection-in-python/, Accessed at 2020.

[48] Dario Radecic, "Feature Selection in Python - Recursive Feature Elimination", Available at https://towardsdatascience.com/feature-selection-in-python-recursive-feature-elimination-19f1c39b8d15, Accessed at 2020.

[49] R.E. Schapire, "The Boosting Approach to Machine Learning: An Overview", *Nonlinear Estimation and Classification*, Vol. 12, No. 2, pp. 149-171, 2003.

[50] J. Yoon, "Forecasting of Real GDP Growth using Machine Learning Models: Gradient Boosting and Random Forest Approach", *Computational Economics*, Vol. 57, No. 1, pp 247-265, 2021.

[51] Mahmoud Abbasi, and Oystein Haugen, "Boosting Algorithms for Network Intrusion Detection: A Comparative Evaluation of Real AdaBoost, Gentle AdaBoost and Modest AdaBoost", *Engineering Applications of Artificial Intelligence*, Vol. 94, pp. 1-16, 2020.

[52] R.E. Schapire, "The Strength of Weak Learnability", *Machine Learning*, Vol. 5, No. 2, pp. 197-227, 1990.

[53] Jason Brownlee, "Boosting and AdaBoost for Machine Learning", Available at https://machinelearningmastery.com/boosting-and-adaboost-for-machine-learning/, Accessed at 2020.

[54] Zulaikha Lateef, "A Comprehensive Guide to Boosting Machine Learning Algorithms", Available at https://www.edureka.co/blog/boosting-machine-learning/, Accessed at 2020.

[55] *Tianqi Chen and Guestrin, Carlos, "XGBoost: A Scalable Tree Boosting System", Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794, 2016.*

[56] Wei Dong, "XGBoost Algorithm-based Prediction of Concrete Electrical Resistivity for Structural Health Monitoring", *Automation in Construction*, Vol. 11, No. 1, pp. 1-14, 2020.

[57] A.B. *Parsa, A. Movahedi and H. Taghipour, "Toward Safer Highways, Application of Xgboost and Shap for Real-Time Accident Detection and Feature Analysis", Accident Analysis and Prevention, Vol. 136, pp. 1-18, 2020.*

[58] Tianqi *Chen, "Xgboost: Extreme Gradient Boosting", Available at https://cran.r-project.org/web/packages/xgboost/vignettes/xgboost.pdf, Accessed at 2021.*

[59] *Aarshay Jain, "Complete Guide to Parameter Tuning in XGBoost with codes in Python", Available at* https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/, Accessed at 2021.

[60] Jason Brownlee, "A Gentle Introduction to Xgboost for applied machine learning", Available at https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/, Accessed at 2021.

[61] Guolin Ke, Qi Meng, Thomas Finley and Taifeng Wan, "Light GBM: A Highly Efficient Gradient Boosting Decision Tree", *Advances in Neural Information Processing Systems*, Vol. 23, No. 1, pp. 1-16, 2017.

[62] B. Shaker, "Light BBB: Computational Prediction Model of Blood-Brain-Barrier Penetration based on Light GBM", *Bioinformatics*, Vol. 37, No. 8, pp. 1135-1139, 2020.

[63] Mingxi Liu and Zeqian Sima, "A Novel Cryptocurrency Price Trend Forecasting Model based on Light GBM", *Finance Research Letters*, Vol. 32, pp. 1-19, 2020.

[64] Pranjal Khandelwal, "Which Algorithm takes the Crown: Light GBM vs XGBOOST?", Available at https://www.analyticsvidhya.com/blog/2017/06/which-algorithm-takes-the-crown-light-gbm-vs-xgboost/, Accessed at 2021.

[65] Pushkar Mandot, "What is LightGBM, How to Implement it? How to Fine Tune the Parameters?", Available at https://medium.com/@pushkarmandot/https-medium-com-pushkarmandot-what-is-lightgbm-how-to-implement-it-how-to-fine-tune-the-parameters-60347819b7fc, Accessed at 2021.

[66] Sarang Narkhede, "**Understanding Confusion Matrix", Available at** https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62, Accessed at 2021.

[67] Yasen Jiao and Pufeng Du, "Performance Measures in Evaluating Machine Learning based Bioinformatics Predictors for Classifications", *Quantitative Biology*, Vol. 4, No. 4, pp. 320-330, 2016.

[68] D. Erika, "Accuracy, Recall and Precision", Available at https://medium.com/@erika.dauria/accuracy-recall-precision-80a5b6cbd28d, Accessed at 2021.

[69] Jason Brownlee, "A Gentle Introduction to the Fbeta-Measure for Machine Learning", Available at https://machinelearningmastery.com/fbeta-measure-for-machine-learning/#:~:text=A%20default%20beta%20value%20is,the%20calculation%20of%20the%20score, Accessed at 2021.

[70] Y. Pang, "Multi-Scale Interactive Network for Salient Object Detection", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1-13, 2020.

[71] Bharathi Raja and Vigneshwaran Muralidaran, "Findings of the shared task on Hope Speech Detection for Equality, Diversity, and Inclusion", *Proceedings of 1st Workshop on Language Technology for Equality, Diversity and Inclusion*, pp. 1-12, 2021.

[72] D. Devarriya, "Unbalanced Breast Cancer Data Classification using Novel Fitness Functions in Genetic Programming", *Expert Systems with Applications*, Vol. 140, pp. 1-15, 2020.

[73] Parul Pandey, "Simplifying the ROC and AUC Metrics", Available at https://towardsdatascience.com/understanding-the-roc-and-auc-curves-a05b68550b69, Accessed at 2021.

[74] Mary L. McHugh, "Interrater Reliability: The Kappa Statistic", *Biochemia Medica*, Vol. 22, No. 3, pp. 276-282, 2012.

[75] P.K. Chan and S.J. Stolfo, "On the Accuracy of Meta-learning for Scalable Data Mining", *Journal of Intelligent Information Systems*, Vol. 8, pp. 5-28, 1997.

[76] Lakshmi Devasena, "Effectiveness Analysis of Zero R, RIDOR and Part Classifiers for Credit Risk Appraisal", *International Journal of Advances in Computer Science and Technology*, Vol. 3, No. 2, pp.6-11, 2014..