International Journal of Advanced Technology and Engineering Exploration, Vol 9(90)
ISSN (Print): 2394-5443 ISSN (Online): 2394-7454
http://dx.doi.org/10.19101/IJATEE.2021.875402

Research Article

Efficient ensemble machine learning techniques for early prediction of diphtheria diseases based on clinical data

Bilal Abdualgalil[1*], Sajimon Abraham[1] and Waleed M. Ismael[2]
School of Computer Sciences, Mahatma Gandhi University, Kottayam, Kerala, India[1]
Hohai University, Chanzhou campus, Jiangsu, China[2]

Received: 09-December-2021; Revised: 08-May-2022; Accepted: 11-May-2022

Abstract
Diphtheria is a worldwide concern, particularly in Yemen. Early detection is important for reducing diphtheria deaths. In fact, proper diphtheria diagnosis takes time due to various clinical examinations. This problem requires the development of a new diagnostic system. With machine learning (ML) techniques, continuing to be proposed, ensemble learning techniques have been introduced into healthcare applications. Efficient ensemble ML techniques (EEMLT) are used to develop prediction models for diphtheria disease in this study. Five ensemble ML models i.e., random forest classifier (RFC), gradient boosting classifier (GBC), extra tree classifier (ETC), eXtreme gradient boosting (XGB), and light gradient boosting machine (LightGBM) were used. Moreover, five popular baseline classifiers, i.e., logistic regression (LR), k-nearest neighbors (KNN), support vector classifier (SVC), decision tree classifier (DTC), multilayer perceptron (MLP), were used as benchmarks. All ensemble and baseline classifiers are trained and tested in the dataset using 10-fold cross-validation (CV) and holdout CV approaches. All models were evaluated on a test set using different metrics including accuracy, F1-sore, Recall, Precision, and area under curve (AUC) measures. According to the results of this study, the ETC model achieved high accuracy with 98.92% and 99.2% in holdout and 10-fold CV, respectively. It is found that the ETC achieved high accuracy of 99.2% in 10-fold and holdout CV approach. Finally, the experimental results reveal that the performance of ensemble classifiers has outperformed those of baseline classifiers. We believe that the proposed diphtheria prediction system will help doctors accurately predict diphtheria disease.

Keywords
Ensemble machine learning, Baseline classifiers, Diphtheria disease, SMOTE+ENN, Multiclass classification.

1.Introduction
An outbreak of diphtheria, declared in Yemen in October 2017, is ongoing [1]. Diphtheria is an infection caused by the bacteria Corynebacterium diphtheria. It usually starts in the upper respiratory tract and spreads to other parts of the body because of the spread of the bacterial toxin. In the temperate parts of the world, this disease still happens the most. It is more common in the colder months of the year and mainly affects children under the age of 10 [2]. The number of reported diphtheria cases around the world has been gradually growing. There were 16,651 recorded cases in 2018, which was more than double the average number from 1996 to 2017 (8,105 cases) [3].

Diphtheria symptoms include swollen glands (enlarged lymph nodes) in the neck, trouble breathing or rapid breathing, nasal discharge, weariness, fever, and chills, in addition to a thick, gray membrane covering the throat and tonsils and a sore throat [4].

It is critical to obtain an accurate diagnosis as soon as possible because failure to receive specific therapy may result in death.

Diphtheria early detection is not without its drawbacks, such as the fact that it takes a long time to correctly diagnose diphtheria [5] due to the vast number of clinical exams required. Consequently, early detection of diphtheria is necessary for doctors to choose the best treatment for patients. This is one of the challenges still faced by them. As a result, efficient ensemble machine learning techniques (EEMLT) has recently emerged as a widely used diagnostic method. By using clinical data, these

*Author for correspondence

approaches are able to develop an automated mechanism for detecting diphtheria on its own.

Also, in medical data such as epilepsy [6, 7], neuromuscular diseases [8, 9], heart rhythms [10, 11], etc., Machine learning (ML) classifiers are very effective in interpreting such diseases. Furthermore, studies have proven that ML techniques are also effective in predicting clinical data such as biomedical studies [12, 13], viral disease [14], and cancer [15]. These techniques also work well for predicting diphtheria diseases; however, there is still a challenge in using ML techniques to predict diphtheria diseases based on clinical data that is mostly imbalanced, as well as the selection of important findings; all of these factors affect the accuracy of models in predicting diphtheria diseases.

To our knowledge, no study has yet used artificial intelligence approaches for the rapid diagnosis of diphtheria, and this work requires the use of effective ML techniques to improve results.

The objective of this study is to perform a comparative analysis of five EEMLT for the early prediction of diphtheria diseases and compare them with five baseline classifiers. The primary contributions of this study are:

- Implementing the prediction system for diphtheria disease based on efficient ensemble ML.
- Identifying the most precise and efficient ML method for predicting diphtheria disease by comparing five efficient techniques.
- Carrying out a comparison between the ensemble classifiers and baseline classifiers to identify the efficiency of ensemble classifiers in prediction.

This paper is organized as follows: Section 2 presents the literature review on diphtheria disease prediction. Section 3 explains the proposed methodology. The results of the study are described in section 4 and discussed in section 5. Section 6 concludes with a conclusion and suggestions for future research.

2.Literature review

There are few recent studies on the diphtheria disease because it only emerged at the end of 2017 and the beginning of 2018.

Anggraeni et al. [16] used the radial basis function neural network (RBFNN) for forecasting the diphtheria case number in Indonesia, this method achieved good performance for forecasting in Malang, Surabaya, and Sumenep with mean absolute scaled error (MASE) values of 0.84, 0.817, and 0.820, of which all MASE values are less than 1.

Park et al. [17] developed and validated ML models for the classification of carpal tunnel syndrome (CTS) severity. The CTS was multiclass classified into three grades: 507 mild, 276 moderates, and 254 severe, this study achieved 76.6% accuracy with the eXtreme gradient boosting (XGB) model compared with random forest (RF) and K-nearest neighbors (KNN).

Zhang et al. [18] proposed a number of imputation algorithms based on different classifiers (naive Bayes (NB), KNN, decision tree (DT), and multilayer perceptron (MLP)) to process missing values in their clinical heart failure (HF) dataset. The results of a study show there is no universal imputation technique that outperforms all other classifiers.

The thyroid gland dataset, which consists of 215 samples and three classes, was classified by Diri and Albayrak [19]. It has been classified using four different ML classifiers, including NB, k-NN, k-Means, and 2-D SOM (hyperthyroid 35 samples, hypothyroid 30 samples, and euthyroidism 150 samples). The results of the study were 95.83 % for NB, 91.67 % for KNN, 84.72 % for 2-D SOM, and 72 % for k-Means. NB classifier achieved better accuracy than K-NN, 2-D SOM, and k-means.

To improve the accuracy of predicting cardiovascular disease, Mohan et al. [20] proposed hybrid HRFLM approaches that combine RF and linear method (LM) characteristics. The Cleveland dataset used in the study has 13 features, 303 samples, and is multiclass (0-4) variable represents patients, with the scaling referring to disease severity (4 being the highest). HRFLM was found to be quite accurate in predicting heart disease.

Chaudhary et al. [21] presented an improved random forest classifier (RFC) technique for multiclass disease classification problems using five benchmark datasets. The improved model combines an RFC ML approach, an attribute evaluator method, and an instance filter method. The proposed model in the study outperforms the RFC with 97.80% accuracy.

Jacob and Ramani [22] presented data mining algorithms to classify breast tissue data. The Wisconsin Breast tissue dataset was obtained from the UCI ML Repository, it contains 11 attributes and 106 samples. The dataset multiclass classification

contains 6 classes. This study achieved 100% accuracy in the random tree algorithm.

Altaf et al. [23] proposed a hybrid feature extraction method that would utilize magnetic resonance imaging (MRI) and clinical data to automatically classify Alzheimer's disease. The sample consists of 287 subjects, divided into three output classes: MCI (105 subjects), AD (92 subjects), and norm (90 subjects). Based on key evaluation criteria such as accuracy, the proposed algorithm outperforms state-of-the-art techniques by 79.8%.

Iqbal and Islam [24] proposed a dengue prediction machine learning model. Dengue fever is one of the most well-known viral illnesses in humans. More than 33% of the world's population are at risk, including numerous cities in India.

Yang and Man [25] suggested an improved feature selection method based on feature item length information. The suggested method reduces the importance of infrequent feature items while emphasizing the importance of negative features in the classification.

Ucar et al. [26] suggested a new hybrid ML approach for diagnosing tuberculosis disease. An adaptive neuro-fuzzy inference system and rough sets are combined with the proposed ML method. The results show that the proposed technique shows more viable results than the other algorithms.

Fariza et al. [27] used the hierarchical clustering method to provide a new method to diphtheria risk analysis in Surabaya based on multiple parameters, including (Diphtheria, Tetanus Pertussis) immunization, number of diphtheria sufferers, and population density. The 2019 predictions reveal that using a single linkage rather than an average or complete linkage resulted in a lower diphtheria risk level of susceptibility with the least variance of 4.43 10-5. This demonstrates very good clustering results.

Agglomerative hierarchical clustering was used by Singh et al. [28] for the purpose of locating clusters in a patient population. According to the results of a study, clustering revealed the presence of nine different and clinically relevant cohort/multi-morbidity groups in the patient population.

Fatoni et al. [29] used the KNN approach to create an expert system for diagnosing diphtheria. This method compares the similarity of each diphtheria symptom to provide an early diphtheria diagnosis. The accuracy of this study's diphtheria diagnosis was 93.056%.

Chumachenko et al. [30] developed an intelligent multiagent model of the dynamics of diphtheria infectious morbidity spreading. The developed intelligent multiagent model overcomes the limitations of previous epidemic dynamics models.

As per the literature review, most of the studies on disease prediction did not use techniques to rebalance data to obtain the best generalizable models for the prediction of diseases, because healthcare data are often imbalanced, and contain missing values as well as unimportant features (e.g., name and address), which often are processed in the pre-processed stage. These issues are attributed to generalization and data imbalance. Making most disease prediction models biased toward a specific disease. Therefore, there are important research gaps in the healthcare prediction system: missing values, features selections, and data rebalancing. Based on dataset from the Epidemiological Surveillance Sector of the Ministry of Public Health and Population in Sana'a, Yemen, this study aims to fill these gaps. Improvement of the problem of imbalance and generalization by using ML or deep learning techniques for the early detection of diphtheria disease. This was accomplished in this study through the use of ML techniques. Based on clinical data, our proposed classifier will be effective in predicting any disease.

3.Proposed methodology

Predicting diphtheria disease is the purpose of this study, as illustrated in *Figure 1*. The proposed system includes the following steps:

1) Obtaining the diphtheria dataset,
2) Pre-processing and cleaning the obtained dataset before being used to build models,
3) A cleaned dataset is divided into training and testing sets
4) Applying five ensemble ML algorithms to build predictive models for predicting diphtheria,
5) Feeding the testing set into the model to evaluate its performance.
6) Applying the built model to predict diphtheria at this point,
7) Evaluating the obtained results from all algorithms, and
8) Carrying out a comparison to determine the best algorithm.

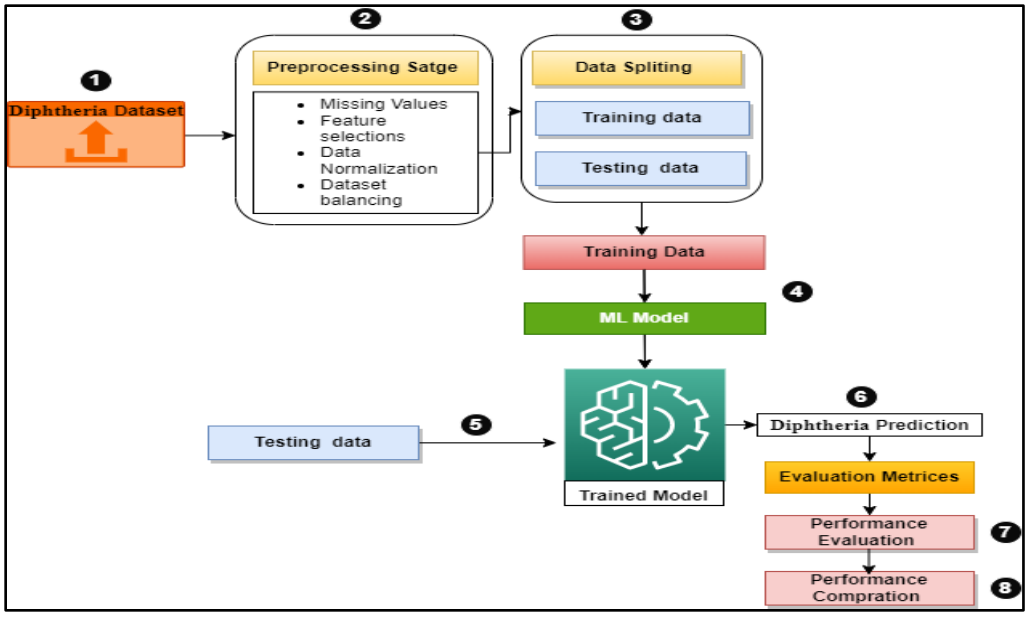

Figure 1 The proposed system for diphtheria disease prediction

3.1Experimental setup

Using Python 3.7.1 software, we implemented our experimental models on Windows 10 OS, running on an Intel® Core(TM) i5- 8250U CPU 1.80 GHz 4 processor, 8-GB-RAM, and a 2-TB hard drive.

3.2Dataset

In this study, the dataset was obtained from the Epidemiological Surveillance Sector of the Ministry of Public Health and Population, Sana'a, Yemen [31]. The clinical data were collected during the 2019 year. The dataset contains 2032 samples and 11 features. *Table 1* shows the distribution of clinical features. *Figure 2* represents the histogram of value distribution of all features in the diphtheria dataset.

Table 1 Normalized values of different attributes of diphtheria data set

No.	Feature name	Values	Missing values
1	sex	{1=M, 0=F}	0
2	Age (Years)	Continuous	19
3	Throat swab taken (TST)	1 = Yes, 0 = No	466
4	Pseudomembrane (PM)	1 = Yes, 0 = No	0
5	Difficulty of Breathing (DOB)	1 = Yes, 0 = No	0
6	Difficulty of Swallowing (DOS)	1 = Yes, 0 = No	0
7	Cervical L.N. Swelling (CLNS)	1 = Yes, 0 = No	0
8	Upper respiratory tract infection (URTI)	1 = Yes, 0 = No	0
9	Total number of diphtheria vaccine received (Penta/DPT/DT) (TNODV)	Zero dose: 0 1 dose : 1 3 dose : 3 2 dose : 2	0
10	Treatment Received (TR)	1= antibiotic & antitoxin (ABAT) 2= antibiotic alone (AB) 3= antitoxin alone(AT) 0 = no treatment	2
11	Class : C.diphtheriae culture result	{ 0= no corynebacterium diphtheria isolated 1= corynebacterium diphtheria isolated 2= Positive C.diphtheriae Isolated 3= No C\S from C.D. Center 4= No C.diphtheriae Isolated }	1402

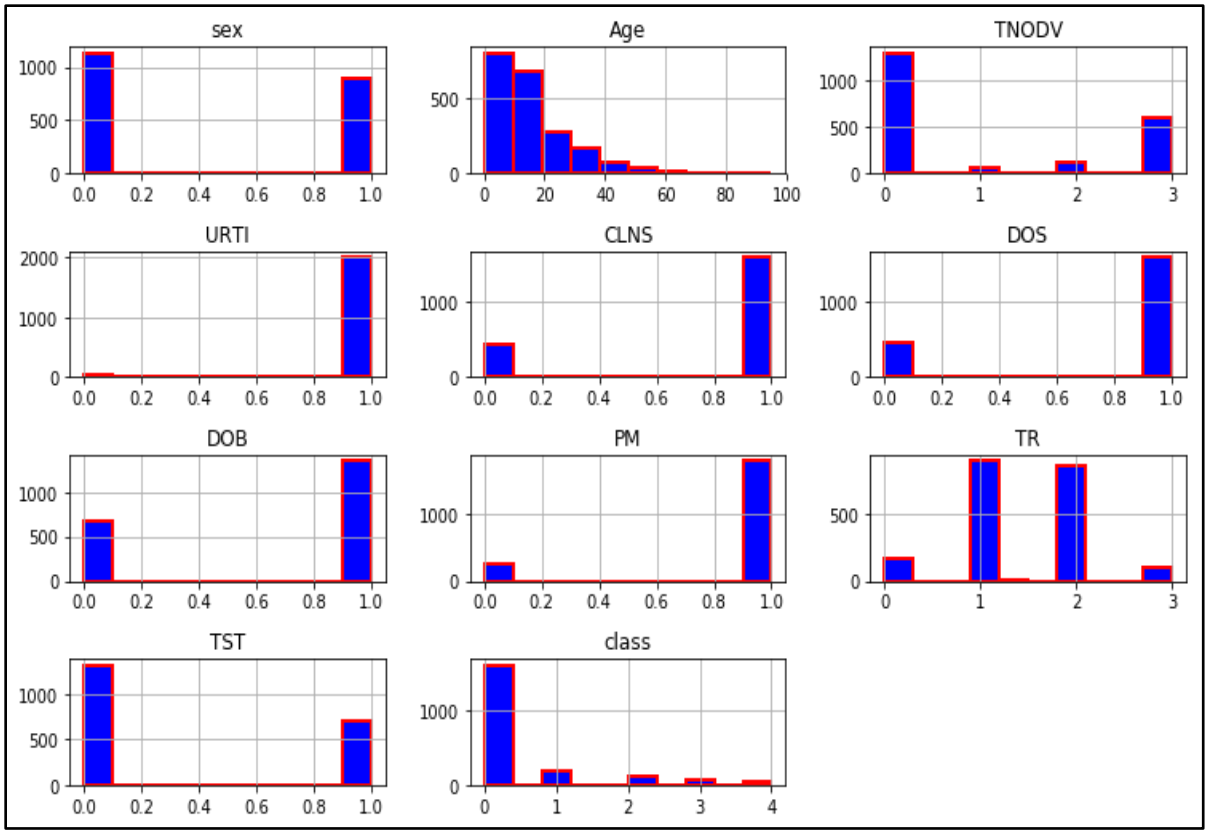

Figure 2 Histogram representation of the feature value distribution

3.3Pre-processing

Data pre-processing and cleaning are important steps in handling data before it is used in ML algorithms. The diphtheria dataset can be downloaded in .xlsx format. This stage consists of a series of steps which are as under:

3.3.1Missing data handling

Missing values and noise are present in real-world data, which is also in a raw format that cannot be directly used to build ML models. Data pre-processing processes, such as data cleaning and formatting, are required to turn such noisy data into a machine-understandable format. The handling of missing data was the initial stage in data pre-processing.

In this dataset, we found that the age feature had 19 missing values, TST had 466 missing values, TR feature had 2 missing values, and class had 1402 missing values as shown in *Figure 3(a).* The mean and mode methods were used to replace the missing values for handling these features. We replaced missing values in Age and TR features with mean values, while TST and class features, the missing was replaced by the Mode values as shown in *Figure 3(b).*

3.3.2Feature selection

The two-dimensional ranking of diphtheria dataset features using Pearson rank correlations [32] as shows in *Figure 4* to demonstrate how highly the features correlated with the diagnosis results, which was used to select the features for the dataset. It shows that the features are highly correlated with diagnostic results by the darker red color. In the end, 11 features were included in the final dataset, TST, TR, PM, DOB, DOS, CLNS, URTI, TNODV, Age, sex, and class.

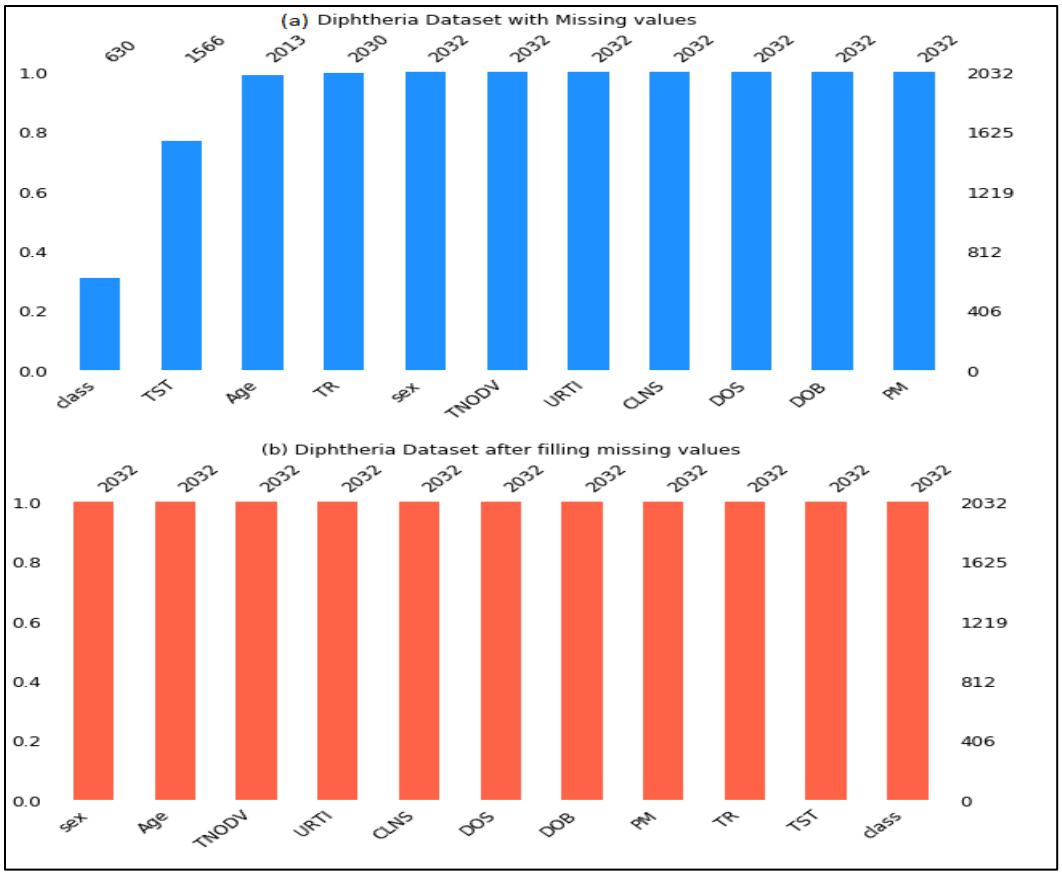

Figure 3 Diphtheria dataset (a) diphtheria dataset with missing values and (b) diphtheria dataset after filling missing values

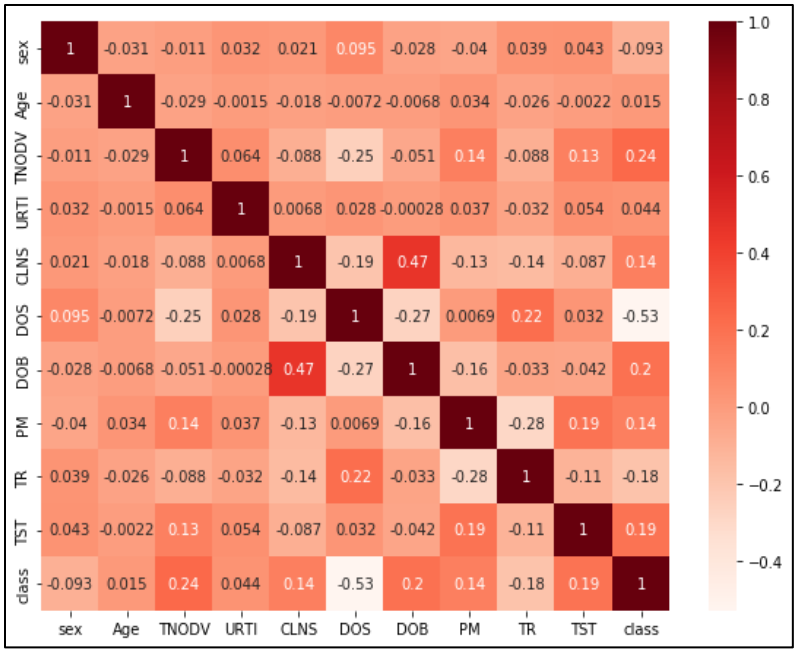

Figure 4 Pearson-ranking visualization of the diphtheria dataset after processing of missing values

3.3.3Data normalization

Normalization is a way used to bring all qualities under a single scale of minimum, maximum, and medium values without distorting the values within them. The diphtheria dataset is normalized using Z-scores [33]. This method normalizes input feature vectors using the mean and standard deviation of each feature. In Equation (1), Z is the normalized feature value, x_i is the original feature value, σ is the standard deviation, and μ is the mean.

$$Z = \frac{x_i - \mu}{\sigma} \quad (1)$$

3.3.4Data balancing

Using an imbalanced dataset to train ML models can lead to a bias towards the majority class. To avoid this bias, it is necessary to use a more balanced dataset. In this work, the SMOTE+ENN hybrid technique is used to rebalance the dataset. It was developed by [34]. It is a hybrid of the synthetic minority oversampling technique (SMOTE) and edited nearest neighbours (ENN) technique. SMOTE is the most popular oversampling technique and can be combined with many different under-sampling techniques. A random selection of minority class examples is selected by SMOTE. To build a synthetic example, we took the sample's nearest k neighbours and randomly selected a point inside that region.

ENN works by selecting examples to be deleted. This rule entails locating and deleting misclassified examples in a dataset using k=3 nearest neighbors. After applying the SMOTE+ENN hybrid technique, the new balanced dataset becomes as the following: Resampled dataset shape Counter ({0.0: 1246, 3.0: 618, 1.0: 615, 4.0: 212, 2.0: 96}), as shown in *Figure 5 (b)* compared with the original dataset which was as the following :

Original dataset shape Counter ({0.0: 1618, 1.0: 192, 2.0: 112, 3.0: 64, 4.0: 46), as shown in *Figure 5 (a).* We also implemented SMOTE and its other extensions, such as SMOTE + Tomek and adaptive synthetic sampling (ADASYN). The best results, however, were obtained by combining SMOTE with an ENN modification.

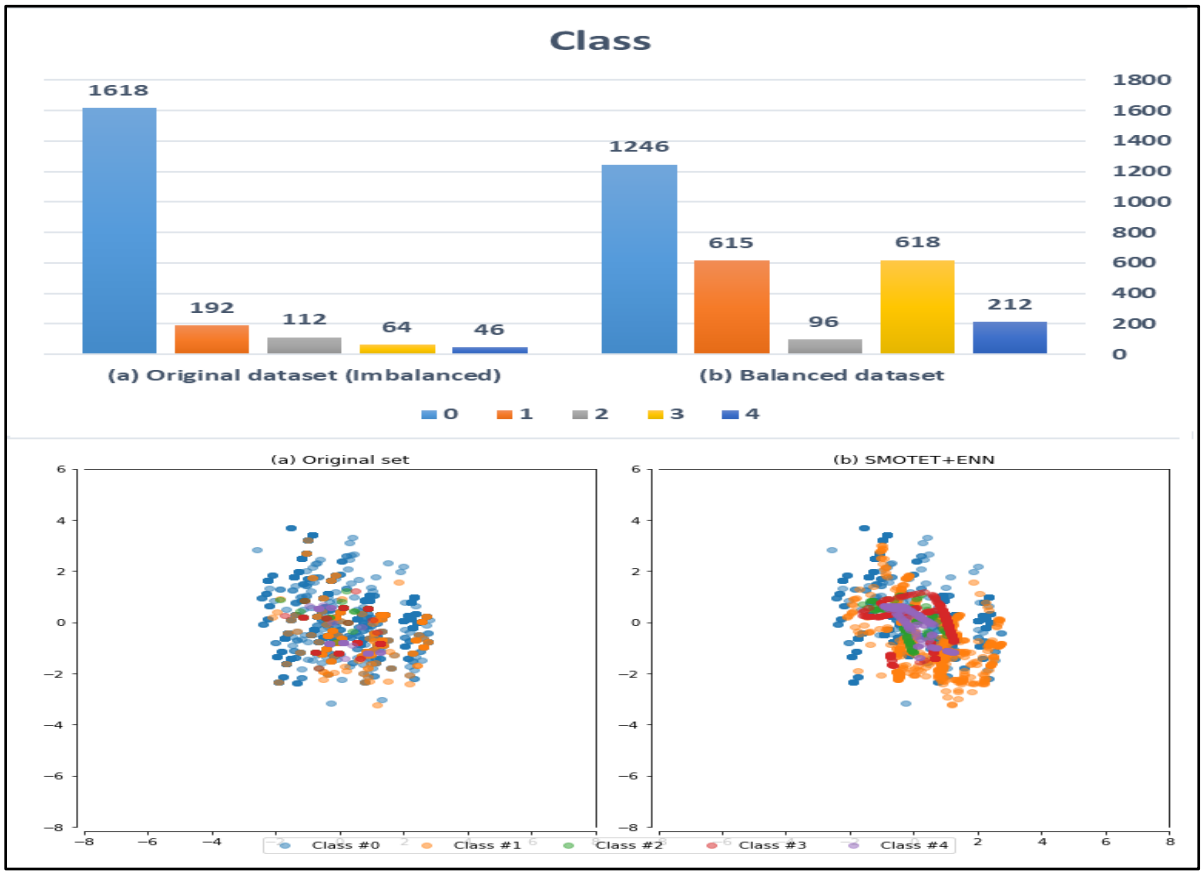

Figure 5 Dataset (a) Original (Imbalanced) dataset, and (b) dataset after SMOTE+ENN hybrid technique (Balanced data set)

3.3.5Data splitting

In the stage of data splitting, we used two methods to split the dataset into a training set and a testing set after preprocessing. The first method is holdout cross-validation (CV), in which we divided the data set into 70% for the training set and 30% for the testing set, and the second method is 10-fold CV. The training data is fed into the ensemble ML model to train the model. The diphtheria class (Class: C.diphtheriae culture result) feature is used as the target variable in the prediction classifier.

3.4Methods

The ensemble method combines two or more classification algorithms to improve or boost overall performance. Bagging and Boosting are the two most ensemble-based strategies. Boosting follows a sequential process in which the subsequent model corrects the previous model's errors while Bagging works by combining the results of multiple models to get the final result. Five EEMLT used in this study are: RFC, gradient boosting classifier (GBC), extra tree classifier (ETC), XGB, and light gradient boosting machine (LightGBM). Five baseline ML models used in this study are LR, KNN, support vector classifier (SVC), decision tree classifier (DTC), and MLP.

RFC [35] is a decision tree-based classification algorithm. In uncorrelated forests, the algorithm builds each tree randomly to promote accurate decision-making. GBC [36] combines each weak learning model to generate a strong predictive model. Gradient boosting frequently use decision trees. ETC [37] is a decision tree-based ensemble learning approach. Random Forest's Extra Trees Classifier randomises certain decisions and data sets to avoid overfitting and over-learning from the data. And XGB [38] is a gradient boosting decision-tree-based ensemble ML algorithm. Unstructured data prediction problems (images, text, etc.). Wide range of uses like user-defined prediction and regression problems. LightGBM [39] is an efficient Gradient Boosting Decision Tree (GBDT). Gradient-based one side sampling (GOSS) and exclusive feature bundling (EFB) are employed to overcome the constraints of the histogram-based technique utilized in all GBDT frameworks and EFB. The GOSS and EFB approaches form the LightGBM Algorithm's properties. They work together to make the model operate efficiently and outperform rival GBDT frameworks. In LR [40] classifier, independent variables (x) are used to predict one or more dependent data variables (y). KNN [41] is also known as a "lazy learner" due to its lengthy and limited training period. The training set is used to evaluate a new instance. The distance between the new instance and the training instances is measured, and the result is calculated based on the new instance's proximity to the training instances.

SVC [42] employs classification algorithms to solve two-group classification problems. After feeding an SVC model, they can categorize new text a set of labeled training data for each category. DTC [43] this algorithm creates a tree from the input dataset based on conditions. The tree is refined and made top-down. Conditions are used to build the branches. For example, if the dataset meets the condition, it is refined on the left branch.

MLP [44] is a fully connected layer model. The fully connected layer model structure includes three layers: input, hidden, and output layers, composed of the activation function, weights, and biases.

In this study, baseline classifiers were built using Python's Scikit-learn package. Whereas ensemble classifiers such as XGBoost and LightGBM were implemented using Python libraries "xgboost" and "lightgbm," which include classification, regression, and clustering tools for machine learning and modeling. To achieve maximum accuracy, users can fine-tune the classification parameter settings using the training methods included in the package. The Hyper-parameters settings are used to train each ensemble and baseline ML classifier as shown in *Table 2* in detail. Using the testing data, the model predicts diphtheria disease after training the classifiers.

Table 2 Setting the hyper-parameters for classification methods

No.	Model	Ensemble classifiers
1	RFC	criterion='entropy',n_estimators=25,max_depth=7,random_state=33
2	GBC	n_estimators=40,learning_rate=0.1,max_depth=5,random_state=33
3	XGB	learning_rate =0.75, n_estimators=1000, max_depth=3, min_child_weight=1, gamma=0.1, subsample=0.8, colsample_bytree=0.8, objective= 'multi:softprob', nthread=4, scale_pos_weight=1, seed=27

4	ETC	n_estimators=100, max_features=9
5	LightGBM	boosting_type='gbdt',n_estimators=1000, objective= 'multi_logloss', learning_rate=0.1
No.	**Model**	**Baseline classifiers**
1	LR	penalty='l2',solver='sag', C=1.0,random_state=33, multi_class='ovr'
2	KNN	n_neighbors= 3,weights ='uniform', algorithm='auto'
3	SVC	kernel= 'rbf', max_iter=1000,C=1,gamma=1,probability=True, decision_function_shape='ovr'
4	DTC	criterion='entropy',max_depth=3,random_state=33
5	MPL	activation='relu', solver='lbfgs', learning_rate='constant', early_stopping=True,alpha=0.0001 ,hidden_layer_sizes=(100,4),random_state=33

3.5Evaluation metrics

In supervised machine learning (SML), there are several methods for evaluating the performance of learning models. In this study, accuracy, recall, precision, F1 scores, AUC, and error rate (ER) metrics were used to evaluate all models, which are mathematically represented by the *Equations* 2-7 (*Table 3)*. Because the dataset contains multiclass classifications, the metrics for binary classification do not entirely apply in this study. N × N is the number of various classes C_0, C_1,..., C_N that is included in the multiclass confusion matrix (e.g., 5 classes in the diphtheria dataset in our study). Therefore, this cannot be classified as a true positive (TP), true negative (TN), false positive (FP), or false negative (FN). It is possible to analyze a specific type of data rather than the entire dataset using a multiclass confusion matrix. Based on this method, a set of metrics can be defined for each class. After then it's possible to provide metrics for the full matrix depending on this combination of metrics. Accuracy, recall, precision, and F1-score are defined metrics for a multiclass confusion matrix. In this study the evaluation metrics in *Table 3* were developed using a "macro" method.

Table 3 Evaluation metrics for a multiclass confusion matrix

Metric	Description	Formula	
Time(s)	Calculates the amount of time an ML model will take to execute.	---	
Accuracy (macro average) [45]	The average per-class effectiveness of the classifier	$Accuracy = \frac{\sum_{i=1}^{N} TP(C_i)}{\sum_{i=1}^{N}\sum_{j=1}^{N} C_{i,j}}$	(2)
Recall (macro average) [46]	Average per-class effectiveness of a classifier to identify the class label	$Recall(macro) = \frac{1}{N}\sum_{i=1}^{N} \frac{TP(C_i)}{TP(C_i)+FP(C_i)}$	(3)
Precision (macro average) [47]	Average per-class agreement of the true class labels with those of the classifier's	$Precision(macro) = \frac{1}{N}\sum_{i=1}^{N} \frac{TP(C_i)}{TP(C_i)+FN(C_i)}$	(4)
F1-score (macro average) [48]	Defined as the harmonic mean between precision (macro-average) and recall (macro-average)	$F1 - score(macro) = 2 * \frac{Precision(macro)\times Recall(macro)}{Precision(macro)+Recall(macro)}$	(5)
AUC - Area under the ROC curve (macro average) [49]	AUC score is used to determine which model best predicts classes. AUC is the relationship between true-positive rate and false positive rate	$AUC_{total} = \frac{2}{N(N-1)}\sum_{\{C_i,C_j\}\in N} AUC(C_i, C_j)$	(6)
ER [50]	The ER is calculated by dividing the total number of incorrect predictions on the test set by the total number of predictions on the test set. Because accuracy and ER are complements, we can always calculate one from the other.	$Error\ Rate(ER) = 100 -$ Accuracy (macro average)	(7)

4.Results

The experimental results of diphtheria prediction using five efficient ensemble ML models (RFC, GBC, ETC, XGB, and LightGBM), and five baseline ML models (LR, KNN, SVC, (DTC, and MLP) are presented. As shown in *Table 3*, all models were evaluated using the same metrics on the same dataset. *Table 4* and *Table 5*, show the performance of all

ensemble ML models evaluated using the holdout CV approach and the 10-fold CV approach respectively. *Table 6* and *Table 7* show the run time of all ensemble models used in the study in both holdout CV and the 10-fold CV approach.

This study used five metrics: accuracy, F1-Score, precision, recall, and AUC to compare prediction models' performance for a dataset in the testing phase. We considered all metrics with holdout CV and a 10-fold CV approach. The AUC values were calculated for all ensemble ML models with the holdout CV approach, as shown in *Figure 6*. And all ensemble ML models were evaluated with a 10-fold CV approach shown in *Figure 7*. Also, the Confusion matrix for all ensemble ML models was evaluated with the holdout CV approach and 10-fold CV approach, respectively, as shown in *Figure 8* and *9*.

The ETC model achieved 99.02% accuracy, 98.25% f1-score, 98.35% precision, 98.28% recall, 99.92% AUC, and 0.98 ER with the 10-fold CV approach. In case of holdout CV approach the results are 98.8% accuracy, 98.55% f1-score, 98.8% precision, 98.34% recall, 99.92% AUC, and 1.08 ER.

Finally, in both the 10-fold CV and holdout CV techniques, the ETC found to be efficient for diphtheria prediction based on EEMLT. Based on the results of the whole experiment, all ensemble ML approaches performed well in the prediction of diphtheria data. Moreover, ensemble ML techniques and baseline ML techniques in 10-fold CV and holdout CV approach were compared, as listed in *Table 8* and *Table 9* and visualized in *Figure 10* and *11* respectively. We conclude that ensembles ML techniques outperformed baseline ML techniques for diphtheria disease prediction.

Additional to this work, the performance of all ensemble and baseline ML techniques were compared in the case of original diphtheria data (imbalanced dataset before rebalanced) in both 10-fold, and holdout CV approaches, as shown in *Table 10* and *Table 11* respectively. The results of this comparison are that performance of all ensemble and baseline classifiers in an imbalanced dataset was much less in the balanced dataset case, as shown in *Figure 12* and *13* respectively

We conclude from this addition that all ML classifiers, whether ensemble or baseline classifiers work well performance for diseases prediction in the case of balanced medical data.

In this study, three multi-class imbalanced data sets (contraceptive, Yeast, and Shuttle from Keele's data repository [51]) are presented in *Table 12*. With the ETC model, the performance of the best ensemble ML model with different multiple class imbalanced datasets in both holdout and 10-fold CV approaches has achieved more than 90% accuracy. The results of the competition are presented in *Table 13* and shown in *Figure 14*. These results prove that model which achieved the best performance can also be generalized to different imbalanced medical data.

Table 4 All ensemble ML models were evaluated using the holdout CV approach

Model	Accuracy (%)	F1-Score (%)	Precision (%)	Recall (%)	AUC (%)	ER
RFC	95.1	94.53	96.62	92.75	99.55	4.9
GBC	97.96	96.55	97.04	96.11	99.75	2.04
XGB	98.44	98.4	98.9	97.95	99.87	1.56
ETC	**98.92**	**98.89**	**98.94**	**98.84**	**99.92**	**1.08**
LightGBM	98.44	97.93	98.28	97.6	99.94	1.56

Table 5 All ensemble ML models were evaluated using the 10-fold CV approach

Model	Accuracy (%)	F1-Score (%)	Precision (%)	Recall (%)	AUC (%)	ER
RFC	95.79	93.71	96.36	91.99	99.11	4.21
GBC	96.56	93.21	95.61	91.65	98.6	3.44
XGB	97.94	95.96	96.77	95.39	99.16	2.06
ETC	**99.02**	**98.29**	**98.43**	**98.28**	**99.41**	**0.98**
LightGBM	98.41	97.04	97.78	96.59	99.48	1.59

Table 6 Operating time across different classifiers (unit: second) with holdout CV

Ensemble classifiers	**RFC**	**GBC**	**XGB**	**LightGBM**	**ETC**
Time(s)	0.04	0.54	2.42	1.35	0.29
Baseline classifiers	**LR**	**KNN**	**SVC**	**DTC**	**MLP**
Time(s)	0.05	0.06	0.8	0.003	3.84

Table 7 Operating time across different classifiers (unit: second) with 10-fold CV

Ensemble classifiers	**RFC**	**GBC**	**XGB**	**LightGBM**	**ETC**
Time(s)	1.01	10.82	54.35	26.08	4.39
Baseline classifiers	**LR**	**KNN**	**SVC**	**DTC**	**MLP**
Time(s)	1.1	0.27	6.87	0.12	25.55

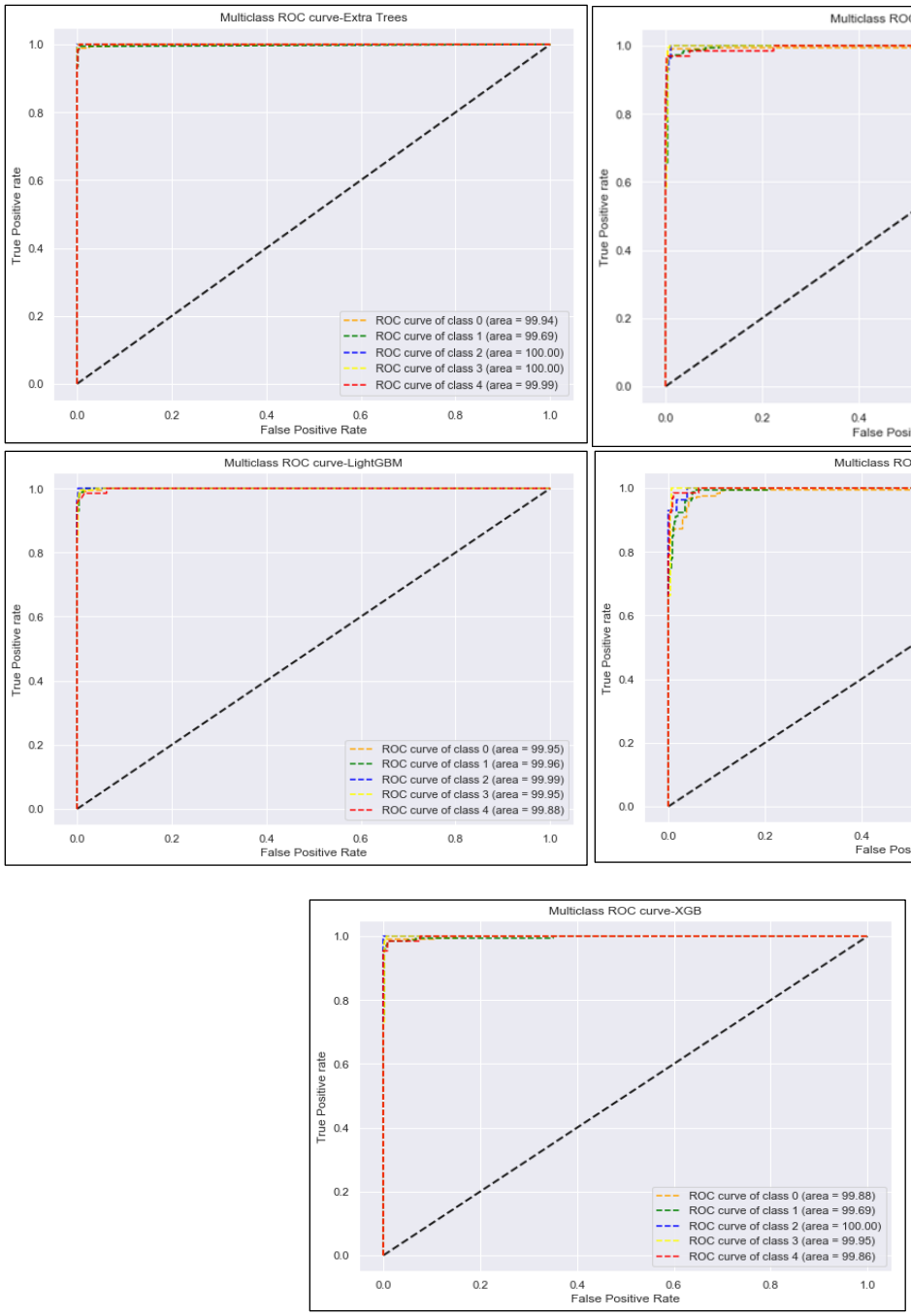

Figure 6 The AUC values were calculated for all ensemble ML models with the holdout CV approach

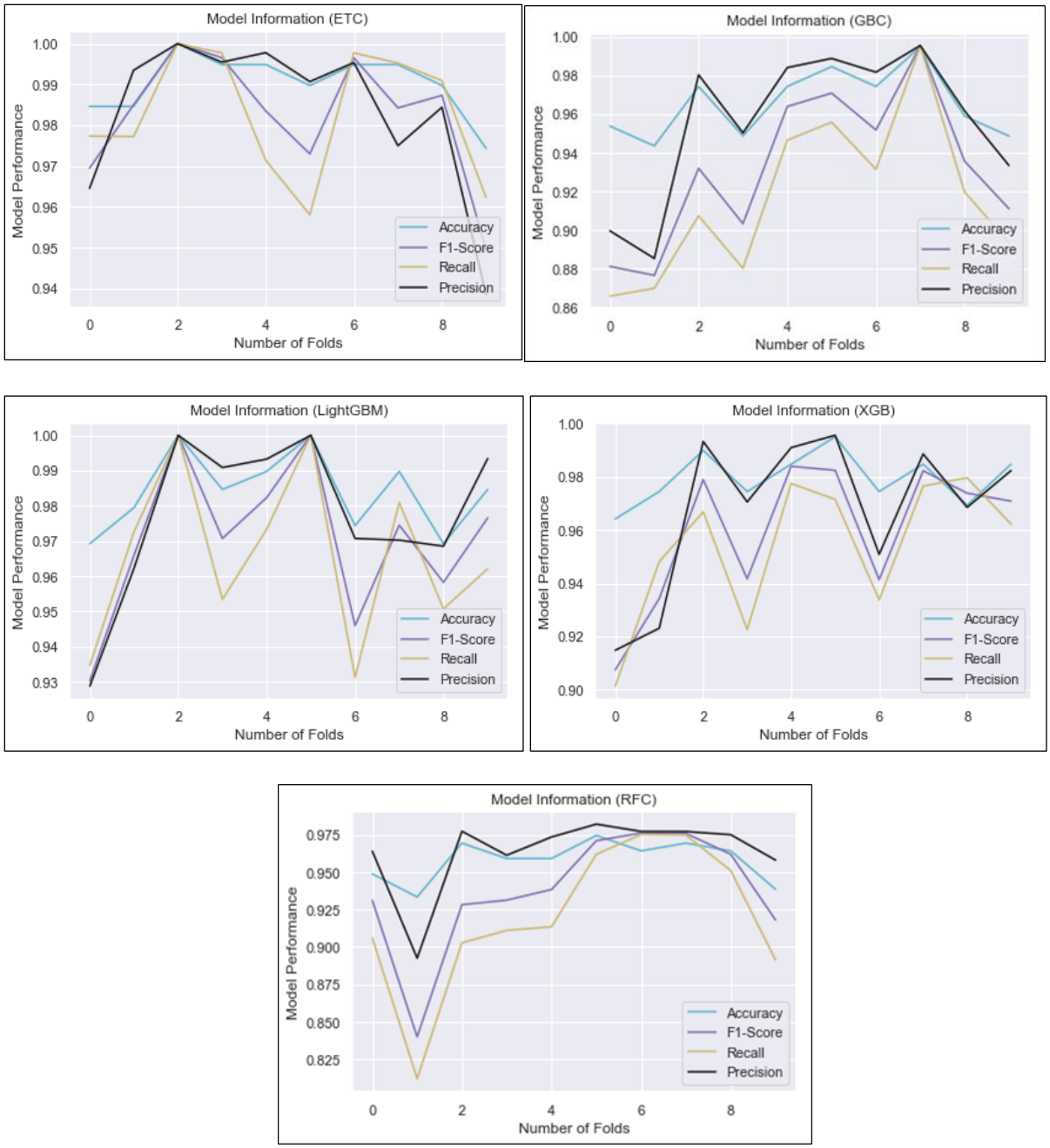

Figure 7 All ensemble ML models were evaluated with a 10-fold CV approach

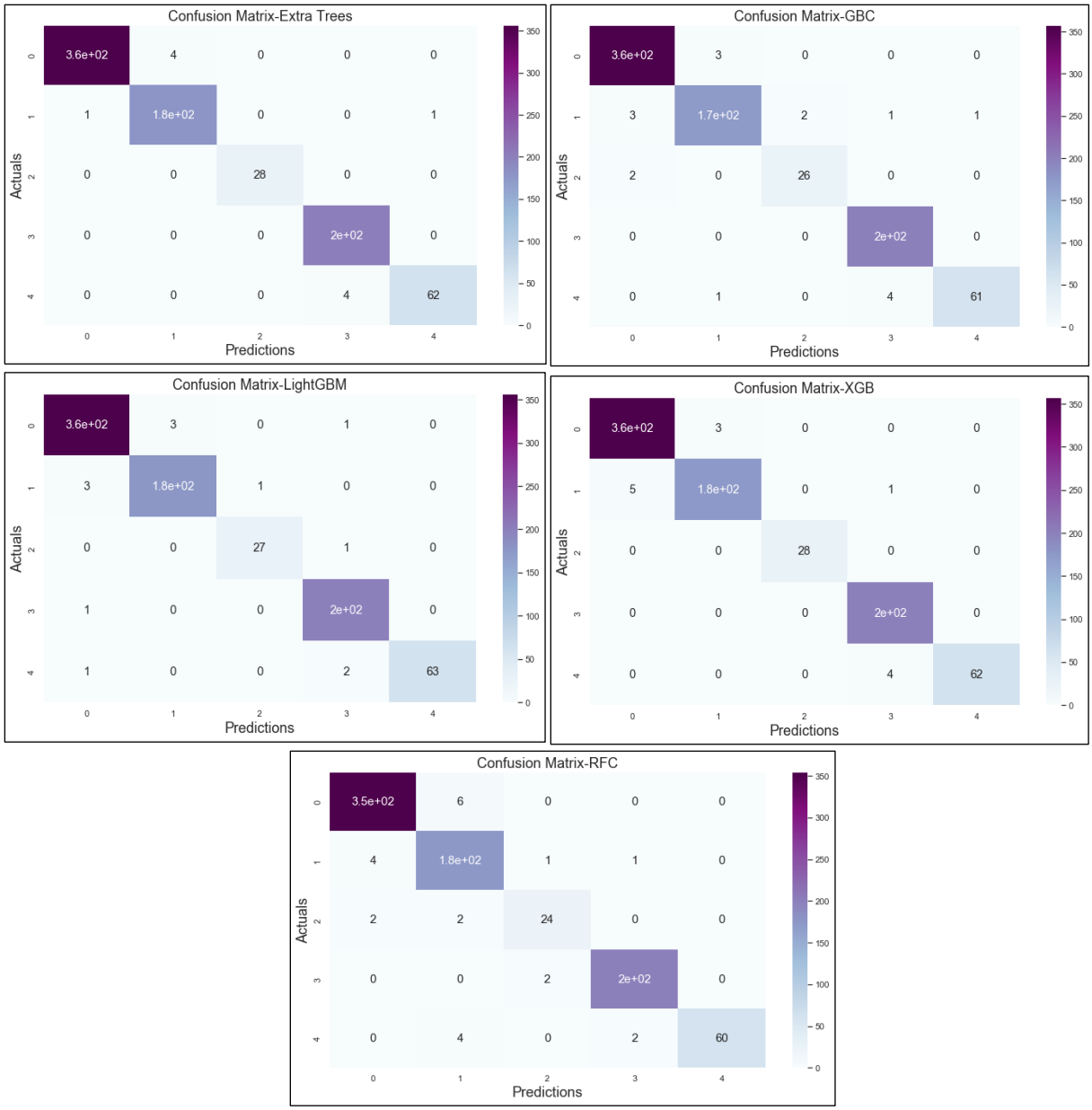

Figure 8 Confusion matrix for all ensemble ML models were evaluated with holdout CV approach

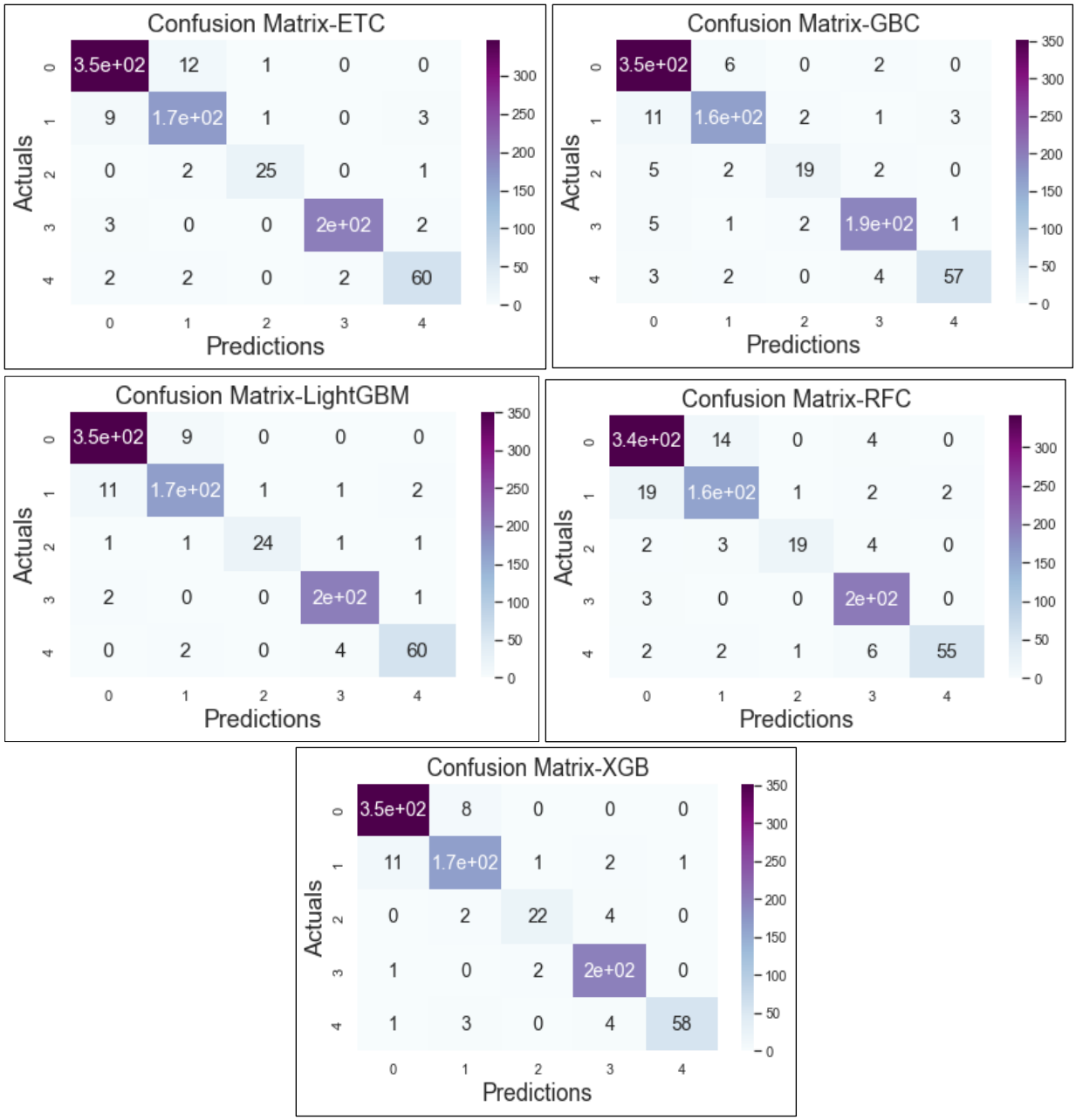

Figure 9 Confusion matrix for all ensemble ML models were evaluated with a 10-fold CV approach

Table 8 Compare ensemble ML model with baseline classifiers models in a 10-fold CV approach

	Ensemble classifiers					**Baseline classifiers**				
Performance measure	**RFC**	**GBC**	**XGB**	**LightGBM**	**ETC**	**LR**	**KNN**	**SVC**	**DTC**	**MLP**
Accuracy (%)	95.79	96.56	97.94	98.41	**99.02**	77.28	97.33	96.3	93.23	96.92
F1-Score (%)	93.71	93.21	95.96	97.04	98.25	53.84	96.51	95.33	88.26	92.89
Precision (%)	96.36	95.61	96.77	97.78	98.35	56.43	97.72	97.92	91.95	95.04
Recall (%)	91.99	91.65	95.39	96.59	98.28	55.4	95.57	93.47	87.37	91.5
AUC (%)	99.11	98.6	99.16	99.48	99.41	87.61	96.61	98.29	95.01	95.23
ER	4.21	3.44	2.06	1.59	0.98	22.72	2.67	3.7	6.77	3.08

Table 9 Compare ensemble ML model with baseline classifiers models in a holdout CV approach

	Ensemble classifiers					Baseline classifiers				
Performance measure	**RFC**	**GBC**	**XGB**	**LightGBM**	**ETC**	**LR**	**KNN**	**SVC**	**DTC**	**MLP**
Accuracy (%)	95.1	97.96	98.44	98.44	**98.92**	75.14	97.84	96.65	93.78	97.13
F1-Score (%)	94.53	96.55	98.4	97.93	98.89	50.37	96.5	96.1	92.01	94.94
Precision (%)	96.62	97.04	98.9	98.28	98.94	53.41	95.91	98.16	93.94	95.86
Recall (%)	92.75	96.11	97.95	97.61	98.84	52.12	97.29	94.29	90.43	94.13
AUC (%)	99.55	99.75	99.87	99.94	99.92	87.62	99.37	99.52	98.46	98.97
ER	4.9	2.04	1.56	1.56	1.08	24.86	2.16	3.35	6.22	2.87

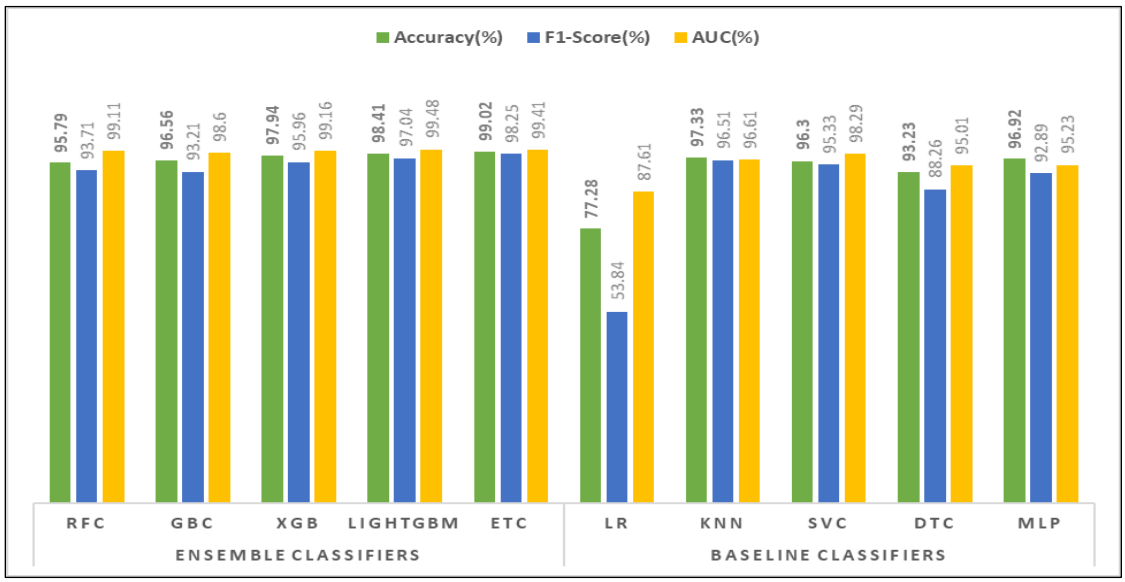

Figure 10 Compare ensemble ML model with baseline classifiers models in a 10-fold CV approach

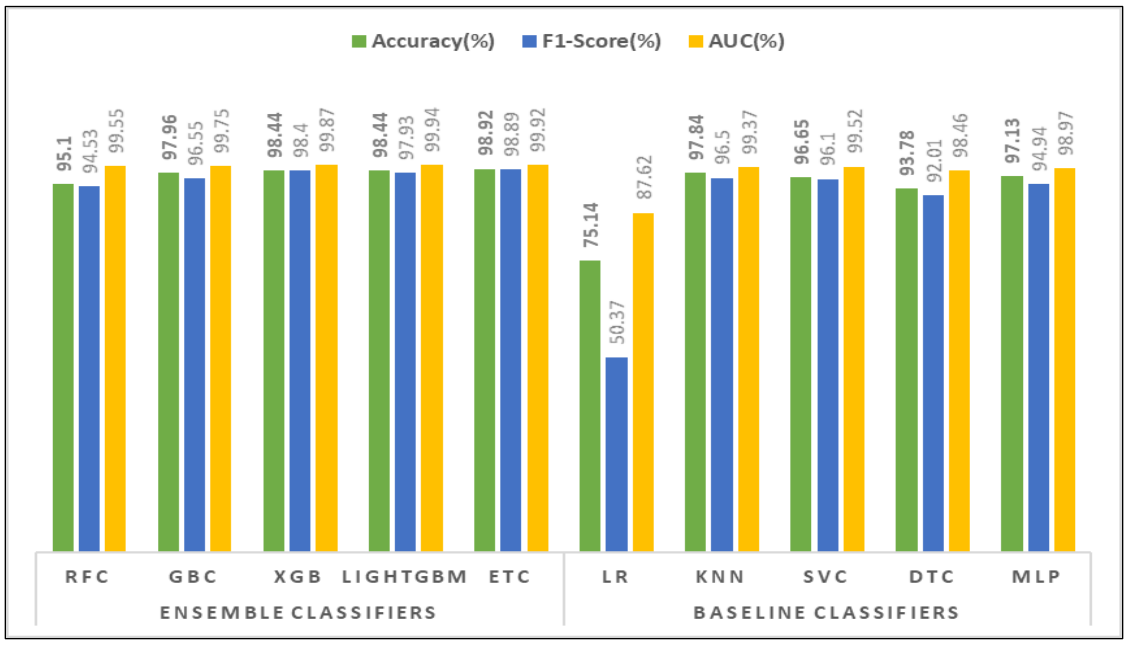

Figure 11 Compare ensemble ML model with baseline classifiers models in a holdout CV approach

Table 10 Compare ensemble ML model with baseline classifiers models in 10-fold CV approach in case of the original dataset (without SMOTE+ENN hybrid technique)

	Ensemble classifiers				Baseline classifiers					
Performance measure	**RFC**	**GBC**	**XGB**	**LightGBM**	**ETC**	**LR**	**KNN**	**SVC**	**DTC**	**MLP**
Accuracy (%)	80.16	79.25	78.41	78.05	79.39	81.43	76.72	79.25	79.46	79.88
F1-Score (%)	25.72	27.31	27.79	26.81	29.78	25.96	31.03	24.6	26.77	30.6
Precision (%)	27.1	30.748	30.06	29.11	35.82	25.12	35.67	28.05	31.17	32.57
Recall (%)	26.22	27.83	28.27	26.97	29.87	27.37	31.31	25.27	27.22	32.31
AUC (%)	84.28	80.73	80.93	79.6	79.6	84.51	70.93	76.78	78.94	82.8

Table 11 Compare ensemble ML model with baseline classifiers models in holdout CV approach in case of the original dataset (without SMOTE+ENN hybrid technique)

	Ensemble classifiers				Baseline classifiers					
Performance measure	**RFC**	**GBC**	**XGB**	**LightGBM**	**ETC**	**LR**	**KNN**	**SVC**	**DTC**	**MLP**
Accuracy (%)	79.18	78.68	79.18	78.68	79.18	79.83	74.91	78.68	78.68	79.67
F1-Score (%)	29.4	31.97	34.68	31.76	33.68	27.57	23.77	31.63	30.03	29.29
Precision (%)	31.97	38.5	44.07	33.15	43.43	44.63	25.67	51.11	31.99	29.42
Recall (%)	30.7	32.28	34.44	32.32	33.57	29.99	23.81	31.98	30.36	34.31
AUC (%)	87.66	86.55	85.67	86.18	85.75	85.61	71.83	76.19	86.21	86.76

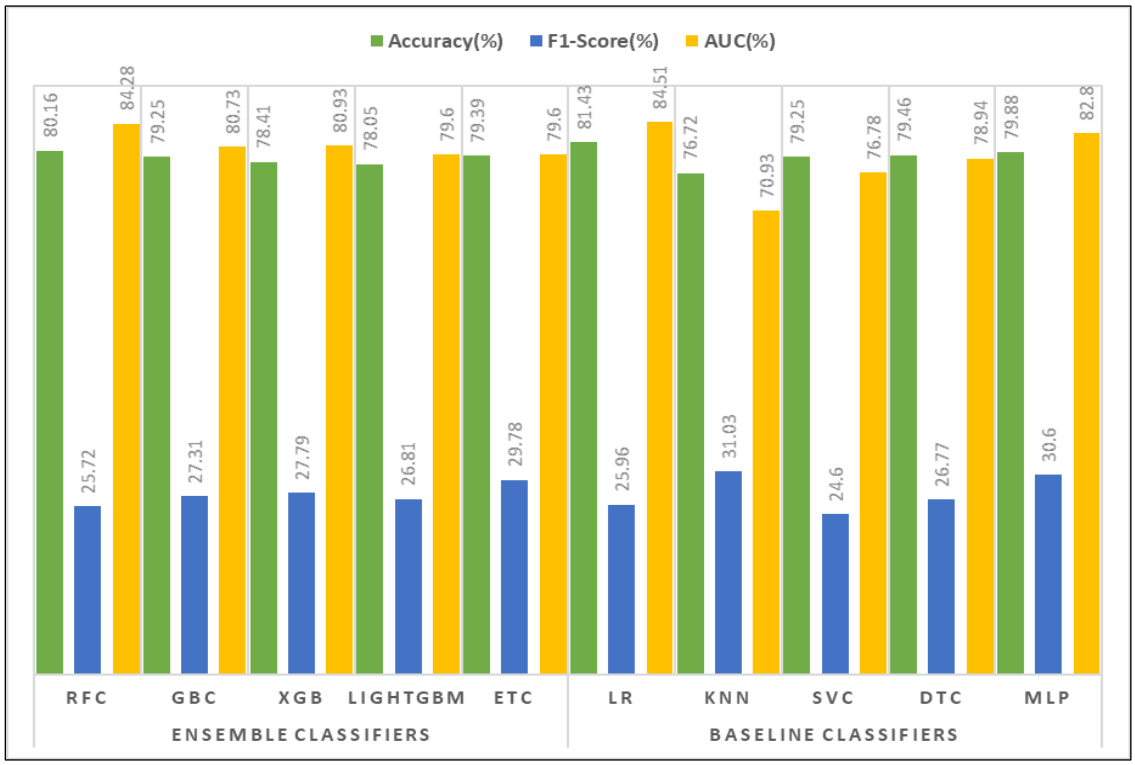

Figure 12 Compare ensemble ML model with baseline classifiers models in 10-fold CV approach in case of the original dataset (without SMOTE+ENN hybrid techniques)

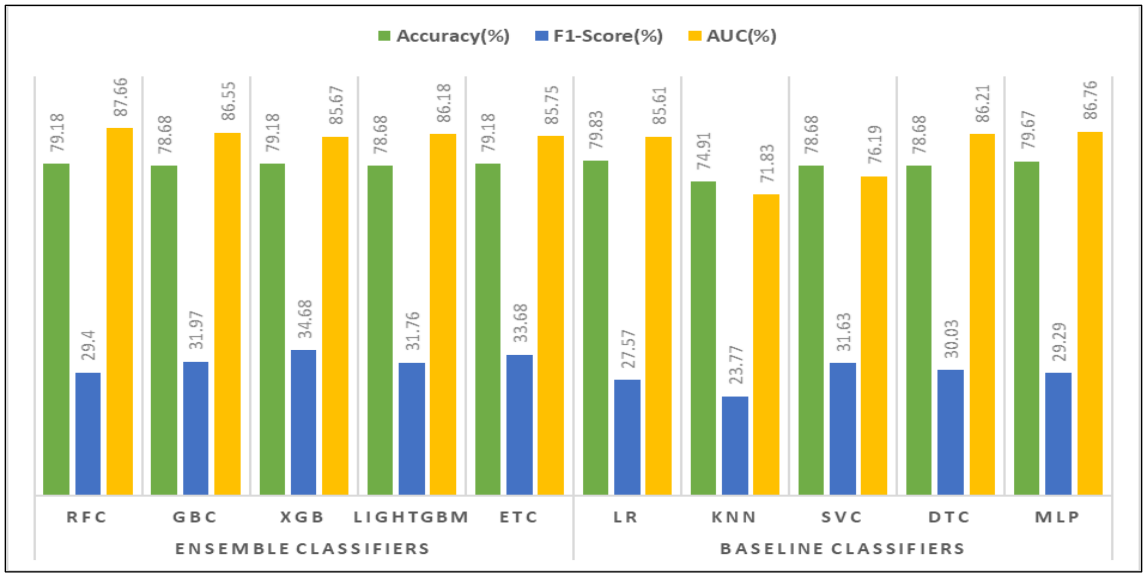

Figure 13 Compare ensemble ML model with baseline classifiers models in holdout CV approach in case of the original dataset (without SMOTE+ENN hybrid techniques)

Table 12 Describes the different multiple class imbalanced datasets which compared with the performance of the best ensemble mode in holdout and 10-fold CV approaches

	Contraceptive dataset	**Yeast dataset**	**Shuttle dataset**
No. of Features	9	8	9
No. of Classes	3	10	7
No. of Instances	1473	1484	2175

Table 13 The accuracy of the performance of the best ensemble ML model in this study with different multiple class imbalanced datasets in holdout and 10-fold CV approaches

Dataset	**Contraceptive**		**Yeast**		**Shuttle**	
Splitting data	**Holdout CV**	**10-Fold CV**	**Holdout CV**	**10-Fold CV**	**Holdout CV**	**10-Fold CV**
Model	ETC	ETC	ETC	ETC	ETC	ETC
Accuracy (%)	**95.1**	**91.55**	**97.81**	**97.15**	**99.7**	**99.5**
Time(s)	0.14	2.78	0.44	11.49	0.71	10.20

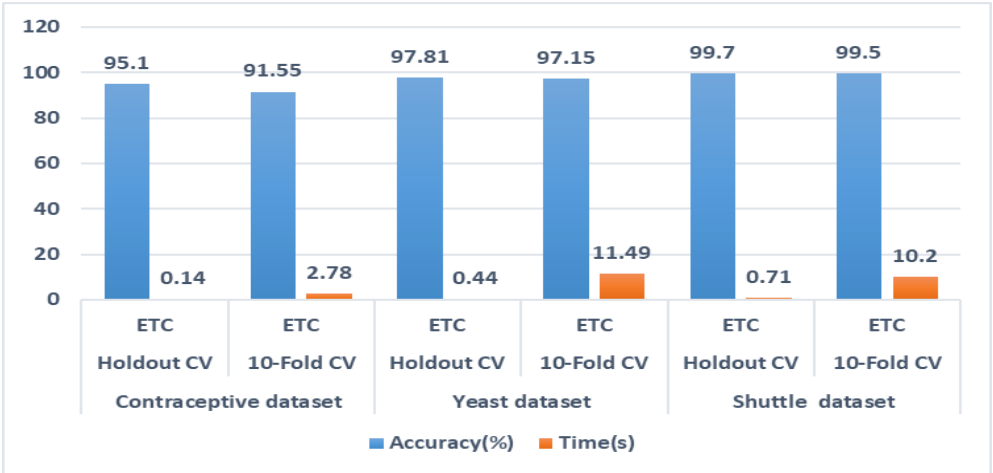

Figure 14 The accuracy of the performance of the best ensemble ML model in this study with different multiple class imbalanced datasets in holdout and 10-fold CV approaches

5.Discussion

Early prediction of individuals at high risk of diphtheria is an essential challenge in the health domain.

In this paper, ten models were used, five of which are efficient ensemble ML models, namely (RFC, GBC, ETC, XGB, and LightGBM) and five baseline ML models namely (LR, KNN, SVC, DTC, and MLP). For prediction of diphtheria diseases using ensemble ML techniques in two scenarios: First scenario, using the hold out CV approach to split the dataset into training and testing data with the balanced dataset (after rebalancing using the SMOTE+ENN method). The ETC model achieved the 98.8% accuracy, 98.55% f1-score, 98.8% precision, 98.34% recall, 99.92% AUC, and 1.08 ER. It is found to be prominent.

Second scenario, using the k-fold (k=10) CV approach with the balanced dataset (after rebalancing with the SMOTE+ENN method), the ETC model achieved the highest values based on the evaluation metrics, with 99.02% accuracy , 98.25% f1-score, 98.35% precision, 98.28% recall, 99.92% AUC, and 0.98 ER.

In comparison with the results of all ensemble models in both holdout and 10-fold approaches to determine the best model that achieved high accuracy. We find the ETC model has achieved the best performance in a 10-fold CV approach.

As similar, we applied five baseline models with the above scenarios and the same performance metrics and balanced dataset (after rebalancing using the SMOTE+ENN method). Based on the evaluation metrics, the best model was the MLP with 97.13% accuracy, followed by 94.94% f1-score, 95.86% precision, 94.13% recall, 98.97% AUC, and 2.87 ER in holdout CV.

And in the k-fold CV approach, the best performing model was the KNN model with 97.33% accuracy followed by 96.51% f1-score, 97.72% precision, 95.57% recall, 96.61% AUC, and 2.67 ER.

We conclude that ensembles ML techniques outperformed baseline ML techniques for diphtheria disease prediction. Also, we applied the same previous methodology to a dataset in case imbalanced (Original Dataset). From our results, we find performance in all (ensemble or baseline) models based on all evaluation metrics used in this study have achieved less than 85% accuracy. With comparison, we conclude from this addition that all ML classifiers, whether ensemble or baseline classifiers work well performance for diseases prediction in the case of balanced healthcare data. For generalization, we applied three multi-class imbalanced data sets (contraceptive, Yeast, and Shuttle) with the ETC model which achieved the best model in this study in both holdout and 10-fold CV approaches. We find the ETC model has achieved an accuracy of more than 90% range of (91.55% - 99.5%). These results prove that the ETC model which achieved the best performance can also be generalized to different multi-class imbalanced medical data.

This study has some implications of this study like the ETC model has the highest performance based on all metrics used. And Performance of the models with all metrics used in this study has high with balanced datasets lowest-performing models with imbalanced datasets whether ensemble or baseline models. Improving the performances of these models may require further adjustments to the hyper parameter values or using effective deep learning techniques.

5.1Limitations

Our work has some limitations which could be addressed in future research. The limitations such as access to data and the lack of existing research studies on diphtheria disease prediction using ML and deep learning approaches. As a result, we could not generalize our findings for the prediction of any disease based on clinical data because our framework system is only for multi-class classification learning. This could be a possible direction for future scope.

A complete list of abbreviations is shown in *Appendix I*.

6.Conclusion and future work

Today, diphtheria infection is a global issue. Early detection and prevention of diphtheria disease may save human lives. In this paper, we developed a framework for diphtheria disease prediction and compared the performance of five ensemble ML models for predicting diphtheria (RFC, GBC, XGB, ETC, and LighGBM) with five baseline ML models (LR, KNN, SVC, DTC, and MLP).

In the initial stage of the work, namely the pre-processing stage, the missing values were processed by mean and the mode method. The selection features technique was applied to select the important

features. The data were normalized by Z-Score. The SMOTE+ENN hybrid technique was used to solve the imbalance problem of the dataset. We applied CV approaches such as 10-fold and holdout CV to split the dataset into training and testing sets. After that, ensemble and baseline ML models were built, and their performance was evaluated using accuracy, F1-score, precision, recall, AUC scores, and ER.

The experimental results show that the ETC model improves the performance of the diphtheria prediction system with 99.2 % accuracy followed by 98.25 % f1-score, 98.35 % precision, 98.28 % recall, 99.92 % AUC, and 0.98 ER in the 10-fold CV approach, and 98.8 % accuracy followed by 98.55 % f1-score, 98.8 % precision, 98.34 % recall, 99.92 % AUC, and 1.08 ER in holdout CV approach, we conclude that the ETC model achieved the best performance with 10-fold CV. And in comparison, we conclude that ensembles' ML techniques outperformed baseline ML techniques for diphtheria disease prediction in the case of the proposed framework in this study.

The results of this study show that combining SMOTE+ENN with this model improved the framework's accuracy in making clinical decisions in accurately predicting diphtheria and any other disease with different datasets.

We intend to expand this research in the future by developing many deep learning approaches with large sizes of data that are likely to properly predict disease kinds.

Acknowledgment

The authors wish to thank the Epidemiological Surveillance Sector of the Ministry of Public Health and Population, Sana'a, YEMEN for giving us the dataset.

Conflicts of interest

The authors have no conflicts of interest to declare.

Author's contributions statements

Bilal Abdualgalil: Conducted the work. **Sajimon Abraham**: Analyzed the data. **Waleed M. Ismael**: Wrote the paper. All authors had approved the final version.

References

[1] Badell E, Alharazi A, Criscuolo A, Almoayed KA, Lefrancq N, Bouchez V, et al. Ongoing diphtheria outbreak in Yemen: a cross-sectional and genomic epidemiology study. The Lancet Microbe. 2021; 2(8):e386-96.

[2] https://www.britannica.com/science/diphtheria. Accessed 28 November 2021.

[3] https://www.downtoearth.org.in/news/health/study-warns-diphtheria-could-become-a-major-global-threat-75866. Accessed 29 November 2021.

[4] Diphtheria, https://www.mayoclinic.org/diseases-conditions/diphtheria/symptoms-causes/syc-20351897. Accessed 29 November 2021.

[5] Mistry M, Bhattacharya A. Emergence of diphtheria in western part of gujarat-a microbiological case series from a tertiary care hospital of Rajkot. Saudi Journal of Pathology and Microbiology. 2021; 6(7):246-9.

[6] Alakus TB, Turkoglu I. Detection of pre-epileptic seizure by using wavelet packet decomposition and artifical neural networks. In international conference on electrical and electronics engineering 2017 (pp. 511-5). IEEE.

[7] Vickers NJ. Animal communication: when i'm calling you, will you answer too? Current Biology. 2017; 27(14):R713-5.

[8] Yousefi J, Hamilton-wright A. Characterizing EMG data using machine-learning tools. Computers in Biology and Medicine. 2014; 51:1-13.

[9] Karthick PA, Ghosh DM, Ramakrishnan S. Surface electromyography based muscle fatigue detection using high-resolution time-frequency methods and machine learning algorithms. Computer Methods and Programs in Biomedicine. 2018; 154:45-56.

[10] Alfaras M, Soriano MC, Ortín S. A fast machine learning model for ECG-based heartbeat classification and arrhythmia detection. Frontiers in Physics. 2019.

[11] Ledezma CA, Zhou X, Rodriguez B, Tan PJ, Diaz-zuccarini V. A modeling and machine learning approach to ECG feature engineering for the detection of ischemia using pseudo-ECG. PloS one. 2019; 14(8):1-21.

[12] Munir K, Elahi H, Ayub A, Frezza F, Rizzi A. Cancer diagnosis using deep learning: a bibliographic review. Cancers. 2019; 11(9):1-36.

[13] Andriasyan V, Yakimovich A, Georgi F, Petkidis A, Witte R, Puntener D, et al. Deep learning of virus infections reveals mechanics of lytic cells. BioRxiv. 2019:1-18.

[14] Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, et al. Improved protein structure prediction using potentials from deep learning. Nature. 2020; 577:706-10.

[15] Petrosino A, Loia V, Pedrycz W. Fuzzy logic and soft computing applications. 11th international workshop, WILF 2016; 2017.

[16] Anggraeni W, Nandika D, Mahananto F, Sudiarti Y, Fadhilla CA. Diphtheria case number forecasting using radial basis function neural network. In international conference on informatics and computational sciences 2019 (pp. 1-6). IEEE.

[17] Park D, Kim BH, Lee SE, Kim DY, Kim M, Kwon HD, et al. Machine learning-based approach for disease severity classification of carpal tunnel syndrome. Scientific Reports. 2021; 11(1):1-10.

[18] Zhang Y, Kambhampati C, Davis DN, Goode K, Cleland JG. A comparative study of missing value imputation with multiclass classification for clinical

heart failure data. In 9th international conference on fuzzy systems and knowledge discovery 2012 (pp. 2840-4). IEEE.
[19] Diri B, Albayrak S. Visualization and analysis of classifiers performance in multi-class medical data. Expert Systems with Applications. 2008; 34(1):628-34.
[20] Mohan S, Thirumalai C, Srivastava G. Effective heart disease prediction using hybrid machine learning techniques. IEEE Access. 2019; 7:81542-54.
[21] Chaudhary A, Kolhe S, Kamal R. An improved random forest classifier for multi-class classification. Information Processing in Agriculture. 2016; 3(4):215-22.
[22] Jacob SG, Ramani RG. Discovery of knowledge patterns in clinical data through data mining algorithms: multi-class categorization of breast tissue data. International Journal of Computer Applications. 2011; 32(7):46-53.
[23] Altaf T, Anwar SM, Gul N, Majeed MN, Majid M. Multi-class Alzheimer's disease classification using image and clinical features. Biomedical Signal Processing and Control. 2018; 43:64-74.
[24] Iqbal N, Islam M. Machine learning for Dengue outbreak prediction: an outlook. International Journal of Advanced Research in Computer Science. 2017; 8(1):93-102.
[25] Yang R, Man S. Improved text feature selection algorithms in classification search of environmental protection information. Journal of Environmental Protection and Ecology. 2019; 20(3):1462-9.
[26] Uçar T, Karahoca A, Karahoca D. Tuberculosis disease diagnosis by using adaptive neuro fuzzy inference system and rough sets. Neural Computing and Applications. 2013; 23(2):471-83.
[27] Fariza A, Jalilah H, Basofi A. Spatial mapping and prediction of diphtheria risk in surabaya, Indonesia, using the hierarchical clustering algorithm. In proceedings of the 1st international conference on electronics, biomedical engineering, and health informatics 2021 (pp. 251-68). Springer, Singapore.
[28] Singh SP, Karkare S, Baswan SM, Singh VP. Agglomerative hierarchical clustering analysis of co/multi-morbidities. arXiv preprint arXiv:1807.04325. 2018.
[29] Fatoni CS, Utami E, Wibowo FW. Expert system for diagnosing diphtheria with k-nearest neighbor method. International Journal Artificial Intelligent and Informatics. 2018; 1(2):45-56.
[30] Chumachenko D, Meniailov I, Bazilevych K, Chukhray A. Intelligent multiagent approach to diphtheria infection epidemic process simulation. In Ukraine conference on electrical and computer engineering 2019 (pp. 833-6). IEEE.
[31] MOH, https://moh.gov.ye/en/home.aspx, Accessed 29 March 2022.
[32] Kuhn M, Johnson K. Feature engineering and selection: a practical approach for predictive models. CRC Press; 2019.
[33] Malik S, Harous S, El-sayed H. Comparative analysis of machine learning algorithms for early prediction of diabetes mellitus in women. In international symposium on modelling and implementation of complex systems 2020 (pp. 95-106). Springer, Cham.
[34] Batista GE, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explorations Newsletter. 2004; 6(1):20-9.
[35] Zhang B, Lu L, Hou J. A comparison of logistic regression, random forest models in predicting the risk of diabetes. In proceedings of the third international symposium on image computing and digital medicine 2019 (pp. 231-4).
[36] Lino FDSBMH, Oliveira AG, Morais FSL, Da SRE, Lorenzato DOlJF, Lynn T, et al. Benchmarking machine learning models to assist in the prognosis of tuberculosis. Informatics 2021: 8(2):1-17. Multidisciplinary Digital Publishing Institute.
[37] Sharaff A, Gupta H. Extra-tree classifier with metaheuristics approach for email classification. In advances in computer communication and computational sciences 2019 (pp. 189-97). Springer, Singapore.
[38] Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining 2016 (pp. 785-94).
[39] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. Lightgbm: a highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems. 2017.
[40] Zhu C, Idemudia CU, Feng W. Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. Informatics in Medicine Unlocked. 2019.
[41] Vapnik V. The nature of statistical learning theory. Springer Science & Business Media; 1999.
[42] Soumaya Z, Taoufiq BD, Benayad N, Yunus K, Abdelkrim A. The detection of Parkinson disease using the genetic algorithm and SVM classifier. Applied Acoustics. 2021.
[43] Kumar A, Das S, Tyagi V, Shaw RN, Ghosh A. Analysis of classifier algorithms to detect anti-money laundering. In computationally intelligent systems and their applications 2021 (pp. 143-52). Springer, Singapore.
[44] Ladić T, Mandekić A. Face mask classification using MLP classifier. Ri-STEM-2021. 2021; 10(68).
[45] Abdualgalil B, Abraham S. Applications of machine learning algorithms and performance comparison: areview. In international conference on emerging trends in information technology and engineering 2020 (pp. 1-6). IEEE.
[46] Grandini M, Bagli E, Visani G. Metrics for multi-class classification: an overview. arXiv preprint arXiv:2008.05756. 2020.
[47] Li H, Jiao R, Fan J. Precision of multi-class classification methods for support vector machines. In

9th international conference on signal processing 2008 (pp. 1516-9). IEEE.
[48] Altuve M, Alvarez AJ, Severeyn E. Multiclass classification of metabolic conditions using fasting plasma levels of glucose and insulin. Health and Technology. 2021; 11(4):953-62.
[49] Hassan MR, Huda S, Hassan MM, Abawajy J, Alsanad A, Fortino G. Early detection of cardiovascular autonomic neuropathy: a multi-class classification model based on feature selection and deep learning feature fusion. Information Fusion. 2022; 77:70-80.
[50] Mary-huard T, Perduca V, Martin-magniette ML, Blanchard G. Error rate control for classification rules in multiclass mixture models. The International Journal of Biostatistics. 2021.
[51] https://sci2s.ugr.es/keel/index.php. Accessed 28 December 2021.

Bilal Abdualgalil is PhD a Research Scholar in AI, Mahatma Gandhi University, Kerala, India. He received his Master degree in Master of Computer Application (MCA) from JNT University, Hyderabad, India in 2018, And his B.Sc. degree from Thamar University, Yemen in 2009. His research interests include Artificial Intelligence (Deep learning and Machine learning) in the healthcare domain.
Email: bsaa85@gmail.com

Dr. Sajimon Abraham is Professor in Computer Applications in School of Management and Business Studies, Mahatma Gandhi University Kottayam Kerala. He received Ph.D in Computer Science from Mahatma Gandhi University in 2015 in the area of Spatiotemporal Data Mining. He has additionally held the position of Director of University Centre for International Co-operation and worked in the Royal University of Bhutan for 3 years as a Computer Science Faculty Member and Data Base Architect under the Ministry of External Affairs, Government of India. He has published more than 90 research articles in various journals and his research area includes Data Analytics, Spatiotemporal Data Mining, E-learning and applications of AI in the Business Domain.
Email: sajimabraham@rediffmail.com

Waleed M. Ismael is an Assistant Professor in the Information and Communication Engineering, majoring in IoT Engineering. He received his BS.c in Computer Science from Thamar University, Yemen, in 2006. In 2009, he received a postgraduate diploma in Geoinformatics from ITC institute, Hollande. He completed his MS.c in Geoinformatics, Osmania University, India. His research interests include WSN reliability, Data fusion, Geoinformatics, and Deep Learning.
Email: Waleed.m@hhu.edu.cn

Appendix I

S. No.	Abbreviation	Description
1	AUC	Area Under Curve
2	CLNS	Cervical L.N. Swelling
3	CTS	Carpal Tunnel Syndrome
4	CV	Cross-Validation
5	DOB	Difficulty of Breathing
6	DTC	Decision Tree Classifier
7	EFB	Exclusive Feature Bundling
8	EEMLT	Efficient Ensemble Machine Learning Techniques
9	ER	Error Rate
10	ETC	Extra Tree Classifier
11	FN	False Negative
12	FP	False Positive
13	GBC	Gradient Boosting Classifier
14	GBDT	Gradient Boosting Decision Tree
15	GOSS	Gradient-based One Side Sampling
16	HF	Heart Failure
17	KNN	K-Nearest Neighbors
18	LightGBM	Light Gradient Boosting Machine
19	LM	Linear Method
20	LR	Logistic Regression
21	MASE	Mean Absolute Scaled Error
22	ML	Machine Learning
23	MLP	Multilayer Perceptron
24	NB	Naive Bayes
25	PM	Pseudomembrane
26	RBFNN	Radial Basis Function Neural Network
27	RFC	Random Forest Classifier
28	SML	Supervised Machine Learning
29	SVC	Support Vector Classifier
30	SMOTE+ ENN	Synthetic Minority Oversampling Technique+ Edited Nearest Neighbours
31	TN	True Negative
32	TNODV	Total Number of Diphtheria Vaccine Received (Penta/DPT/DT)
33	TP	True Positive
34	TR	Treatment Received
35	TST	Throat Swab Taken
36	URTI	Upper Respiratory Tract Infection
37	XGB	eXtreme Gradient Boosting