Volume 66, Issue 2, 2022

Journal of Scientific Research

of

The Banaras Hindu University

An Efficient Machine Learning Techniques as Soft Diagnostic for Tuberculosis Classification Based on Clinical Data

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

Bsaa85@gmail.com

Abstract. TB infection is a global problem, especially in Yemen. Early detection is critical to reducing TB deaths. As a result, accurate tuberculosis diagnosis takes time due to numerous clinical examinations. This problem requires a new diagnosis schema. In this study, we proposed classification models based on Efficient Machine Learning Techniques (EMLT), which predict whether the patient is TB-positive or TB-negative. Nine Different Efficient Machine learning models were trained and tested in two imbalance dataset cases using Stratified 10-Fold Cross-Validation and Holdout Cross-Validation and balanced dataset case using Holdout Cross-Validation with Synthetic Minority Oversampling Technique (SMOTE). The best model was evaluated on a test set using F1-score measure in imbalanced dataset case and accuracy measure in balanced dataset case. Based on the obtained results, the models that achieved the highest value of the F1-Score measure in the imbalanced dataset were LR and GBC with 99.826% value in Stratified Cross-Validation approach and GBC with 86.0334 in the Holdout Cross-Validation approach. And the models that achieved the highest value of the accuracy measure in the balanced dataset case (SMOTE) and Holdout Cross-Validation, were LR and GBC with a 99.725% value.

Keywords: Tuberculosis. Classification. Machine learning. Imbalance. Balanced. Early detection. SMOTE.

1 Introduction

According to a recent report of the National Tuberculosis Control Program in Taiz Governorate in 2020, the prevalence of Tuberculosis (TB) has increased in Taiz city between 2017–2020. The report showed that 495 TB cases in 2017 compared to 900 TB cases from 2018 to 2020 However, TB death rates in Taiz decreased over the four years from 2017 to 2020 [1].

Early detection of tuberculosis (TB) is critical in reducing the death rate associated with the disease. However, early detection of tuberculosis has some limitations, such as the fact that it takes a long time to correctly diagnose tuberculosis [2] because it necessitates a large number of clinical examinations. As a result, accurate and rapid early detection of tuberculosis is required to assist clinicians in selecting the most appropriate treatment for their patients. Effective machine learning techniques have recently developed several widely used techniques for diagnosing diseases that are widely used to identify diseases. This approach uses a set of clinical data to develop a model that can be used to detect tuberculosis (TB) on its own. Several studies have been conducted to detect tuberculosis (TB) using a variety of features.

Olatunji et al. [3] used a genetic neuro-fuzzy inferential model-based neural network for the diagnosis of tuberculosis (TB).

Bobak et al. [4] used machine learning techniques that are commonly used for TB classification based on transcriptional biomarkers, such as Support Vector Machine (SVM), Partial Least Squares (PLS), and Random Forest (RF).

A Gaussian Fuzzy Neural Network has been proposed by Mithra and Emmanuel [5] to diagnose tuberculosis (TB) based on microscopic images of sputum smear. To detect mycobacterium TB in patients, Uçar and Karahoca [2] proposed the Adaptive Neuro-Fuzzy Inference System (ANFIS). However, the challenges still faced by them, it a challenge remains in using machine learning techniques to classify TB based on clinical data which are mostly imbalanced and also the importance of features selection were used which affect the accuracy of this classification. This was accomplished in this study.

The purpose of this study is to evaluate the performance of efficient machine learning techniques for tuberculosis (TB) classification based on clinical data, specifically classification techniques. The primary contribution of this study is:

- Developing efficient machine learning techniques as soft diagnostic for the classification of tuberculosis disease based on clinical balanced and imbalanced data.
- Comparing different efficient machine learning techniques based on clinical balanced and imbalanced datasets.
- Identifying the best-performing model for the classification of TB disease in balanced and imbalanced datasets.

This paper is structured as follows. Section 2 describes the dataset and methods. Section 3 contains the findings and discussion. Section 4 contains the conclusion and recommendations for future work.

DOI: 10.37398/JSR.2022.660209

2 Dataset and Methods

2.1 Dataset

The National Tuberculosis Control Program (NTCP) in Taiz, Yemen, provided the datasets for this study. The clinical data were collected during three years, from 2017 to 2020. The dataset contains 1395 samples, which are divided into two categories: positive and negative. The patient with a tuberculosis diagnosis (referred to like 1 in the data set) is of a positive class (TB Positive). The negative class (as 0 in the dataset) refers to patients who are tuberculosis free (TB Negative). There are 938 samples in the positive class and 457 samples in the negative class. Each sample has ten clinical features, all of which are binary. Sex, year, and cough in two weeks, bleeding cough, sweating at night, fever at night, easily tired, weight loss, dyspnea, and decreased appetite are the binary features in the dataset used in this study. The distribution of clinical features are listed in Table 1.

Table 1. TB dataset format; 1: positive; 0: negative

No	Feature name	Values	Data type
1	sex	{M,F}	C
2	cough in two weeks	{1, 0}	C
3	bleeding cough	{1, 0}	C
4	sweating at night	{1, 0}	C
5	fever at nigh	{1, 0}	C
6	Easily tired	{1, 0}	C
7	Weight Loss	{1, 0}	C
8	Dyspnea	{1, 0}	C
9	Decrease appetite	{1, 0}	C
10	Class	{TB Positive, TB Negative}	C

Table 1. In the distribution of the binary values. For a sex feature, value 1 denotes a male whereas value 0 denotes a female, and about all features values, the value 1 represents the positive case, and the value 0 represents the negative case.

2.2 Experiment Design

The proposed framework for designing the experiments used in this study is illustrated in (see Fig. 2) to benchmark the machine learning models. The goal was to find the most effective model for TB prognosis. The framework utilized in this work included pre-processing is applied to the TB clinical data, and features selection, shown in (see Fig. 1) that represented the Two-dimensional ranking of TB dataset features that use Pearson rank correlation to demonstrate how well the features correlated with the results of the diagnosis. The color in darker red shows that the features are highly correlated and to the diagnostic results, for example easily tired, weight loss variables, fever at night, and sweating at night. Training and testing models were used in two cases of dataset: an imbalanced dataset case and a balanced data set case. This study used two scenarios designed for experiment and evaluation: (1) using the imbalanced dataset, and (2) using a balanced dataset. The first scenario is the original data set. The pre-processed data set was imbalanced (939 positive class and 475 negative class), as shown in (see Fig.3 (a)). The imbalance dataset case used two ways to split the dataset for training and testing for all models. The first way is stratified k-fold cross-validation, and the second way is Holdout cross-validation. 10-fold cross-validation is the first way. This means the clinical imbalanced data set are divided into 10 times, with two separate sets for every number of each fold (training and testing sets). The second way is Holdout cross-validation, which is used to split the imbalance dataset into training and testing datasets, where 70% of data is for training and 30% for testing. The second scenario in this study was designed for this experiment and evaluation. The balanced dataset was generated by resampling the imbalance dataset using Synthetic Minority Oversampling Technique (SMOTE), comprising (939 positive class and 939 negative classes), and (see Fig. 3 (b). This scenario used the Holdout cross-validation way to split the balanced dataset into training and testing datasets, where 70% of data is used for training, and 30% is used for testing. During the training phase, every classification model was evaluated using evaluation metrics. As a trained model, the best classification method with the highest accuracy was selected. The trained models were then used to predict the "unseen" clinical data in the testing data. In all different scenarios, the prediction results of the test data sets were evaluated using the same methods as in the training phase.

Finally, in the testing phase, the best models were selected based on the accuracy measure in the balanced dataset case and the f1-score measure in the imbalanced dataset case.

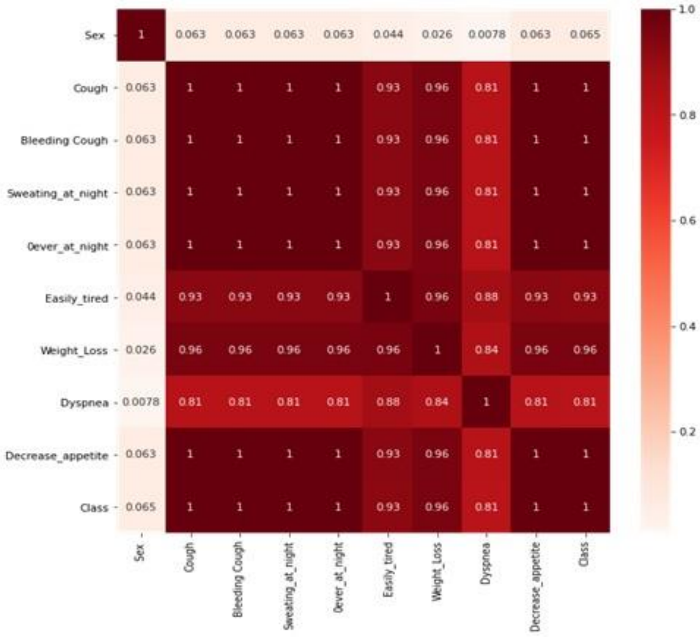

Fig. 1. Pearson-ranking visualization of the TB dataset.

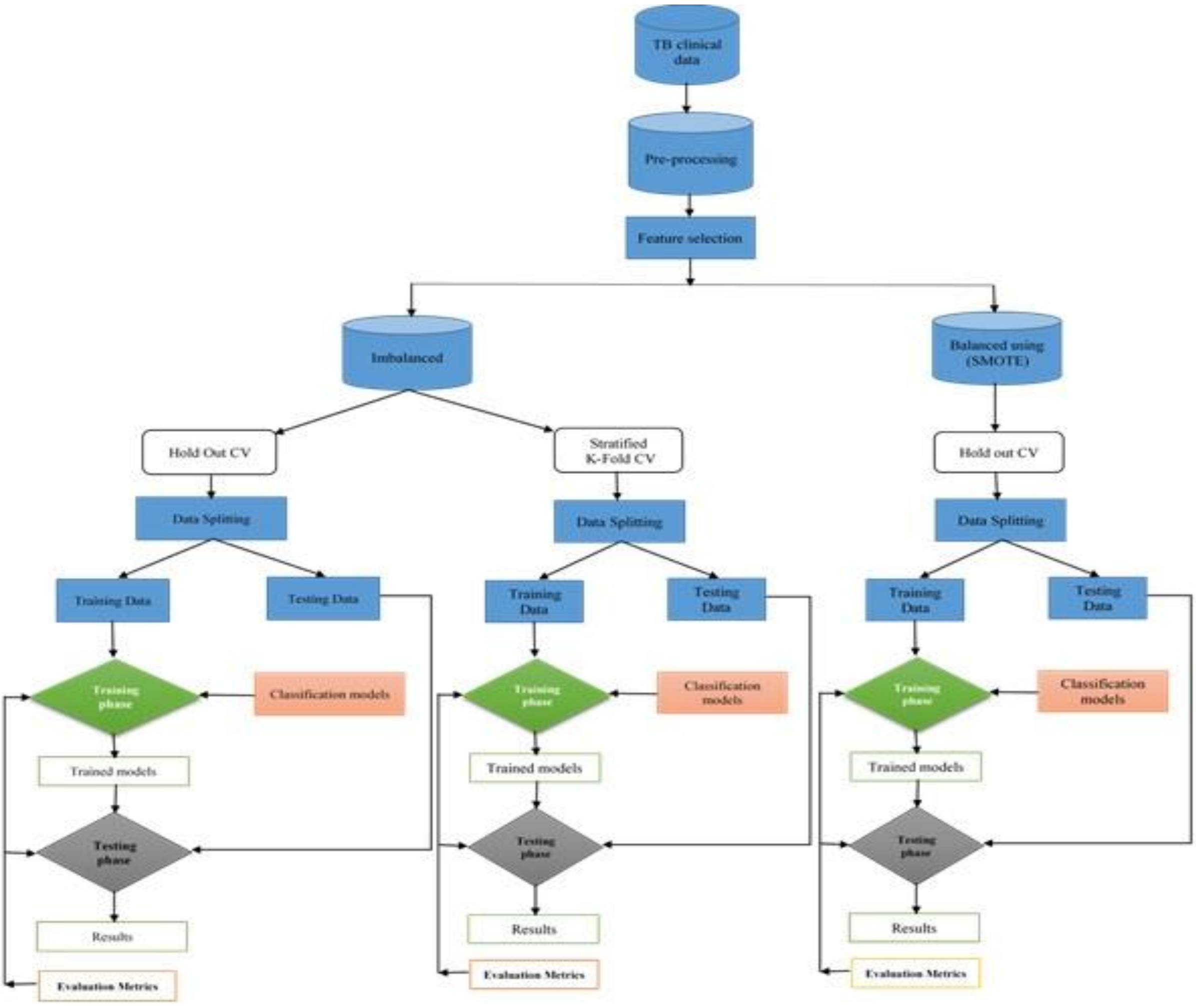

Fig. 2. Proposed framework

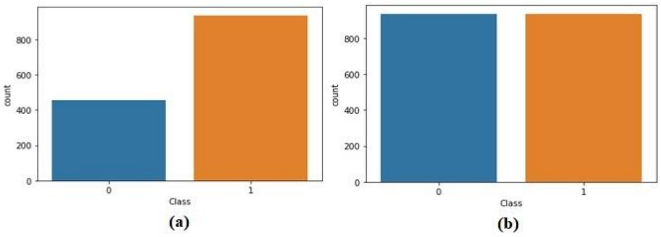

Fig. 3. (a) Imbalanced data set and (b) Balanced data set.

2.3 Classification Methods

In this study, nine efficient machine learning models for classification are used, including Logistic Regression (LR), Naive Bayes (NB), K-Nearest Neighbors (KNN), Support Vector Machine, Decision Tree (DT), Random-Forest, Multilayer Perceptron (MLP), Linear Discriminant Analysis (LDA), and Gradient Boosting classifier (GBC). Various models used in this study are made available in the Python-based scikit-learn package, providing a set of efficient machine learning and modelling tools, including classification, regression, and clustering. The training methods accompanying the package enable users to fine-tune classification parameter settings to achieve maximum accuracy [6].

Table 2 lists the Hyper-parameters settings for each classification model in detail.

Table 2. Hyper-parameters settings of classification methods

No.	Model	Hyper-parameters settings setting
1	LR	penalty='l2',solver='sag', C=1.0,random_state=33
2	NB	priors=None, var_smoothing=1e-09
3	KNN	n_neighbors= 10,weights ='uniform', algorithm='auto'
4	SVC	kernel= 'rbf', max_iter=100,C=2.0, gamma=1
5	DT	criterion='entropy',max_depth=3,random_state=33
6	RF	criterion = 'gini',n_estimators=25,max_depth=5,random_state=33
7	MPL	activation='relu',solver='adam',learning_rate='constant',early_stopping= True,alpha=0.0001 ,hidden_layer_sizes=(100, 4),random_state=33
8	LDA	n_components=1,solver='svd',tol=0.00001
9	GBC	reg_param=0.1,tol=0.0001

2.4 Evaluation metrics

There are several methods for evaluating the performance of learning models in Supervised Machine Learning (SML). This study uses four metrics to compare the models: accuracy, sensitivity, specificity, F1-score. To understand these metrics, a confusion matrix, which is commonly used to determine the performance of a classified as in Table 3. In a confusion matrix, TN represents the number of negative instances correctly classified (True Negatives), FP represents the number of negative instances incorrectly classified as positive (False Positive), FN represents the number of positive instances incorrectly classified as negative (False Negatives), and TP represents the number of positive instances incorrectly classified as negative (False Positives) (True Positives). Many standard evaluation metrics can be defined using the confusion matrix.

Table 3. Confusion matrix

	Predicted negative	Predicted positive
Actual negative	TN	FP
Actual positive	FN	TP

- **Accuracy**. Accuracy is the number of correctly classified samples to the total number of samples [7]. The following equation represents it mathematically as a ratio between the sum of TP and TN and the sum of all samples.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

This measure works well when the number of samples belonging to each class is equal (Balanced).

- **Sensitivity** (Also known as **Recall**). This measure is used to measure completeness. It gives the number of correctly labelled reviews in the test set as positive out of the total number of truly positive reviews [8]. The recall can be written as the following equation:

$$sensitivity = \frac{TP}{TP + FN} \quad (2)$$

- **Specificity.** is the complement to sensitivity or the true negative rate, and summaries how well the negative class was predicted [9], it can be defined as follows:

$$Specificity = \frac{TN}{(TN+FP)} \quad (3)$$

- **Precision.** This metric is used to measure the precision of the reviews. It indicates whether all positive reviews have been correctly labelled as positive to the total number of positive reviews [10]. The precision is calculated using the following equation:

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

- **F1-score**. Measures the recall and precision. It combines the values of recall and precision measures. If F-score is high, the system architecture is reasonable and the proposed techniques are effective [11]. F1- score is measured by the following equation:

$$F1 - score = 2 * \left(\frac{Precision * Sensitivity}{Precision + Sensitivity}\right) \quad (5)$$

3 Results and Discussion

The experimental results of TB classification using nine classification models (Logistic Regression, Nave Bayes, K-Nearest Neighbors, Support Vector Machine, Decision Tree, Random Forest, Multilayer Perceptron, LDA, and Gradient Boosting) are presented. All models are tested on the same dataset and evaluated using the same metrics. Tables 3, 4, and 5 show the classification method evaluation results on the training and testing datasets. This study used two metrics: accuracy and F1-Score, considered for comparison of classification models performance for imbalance and balanced dataset cases in the testing phase. In the imbalanced dataset case, we consider the F1-score measure with 10-Fold Stratified Cross-Validation and Holdout Cross-Validation because its value also accommodates the values of sensitivity and precision. In the balanced dataset case, we considered accuracy measure with Holdout Cross-Validation (SMOTE). As shown in (see Fig. 4), the models that achieved the highest values based on the F1-Score measure were LR and GBC with 99.826% and 99.826 respectively in Stratified Cross-Validation cases. As shown in (see Fig. 5), the model that achieved the highest value based on the F1-Score measure was GBC with 86.0334% with Holdout cross-validation and in an imbalanced dataset case. Also, based on the obtained results shown in (see Fig. 6), the model that achieved the highest value based on the Accuracy measure were LR and GBC with 99.725% in balanced dataset case (with SMOTE) and Holdout cross-validation.

Finally, the LR and GBC models achieved the best models for TB classification based on clinical data with Stratified 10-Fold Cross-Validation in imbalanced data cases and with holdout cross-validation (SMOTE) in balanced data cases. Overall, all machine learning models performed well in the classification of two TB data categories, based on the results of the overall experiment.

The samples in the training and testing datasets are very similar, which contributes to the high accuracy in all experiments.

Because machine learning models lack interpretability, they cannot explain why a sample is classified into a class. The clinician, on the other hand, requires interpretability to make an accurate diagnosis.

Table 4. Performance results in the classification models in training and testing phase in Stratified 10-Fold Cross-Validation with imbalanced data case.

	Training phase				Testing phase			
Model	Accuracy	sensitivity	specificity	F1-score	Accuracy	sensitivity	specificity	F1-score
LR	99.761	100	99.386	99.769	99.692	100	99.236	**99.826**
NB	99.045	99.846	99.693	99.846	99.793	98.958	99.236	99.303
KNN	99.045	99.692	100	99.846	99.794	98.958	99.236	99.303
SVC	99.045	99.846	99.693	99.692	99.589	98.958	99.236	99.303
DT	99.045	99.692	100	99.692	99.589	98.958	99.236	99.303
RF	99.045	99.846	99.693	99.692	99.589	98.9583	99.236	99.303
MLP	99.045	99.692	100	99.844	99.794	98.958	99.236	99.303
LDA	99.045	99.846	99.693	99.846	99.793	98.958	99.236	99.303
GBC	99.761	100	99.386	99.692	99.589	100	99.236	**99.826**

Table 5. Performance results in the classification models in training and testing phase in Holdout Cross-Validation case with balanced data case (with SMOTE).

	Training phase				Testing phase			
Model	Accuracy	sensitivity	specificity	F1-score	Accuracy	sensitivity	specificity	F1-score
LR	99.923	99.846	100	99.923	**99.725**	98.606	99.637	99.424
NB	99.923	100	99.848	99.923	99.165	98.606	99.637	99.124
KNN	99.923	99.846	100	99.923	99.165	98.606	99.637	99.124
SVC	99.923	100	99.848	99.923	99.165	98.606	99.637	99.124
DT	99.795	99.692	100	99.845	99.165	98.958	99.236	99.124
RF	99.923	100	99.848	99.923	99.165	98.606	99.637	99.124
MLP	99.923	99.846	100	99.923	99.165	98.606	99.637	99.124
LDA	99.923	100	99.848	99.923	99.165	98.606	99.637	99.124
GBC	99.923	99.846	100	99.923	**99.572**	98.606	99.637	99.324

Table 6. Performance results in the classification models in training and testing phase in Holdout Cross-Validation case with imbalanced data.

	Training phase				Testing phase			
Model	Accuracy	sensitivity	specificity	F1-score	Accuracy	sensitivity	specificity	F1-score
LR	74.232	31.280	96.997	83.109	74.552	30.113	98.4701	**83.419**
NB	74.232	31.280	96.997	83.109	74.552	30.113	98.470	**83.419**
KNN	74.232	31.280	96.997	83.109	74.552	30.113	98.470	**83.419**
SVC	39.505	89.655	12.924	21.8302	40.159	92.045	12.232	20.997
DT	74.402	28.325	98.825	83.46	73.757	27.272	98.776	83.033
RF	74.402	28.325	98.825	83.461	73.757	27.272	98.776	83.033
MLP	65.358	0	100	79.050	65.009	0	100	78.795
LDA	74.232	31.280	96.997	83.109	74.552	30.113	98.470	**83.419**
GBC	74.402	28.325	98.825	83.46	76.757	27.272	98.776	**86.033**

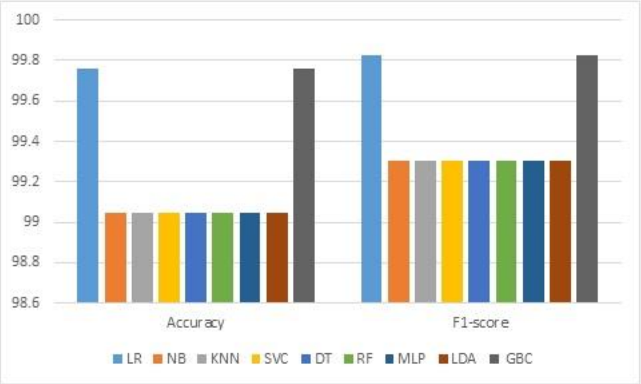

Fig. 4. Comparison accuracy and F1-score of the best classifier in Stratified k-fold cross-validation in the testing phase and data case is imbalanced.

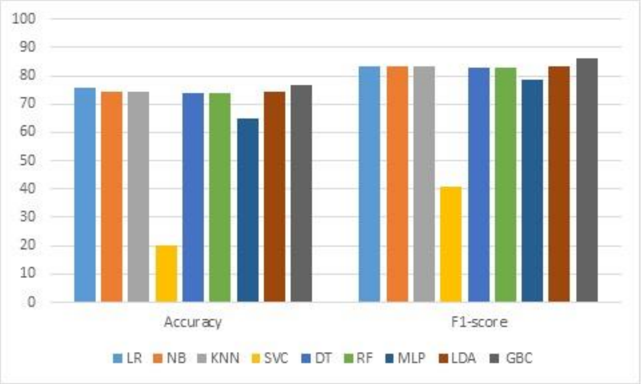

Fig. 5. Comparison accuracy and F1-score of the best classifier in the testing phase with Holdout cross-validation and data case is imbalanced.

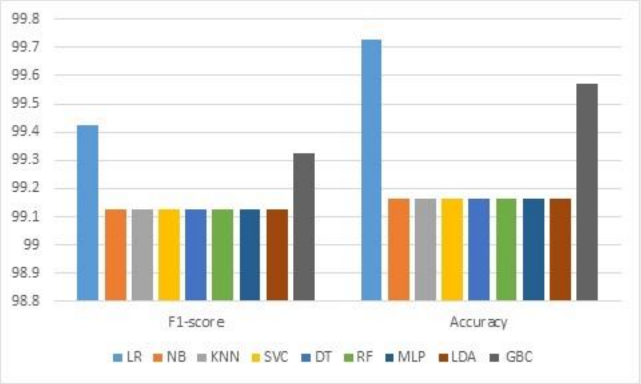

Fig. 6. Comparison accuracy and F1-score to specific of the best classifier in the testing phase with Holdout cross-validation and data case is balanced (with SMOTE).

4 Conclusion and Future Work

Based on patient clinical data, this study evaluates the performance of machine learning models for classifying TB. Nine classification methods used in this study are LR, NB, KNN, SVM, DT, RF, MLP, GBC, and LDA. This study used two scenarios designed for experiment and evaluation: (1) using the imbalanced dataset case, and (2) using a balanced dataset case.

The first scenario is the original dataset that was imbalanced. The imbalance dataset used two ways to split the dataset to evaluate the performance in training and testing phases for all models. The first way is stratified k-fold cross-validation, and the second way is Holdout cross-validation.

The second scenario was the balanced dataset that was generated by resampling the imbalance dataset using Synthetic Minority Oversampling Technique (SMOTE). To evaluate the performance of all models in the training and testing phases, we used accuracy, sensitivity, specificity, and F1-score measures.

The best classification performance of models selected for the imbalance and balanced dataset case was in the testing phase by two metrics: accuracy and F1-Score. In the imbalanced dataset case, we consider the F1-score measure with 10-Fold Stratified Cross-Validation and Holdout Cross-Validation. In the balanced dataset case (with SMOTE) we consider accuracy with Holdout Cross-Validation. The best models that achieved the highest value based on the F1-score measure were LR and GBC with 99.826% and 99.826 respectively in Stratified Cross-Validation cases. Moreover, the best model that achieved the highest value based on the F1-Score measure was GBC with 86.0334%, followed by LR, NB, KNN, and LDA with 83.419% in an imbalanced dataset case. Also, the best models that achieved the highest value based on the accuracy measure were LR and GBC with 99.725% in balanced dataset case.

Clinical data and images of TB patients will be used in the future to classify TB with multimodal features. Another direction in the future is to improve the interpretability of TB-classification machine learning models.

Acknowledgment. The authors wish to thanks the National Tuberculosis Control Program (NTCP), Taiz, YEMEN for giving us the dataset.

References

[1] National Tuberculosis Control Program (2020), https://taizgho.com/taiz/yemen, last accessed 2021/07/21

[2] Uçar, Tamer, and Adem Karahoca.:Predicting existence of Mycobacterium tuberculosis on patients using data mining approaches. Procedia Computer Science (3), 1404-1411 (2011).

[3] Omisore, Mumini Olatunji, Oluwarotimi Williams Samuel, and Edafe John Atajeromavwo.: A genetic-neuro-fuzzy inferential model for diagnosis of tuberculosis. Applied Computing and Informatics 13(1), 27-37(2017).

[4] Bobak, Carly A., Alexander J. Titus, and Jane E. Hill.: Comparison of common machine learning models for classification of tuberculosis using transcriptional biomarkers from integrated datasets. Applied Soft Computing (74), 264-273(2019).

[5] Mithra, K. S., and WR Sam Emmanuel.: GFNN: gaussian-Fuzzy-Neural network for diagnosis of tuberculosis using sputum smear microscopic images. Journal of King Saud University-Computer and Information Sciences (2018).

[6] Fernández-Delgado, Manuel, et al.: Do we need hundreds of classifiers to solve real-world classification problems?. The journal of machine learning research 15(1), 3133-3181, (2014).

[7] Abdualgalil, Bilal, and Sajimon Abraham. : Applications of Machine Learning Algorithms and Performance Comparison: A Review. In International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), pp. 1-6. IEEE, Vellore, India (2020).

[8] Sandaruwan, H. M. S. T., S. A. S. Lorensuhewa, and M. A. L. Kalyani. "Identification of abusive sinhala comments in social media using text mining and machine learning techniques." International Journal on Advances in ICT for Emerging Regions, 13.1 (2020): 1.

[9] Itoo, Fayaz, and Satwinder Singh..: Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection. International Journal of Information Technology, 13.4, 1503-1511(2021).

[10] Alakus, Talha Burak, and Ibrahim Turkoglu.: Comparison of deep learning approaches to predict COVID-19 infection. Chaos, Solitons & Fractals 140,110120(2020).

[11] Chen, Sheng, and Haibo He.: Nonstationary stream data learning with imbalanced class distribution. Imbalanced learning. Foundations, algorithms, and applications, 151-186 (2013).

***