Journal of Robotics and Control (JRC)
Volume 3, Issue 3, May 2022
ISSN: 2715-5072, DOI: 10.18196/jrc.v3i3.14387

Early Diagnosis for Dengue Disease Prediction Using Efficient Machine Learning Techniques Based on Clinical Data

Bilal Abdualgalil [1*], Sajimon Abraham [2], Waleed M. Ismael [3]
[1, 2] School of Computer Sciences, Mahatma Gandhi University, Kottayam, Kerala, India
[3] Hohai University, Chaozhou campus, Jiangsu, China.
E-mail: [1] bsaa85@gmail.com, [2] sajimabraham@rediffmail.com, [3] Waleed.m@hhu.edu.cn
*Corresponding Author

***Abstract*—Dengue fever is a worldwide issue, especially in Yemen. Although early detection is critical to reducing dengue disease deaths, accurate dengue diagnosis requires a long time due to the numerous clinical examinations. Thus, this issue necessitates the development of a new diagnostic schema. The objective of this work is to develop a diagnostic model for the earlier diagnosis of dengue disease using Efficient Machine Learning Techniques (EMLT). This paper proposed prediction models for dengue disease based on EMLT. Five different efficient machine learning models, including K-Nearest Neighbor (KNN), Gradient Boosting Classifier (GBC), Extra Tree Classifier (ETC), eXtreme Gradient Boosting (XGB), and Light Gradient Boosting Machine (LightGBM). All classifiers are trained and tested on the dataset using 10-Fold Cross-Validation and Holdout Cross-Validation approaches. On a test set, all models were evaluated using different metrics: accuracy, F1-sore, Recall, Precision, AUC, and operating time. Based on the findings, the ETC model achieved the highest accuracy in Hold-out and 10-fold cross-validation, with 99.12 % and 99.03 %, respectively. In the Holdout cross-validation approach, we conclude that the best classifier with high accuracy is ETC, which achieved 99.12 %. Finally, the experimental results indicate that classifier performance in holdout cross-validation outperforms 10-fold cross-validation. Accordingly, the proposed dengue prediction system demonstrates its efficacy and effectiveness in assisting doctors in accurately predicting dengue disease.**

Keywords—Dengue Disease; Machine Learning; Extra Tree; SMOTE+ENN; balanced dataset

I. INTRODUCTION

Dengue fever is a mosquito-borne viral disease that spreads quickly in warm weather. It is transferred by a female mosquito known as 'Aedes aegypti.' Variations in rainfall, temperature and unplanned rapid urbanization are the primary causes of dengue disease's extensive prevalence in the tropics. Dengue cases have increased dramatically in recent years over the world. However, the actual number of dengue infections is either never recorded or is classified incorrectly. According to a WHO report, 390 million dengue infections are recorded worldwide each year, with 96 million of these clinically confirmed with disease severity [1].

Only 9 countries experienced severe dengue epidemics before 1970. Nowadays, more than 100 countries in the WHO regions of Africa, the Americas, the Eastern Mediterranean, South-East Asia, and the Western Pacific have the disease now. The Americas, Southeast Asia, and the Western Pacific are the most severely affected, with Asia accounting for 70% of the worldwide disease burden [2].

In another study on dengue disease, dengue viruses can infect more than 3.9 billion people in 128 countries [3]. As stated by the Healthcare sector in Taiz city, Yemen, 4,770 cases of dengue fever had been detected in Taiz governorate in the southwestern part of the country from the beginning of the year through the end of July 2021 [4].

Dengue fever has been recognized as endemic in Yemen for over 25 years, with the first case appearing in Taiz Governorate in 1994. In recent years, the disease has taken a different path, appearing as a severe type of DHF with an increased frequency of outbreaks. A previously unimmune host infected with Dengue develops a secondary antibody response characterized by a delayed and low-titer response. The first immunoglobulin isotype to appear is the IgM antibody. The presence of anti-dengue IgM antibodies in a probable dengue patient indicates a recent infection. One of the most significant advancements has been the detection of anti-dengue IgM using enzyme-linked immunosorbent assay (ELISA), which has become a powerful tool for routine dengue diagnosis [5].

Dengue fever is one of the world's most fatal and widespread viral infections. It is a rapidly spreading tropical virus infection with an increased death rate [6]. Recent years have seen the development of many decision support systems and diagnostic models to help physicians detect and diagnose diseases more accurately. In recent years, artificial intelligence has been increasingly utilized in the field of medical data mining, and numerous decision support systems have been developed by leveraging machine learning and deep learning.

Artificial intelligence-based technologies could improve health outcomes and the quality of life for millions of people in the coming few years. These techniques also work well for predicting dengue diseases, however, there is still a challenge in using machine learning techniques to predict dengue diseases based on clinical data that is mostly imbalanced, as well as the selection of important features; all of these factors affect the accuracy of models in predicting dengue diseases.

To our knowledge, no study has yet used artificial intelligence approaches with data rebalancing methods for the rapid diagnosis of dengue, and this work requires the use of effective machine learning techniques to improve results.

The major contribution of this research is the development of a diagnostic model for the early diagnosis of dengue disease by integrating machine learning techniques into the proposed diagnostic system for precise dengue patient diagnosis. Key steps include:

- Adopting machine learning techniques like KNN, GBC, XGB, LightGBM, and Extra tree for accurate detection of dengue patients.
- Developing a diagnostic model based on machine learning techniques to help physicians in the early detection and diagnosis of dengue disease.
- Validating the results of the proposed diagnostic model using K-fold and holdout cross-validation approaches.

This paper is organized as follows: Section II summarizes the related works in the field of dengue disease diagnosis. The proposed system is explained in Section III. The results of the work are presented in Section IV. Finally, the entire work is concluded in Section V.

II. RELATED WORK

This section discusses the related works of the machine learning algorithm for dengue disease prediction. Marimuthu et al. [7] proposed a bio-computational methodology for mapping gene sequences to construct dengue viral association. It achieved 96.74 % by establishing classification and association rules using standard tools.

Rao, K., K., N., et al. [8] proposed a decision tree-based algorithm to find association rules. The purpose of this research was to explain the significance of association rules in predicting the disease by analyzing the features of affected patients. The model achieved 97% accuracy.

P. Manivannan et al. [9] developed a classification and clustering model to detect dengue viruses using patient data from several Indian states.

Shaukat, K., et al. [10] used the DBSCAN algorithm for dengue fever clustering to illustrate the overall behavior of dengue in the district Jhelum and evaluated several clustering algorithms using graphs based on the dataset. Algorithms such as k-means, K-medoids, DBSCAN, and OPTICS are examples of these.

N. A. Husin et al. [11] proposed a model based on environmental features using the support vector machine. For feature selection, the model used PCA, and for model execution, it used c-SVM with the Gaussian kernel. The model outperformed their earlier efforts in terms of accuracy.

Subitha et al. [12] presented a mining model for dengue fever. in this paper, they implemented KNN, and mining performance was improved. To improve dengue fever results, they used the model to segment microscopic blood images using a neural network. The classification result was analyzed using a backpropagation network. It has an accuracy of 98 %.

Buchade Omkar et al. [13] explained the tests that are conducted on blood samples acquired from patients. Dengue fever is classified into three types in the proposed system: Dengue Fever (DF), Dengue Hemorrhagic Fever (DHF), and healthy patients. Existing work used the PSO technique, which achieved an accuracy of 90.91 %. To achieve high accuracy, they used optimization algorithms such as Spider Monkey Optimization (SMO), and to boost the optimality of the model, they use Probabilistic Neural Networks (PNN). For classification, PNN utilizes the feed-forward technique.

Martinez et al. [14] developed a technique for detecting dengue disease based on blood pressure, viral infection, sex, and age. It trained the model on existing data using Nave Bayesian classification and WAC 55. This model can also be used by patients and nurses to provide features and predict disease occurrence.

M.Bhavani et al. [15] used "a data-driven epidemiological prediction method for dengue epidemics using local and remote sensing data" to develop their method. The author employed Fuzzy Association Rule Mining as a prediction method to extract correlations between clinical, meteorological, climatic, and other variables. The data collection contains dengue case data from 2001 to 2009, obtained from the Peruvian Ministry of Health. Dengue fever is predicted three to four weeks in advance. Positive, negative, sensitivity and specificity values for test data collected 4–7 weeks after prediction were 0.686, 0.976, 0.615, and 0.982, respectively.

Nishanthi et al.[16] conducted a survey titled "A survey Prediction & Detection of Dengue – Mining Methods & Techniques." Classification approaches such as Naive Bayesian, REP Tree, Random Tree, J48, SMO, SVM, Decision Tree Approach, and Spatial Data Analysis, among others, were utilized to arrange datasets. The dengue prediction data set is DNA microarray data, which contains information on gene expression responsible for the dengue virus. When the techniques were compared, it was discovered that Naive Bayes stands out above the others, delivering an accuracy of 92 % with high probability and effectiveness.

Mulyani et al. [17] develop a specific framework for accurate dengue fever prediction that combines the Dempster Shafer (DS) and NB processes. The rules are constructed using ES and Dempster Shafer's theory, and the machine learning part is implemented using NB. The proposed ES achieves 70% accuracy during the training phase and 56% accuracy during the testing phase.

A prediction model developed by Shukat et al. [18] takes into account various symptoms such as fever, bleeding, flu, and a few more. The authors discovered that J48 and NB outperformed REP Tree, Random Tree, and SOM techniques in accurately predicting dengue.

Siriyasatien et al. [19] publish a report on dengue risk variables that uses the K-H model, SVM, and ANN for dengue prediction. When evaluating dengue infection, many factors are considered important, including temperature, rainfall, humidity, wind speed, aedes aegypti larvae infection rate, female mosquito infection rate, male mosquito infection rate, season, and population attributes.

Table I shows a brief overview of work related to dengue. In their research, the authors focused on using machine learning and deep learning techniques in dengue disease prediction, and on dengue datasets from different geographic locations with a specific number of samples, features, and evaluation measures like accuracy, AUC, and F1-score.

However, certain criteria, such as the selection of important features and the handling of an imbalanced dataset, have not been implemented in researchers' previous work. Whereas, in this work, we contributed relevant specific features based on the patient's personal, diagnostic, and symptom data. In addition to selecting the most important features from the dengue dataset and rebalancing it.

III. PROPOSED METHODOLOGY

The objective of this work is to develop a prediction model for predicting dengue disease, as shown in Fig. 1. The steps in the proposed system are as follows.

1) Obtaining the dengue dataset,
2) Pre-processing and cleaning the obtained dataset before being used to build models,
3) Splitting the processed dataset into training and testing sets,
4) Applying five machine learning algorithms to build predictive models for predicting dengue, and
5) Feeding the testing set into the model to evaluate its performance.
6) Using the developed model to predict dengue at this stage,
7) Evaluating the results from all methods, and
8) Comparing the results to determine the best algorithm.

TABLE I. BRIEF REVIEW OF DENGUE RELATED WORK PAPERS

Work	Dengue Dataset Location	No. of Samples	No. of features	Classifier	Accuracy	AUC	F1-Score
[20]	Real-life Hospital data of Dengue patients	**110**	**16**	**PSO-ANN**	**87.27%**	**0.823**	-
[21]	Chittagong Medical College and Dhaka Medical College Hospital.	**209**	**23**	**DT**	**79%,**	-	**0.79**
[22]	Several hospitals in Delhi region	**110**	**17**	**ANN**	**79.09%**	-	-
[23]	Reports of different discharged patients.	**75**	**9**	**Logit Boost**	**92%**	**0.6250**	**0.967**
[24]	Health department, Karuna medical hospital	**100**	**11**	**RF**	**83.3%**	-	-
[25]	Health Center is located in the Thanjavur District of South Tamilnadu, India	**480**	**20**	**EWSORA+MLP**	**98.72%**	**0.89**	**0.94**
[26]	The public health system of Paraguay	**4332**	**38**	**ANN-MLP**	**96%**	-	-

A. Experimental setup

Our experimental models were implemented on Windows 10 O.S, running on Intel® Core(TM) i5- 8250U CPU 1.80 GHz 4 processor, 8-GB-RAM, and a 2-TB hard drive, using Python 3.7.1

B. Dataset Description

In this work, the dataset was obtained from Epidemiological Monitoring Center (EMC) - Public Health and Population Office (PHPO)-Taiz city, YEMEN. The clinical data were collected for three years from 2017 to 2019 as described in Table II. Table III presents the normalized values for all attributes of the dengue dataset. Moreover, the visual representation of the attributes of the dengue dataset is illustratively demonstrated in Fig. 2.

TABLE II. DENGUE DATASET DESCRIPTION

Data set	No. of samples	Input Attribute	Output Attribute	Output Classes	Total No. of Attributes	Missing attributes status	Noisy attributes status
Dengue fever	6694	21	1	2	22	1380	no

TABLE III. NORMALIZED VALUES OF DIFFERENT ATTRIBUTES OF THE DENGUE DATA SET

SN.	Feature Name	Value
1	Age	Continues
2	Sex	1- Male, 0 - Female
3	Fever	1-yes, 0-no
4	Headache	1-Yes,0-no
5	Arthralgia	1-yes, 0-no
6	Myalgia	1-yes, 0-no
7	Conjunctivitis or Pain behind eyes	1-yes, 0-no
8	Skin rash	1-yes, 0-no
9	Generalized weakness	1-yes,0-no
10	Jaundice	1-yes, 0-no
11	Decrease of urine or anuria	1-yes, 0-no
12	Abdominal pain	1-yes,0-no
13	Vomiting	1-yes, 0-no
14	watery diarrhea	1-yes,0-no
15	Ecchymosis	1-yes,0-no
16	meningitis	1-yes 0-no
17	Respiratory tract infection or respiratory insufficieny	1-yes 0-no
18	Convulsions , coma	1-yes,0-no
19	Kidney failure	1-yes, 0-no
20	IgM	1-yes, 0- no
21	IgG	1-yes, 0- no
22	Dengue	1-positive, 0-negative

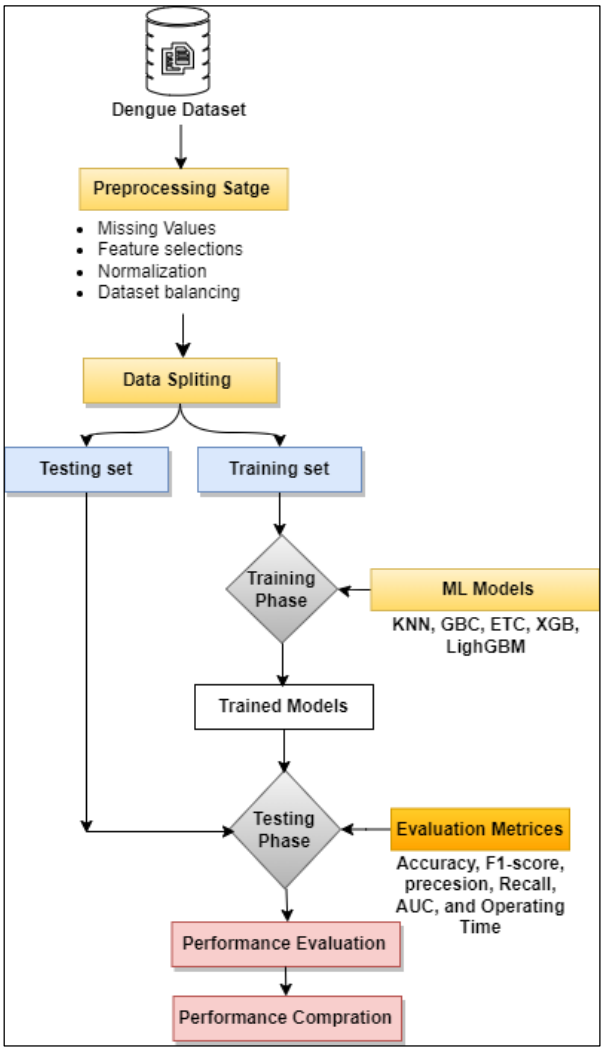


Fig. 1. Proposed dengue disease prediction model

C. *Pre-processing*

Data pre-processing and cleaning are important steps in data handling before it is used in machine learning algorithms. The dengue dataset is available in .xlsx format for download. This stage consists of a sequence of processes that occur after the dataset has been read:

a) Missing Data Handling

Real-world data contains missing values and noise, and they are also in a raw format that cannot be directly used to develop machine learning models. To convert such missing values and noisy data into a machine-readable format, data pre-processing methods such as data cleaning and formatting are required. In this work, the initial stage in data pre-processing was the handling of missing data.

In our experimental dataset, we found the set of features that have missing values, as shown in Fig. 3. The mean method was used to replace the missing values for handling these features.

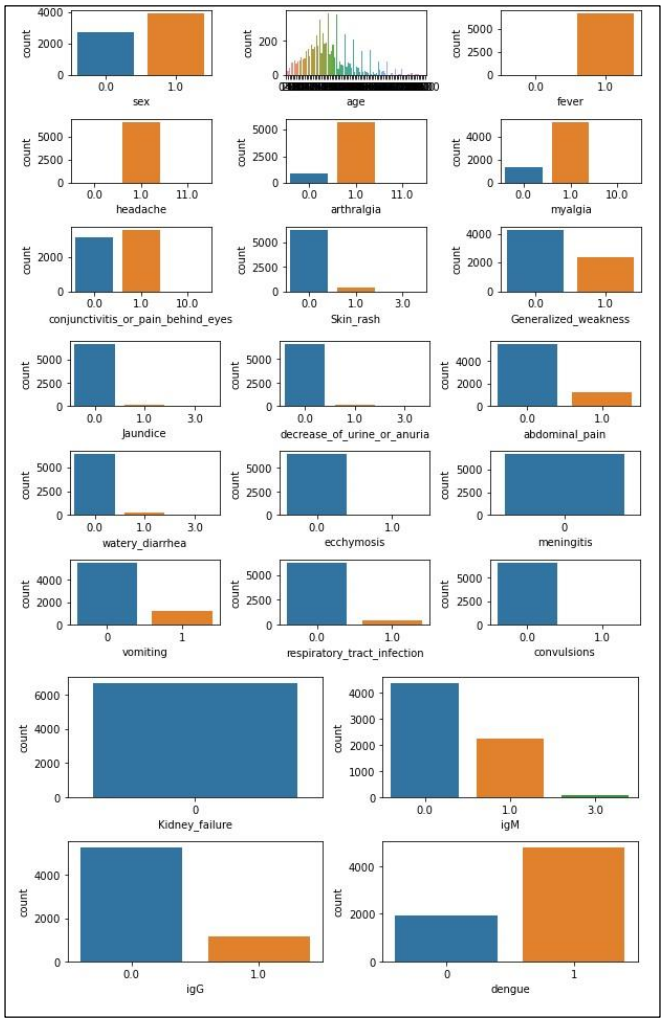


Fig. 2. Visual representation of dengue dataset attributes.

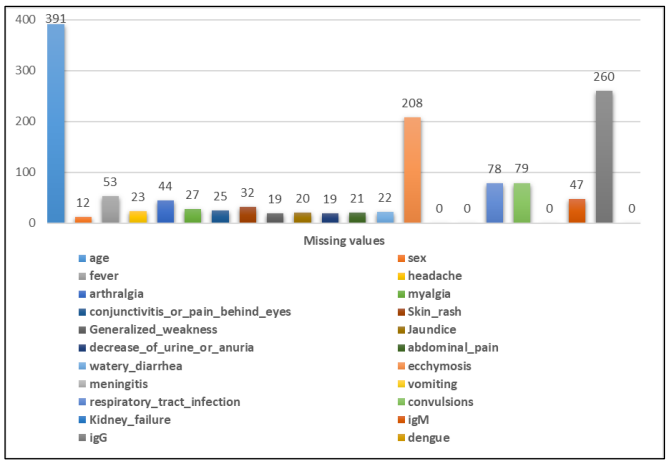


Fig. 3. Dengue dataset with missing values

b) Features Selection

To perform feature selection, the Extra Tree method was used to employ feature ranking. Fig. 4 shows the feature importance predicted by the Extra Tree method. The Extra Tree (ET) method identifies 19 features that are used in this work as the most relevant features except for meningitis and kidney failure features.

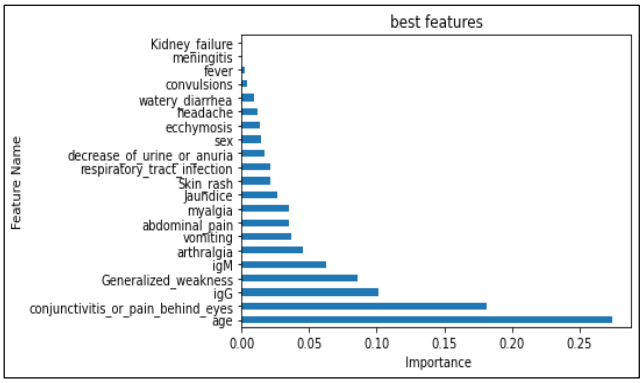


Fig. 4. Important features by the Extra Tree method.

c) *Data Normalization*

A normalization is a pre-processing approach that makes all features to a single scale, that is, to the same minimum, maximum, and medium values, without distorting the differences in the value ranges. To standardize the dengue dataset, we apply the Z-Score Normalization approach [27]. This technique employs the mean and standard deviation of each feature of training data to normalize each input feature vector. Equation (1) gives the mathematical formula for Z-Score normalization, where Z is the normalized attribute value, x_i is the original attribute value, μ is the mean, and σ is the standard deviation.

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

d) *Data balancing*

Using an imbalanced dataset to train machine learning models can result in a bias toward the majority class. To avoid this bias, a more balanced dataset must be used. The SMOTE+ENN hybrid approach is employed in this work to rebalance the dataset. It was developed by [28]. It combines the synthetic minority oversampling method (SMOTE) and the Edited Nearest Neighbors (ENN) method. SMOTE is the most widely used oversampling technique, and it can be used with a variety of under-sampling techniques. SMOTE selects a random sample of minority class examples. To build a synthetic example, we selected the sample's nearest k neighbors and chose a point at random inside that region.

ENN works by selecting examples to be removed. Using k=3 nearest neighbors, this rule locates and deletes misclassified cases in a dataset. The new balanced dataset is as follows after applying the SMOTE+ENN hybrid technique:

- Fig. 5(b) compares the resampled dataset shape Counter to the original dataset, which was as follows:

  (1: 3290, 0: 3211)

- Fig. 5(a) depicts the original dataset shape Counter

  (1: 4782, 0: 1912).

SMOTE and its extensions, such as SMOTE + Tomek and adaptive synthetic sampling, were also implemented (ADASYN). However, the best results were obtained by combining SMOTE with an ENN modification.

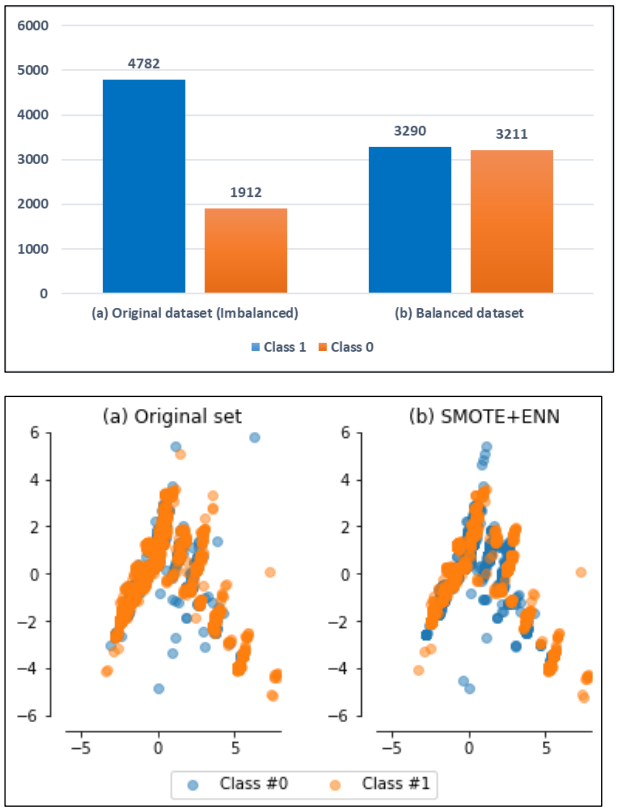


Fig. 5. Dataset (a) Original (Imbalanced) dataset, and (b) dataset after SMOTE+ENN hybrid technique (Balanced data set)

e) *Data Splitting*

In the data Splitting stage, we applied two approaches to split the dataset into a training set and a testing set after pre-processing stage. The first method is Holdout cross-validation, in which we divided the data set into 70% for training and 30% for testing, and the second method is 10-fold cross-validation. The training data is fed into the machine learning model to train the model. The dengue class (Class 1: Positive, Class 0: Negative) feature is used as the target variable in the prediction classifier.

f) *Machine learning Methods*

For data classification, a supervised machine learning model is used to predict the result. This work presents a technique for predicting dengue disease using classification techniques. As described in the data Splitting section, the data has been divided into a training set and a test set. The efficiency of the classifiers is evaluated using test data. The following discusses the specifics of machine learning classifiers are used in this work.

KNN [29] According to this algorithm, data are sampled based on k, which shows the neighbors. Based on similarity measures, new samples are classified based on the stored data. Distance is measured between data points and the nearest most data points are considered neighbors. Distance between data points is measured using different distance measures. For the calculation of distance, we used Euclidean distances. Equation (2) consists of two data points called a and b. The distance between them should be measured.

$$U_d = \sqrt{\sum_{i=1}^{k} (a_i + b_i)^2} \quad (2)$$

GBC [30] combines several weak learners to form a strong ensemble to perform supervised tasks (classification and regression). In GBC, new models are fitted sequentially to improve the accuracy of the response variable estimate. In this algorithm, the new base-learners are constructed to have a negative gradient associated with the loss function of the whole ensemble to be as correlated as possible.

XGB [31] The Extreme Gradient Boosting model (XGBoost) is a supervised classification and regression model. The accuracy of an XGBoost model is determined by the information of both the XGBoost objective function and the basic trainees. Also, the XGBoost model is effective in time-series problems by transforming time-series forecasting data into a supervised learning problem. The mathematical representation of the XGBoost model's formation is given in equation (3).

$$Obj_m = \sum_{i=1}^{n} l((y_i, y_i^{m-1}) + f_m(x_i) + \Omega(f_m) \quad (3)$$

Where n the total number of the trees is, m denotes the number of iterations, and f_m is the error in the m iterations. l is the loss cost function to compute the label and prediction difference in the last stage. Additionally, the output of the new tree, and Ω is the function used for regularization to prevent overfitting in equation (4). T is the number of leaves per tree, and w is the weight of the leaves of each tree.

$$\Omega(f_m) = \gamma T + \frac{1}{2}\gamma \|w\|^2 \quad (4)$$

ETC [32] classifier is a type of ensemble classifier that outperforms all existing tree-based classifiers such as Decision Tree (DT) and Random Forest Classifier in terms of performance (RFC). In this classifier, the decision tree for classification is formed by first forming the root node. The root node is chosen by inspecting the randomly formed subset of available features, as shown in equation (5). Because the ET Classifier depicts both DT and RF, it makes its decision based on entropy and information gain.

$$N = \beta \quad (4)$$

Where N represents the root node, and β represents the number of quad root features provided to the model.

LightGBM [33] is a framework that improves the classification model's efficiency while consuming less memory in a decision tree-based gradient boosting framework. Two novel techniques, Gradient-based One Side Sampling and Exclusive Feature Bundling (EFB) are used to overcome the limitations of the histogram-based algorithm used in all GBDT (Gradient Boosting Decision Tree) frameworks. The two techniques of GOSS and EFB define the characteristics of LightGBM. Their collaboration enables the model to operate efficiently and set itself apart from other GBDT frameworks.

All classifiers models used in this study are made available in the Python-based sci-kit-learn package, while ensemble models such as XGBoost and LightGBM are available in Python library "*xgboost*" and *"lightgbm"*, 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. We used a trial and error method to settings Hyper-parameters to train each machine learning classifier as shown in Table IV in detail. After training the classifiers, the model predicts dengue disease using the testing data.

TABLE IV. HYPER-PARAMETERS SETTINGS OF CLASSIFICATION METHODS

No.	Model	Hyper-parameters settings
1	GBC	n_estimators=100,learning_rate=0.1,max_depth=5,random_state=33
2	XGB	learning_rate =0.75, n_estimators=1000, max_depth=7, min_child_weight=1, gamma=0.1, subsample=0.8, colsample_bytree=0.8, objective= 'binary:logitraw', nthread=4, scale_pos_weight=1, seed=27
3	ETC	n_estimators=100, max_features=19
4	LightGBM	boosting_type='gbdt',n_estimators=1000, learning_rate=0.1
5	KNN	(n_neighbors= 3,weights ='uniform', algorithm='auto'

g) Evaluation metrics

There are some methods for evaluating the performance of machine learning models. Analytical research will be supported by the use of different evaluation tools [34]. In this work, we used six basic metrics (accuracy [35], precision [36], recall [37], F-Score [38], AUC [39], and time) to explore the differences between machine learning algorithms. The confusion matrix [40] helps us in calculating all metrics except for Time. True positive (TP), true negative (TN), false positive (FP), and false-negative (FN) are the elements of the confusion matrix. When it comes to health care data, the most important prediction is a false negative. All performance measures are used to evaluate all models in this work and which are mathematically represented in equations (6)-(11).

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)} \times 100 \quad (6)$$

$$\text{Re}call = \frac{TP}{(TP + FN)} \times 100 \quad (7)$$

$$\text{Pr}ecision = \frac{TP}{(TP + FP)} \times 100 \quad (8)$$

$$F1\text{-}score = 2 \times \left(\frac{Precision * Recall}{Precision + Recall} \right) \times 100 \quad (9)$$

$$True\ Positive\ Rate\ (TPR) = \frac{TP}{TP + FN} \times 100 \quad (10)$$

$$False\ Positive\ Rate\ (FPR) = \frac{FP}{TN + FP} \times 100 \quad (11)$$

Where Time(s) refers to model operating time

IV. RESULTS AND DISCUSSION

Results of all experiments for dengue disease prediction are discussed in this section using five machine learning techniques are (Gradient Boosting classifier (GBC), Extra Tree Classifier (ETC) and eXtreme Gradient Boosting (XGB), and Light Gradient Boosting Machine (LightGBM), k-nearest neighbors (KNN)). All models are tested on the same dataset and evaluated using the same metrics.

A. Experimental Results with Balanced data

The performance of all machine learning models was evaluated using the Holdout cross-validation approach and 10-fold cross-validation approach as shown in Tables V and VI respectively.

This work 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 Cross-Validation and a 10-fold Cross-Validation approach. The AUC values were calculated for all machine learning models with the Holdout cross-validation approach, as shown in Fig. 6. And all machine learning models were evaluated with a 10-fold cross-validation approach shown in Fig. 7.

The models that achieved the highest values based on the evaluation metrics were the ETC model with 99.03% accuracy followed by 99.04% f1-score, 98.92% precision, 99.17% recall, and 97.69% AUC and 9.624 (s) operating time in 10-fold Cross-Validation approach, and 99.12% accuracy followed by 99.13% f1-score, 99.08% precision, 99.18% recall, and 99.12% AUC, and 0.637 (s) with operating time in Holdout Cross-Validation approach.

Also, the Confusion matrix for all machine learning models was evaluated with the Holdout cross-validation approach and 10-fold cross-validation approach as shown in Tables VII and VIII, respectively.

Finally, the ETC model achieved the best models for dengue prediction based on machine learning techniques with both of 10-Fold Cross-Validation and Holdout cross-validation approaches. Overall, all machine learning techniques performed well in the prediction of dengue data, based on the results of the overall experiment.

B. Experimental Results with the original dataset (Imbalanced dataset)

Also, the performance of all machine learning techniques was compared in the case of the original dengue dataset (imbalanced dataset before rebalanced) in both holdout and 10-fold cross-validation approaches, as shown in Tables IX and X respectively. The results of this experiment are the performance of all classifiers in an imbalanced dataset case, as shown in Fig. 8 and 9 respectively.

The model that achieved the highest values based on the evaluation metrics was the GBC model with 85.71% accuracy followed by 90.16% f1-score, 88.19% precision, 92.21% recall, and 81.01% AUC, with 0.653 (s) operating time in Holdout Cross-Validation approach, and 85.72% accuracy followed by 90.25% f1-score, 88.27% precision, 92.34% recall, and 79.03% AUC, with 9.693 (s) operating time in 10-fold Cross-Validation approach.

Overall, the GBC model achieved the best models for dengue prediction based on machine learning techniques with both of 10-Fold Cross-Validation and Holdout cross-validation approaches with the imbalanced dataset. Comparison performance of all classifiers with both of Holdout cross-validation approach and 10-Fold Cross-Validation with the balanced and imbalanced dataset as shown in Fig. 10 and 11 respectively.

Finally, all machine learning techniques with balanced dataset cases outperformed the imbalanced dataset for dengue disease prediction. We conclude from this work, all machine learning models performed well in the prediction of dengue data, based on the results of the overall experiment.

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. Improving the performances of these models may require further adjustments to the hyperparameter values or using effective deep learning techniques.

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 dengue disease prediction using machine learning and deep learning approaches. As a result, we could not generalize our findings for the prediction of any other disease based on clinical data because our framework system is only for dengue prediction using machine learning techniques. This could be a possible direction for future scope.

TABLE V. EVALUATION METRICS FOR ALL CLASSIFICATION MODELS IN THE HOLDOUT CROSS-VALIDATION APPROACH

Model	Accu-racy	F1-Score	Pre-cision	Re-call	AUC	Time (s)
KNN	98.56	98.57	98.17	98.97	98.56	0.187
GBC	97.64	97.63	98.34	96.93	97.64	0.569
XGB	98.51	98.51	98.86	98.15	98.51	2.11
ETC	**99.12**	**99.13**	**99.08**	**99.18**	**99.12**	**0.637**
LightGBM	98.82	98.82	98.87	98.77	98.82	0.88

TABLE VI. EVALUATION METRICS FOR ALL CLASSIFICATION MODELS IN A 10-FOLD CROSS-VALIDATION APPROACH

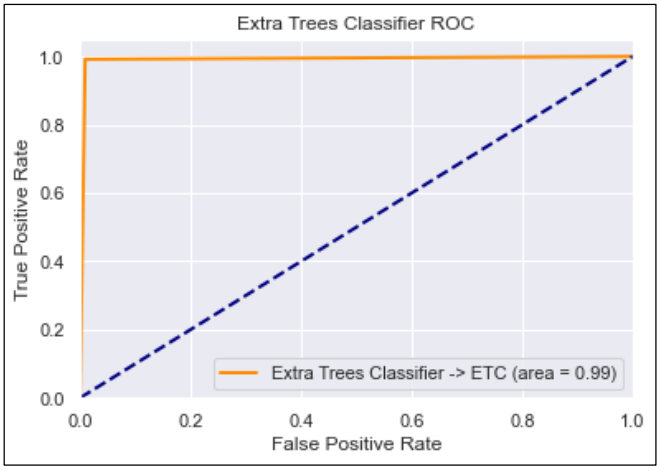
Model	Accu-racy	F1-Score	Pre-cision	Re-call	AUC	Time (s)
KNN	98.26	98.29	98.32	98.27	95.79	0.634
GBC	97.29	97.34	97.37	97.31	96.51	9.934
XGB	98.35	98.37	98.48	98.26	96.67	52.886
ETC	**99.03**	**99.04**	**98.92**	**99.17**	**97.69**	**9.624**
LightGBM	98.74	98.76	98.66	98.87	96.82	17.617

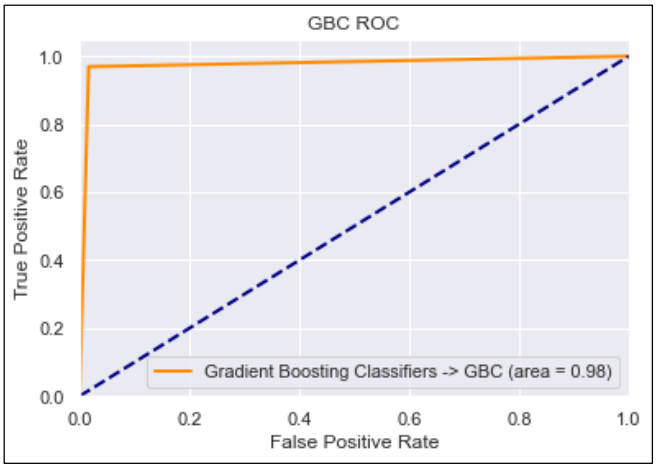
TABLE VII. CONFUSION MATRIX FOR XGB, ETC, GBC, KNN, AND LIGHTGBM APPROACHES IN HOLDOUT CROSS-VALIDATION METHOD.

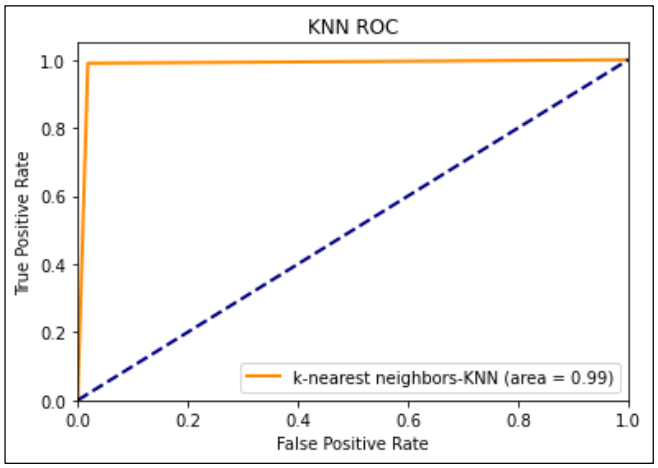
	XGB		ETC		GBC		KNN		LightGBM	
	Predicted		Predicted		Predicted		Predicted		Predicted	
Actual	962	11	964	9	957	16	955	18	962	11
	18	960	8	970	30	948	10	968	12	966

TABLE VIII. CONFUSION MATRIX FOR XGB, ETC, GBC, KNN, AND LIGHTGBM APPROACHES IN THE 10-FOLD CROSS-VALIDATION METHOD

	XGB		ETC		GBC		KNN		LightGBM	
	Predicted		Predicted		Predicted		Predicted		Predicted	
Actual	947	26	951	22	940	33	928	45	941	32
	39	939	23	955	35	943	37	941	30	948







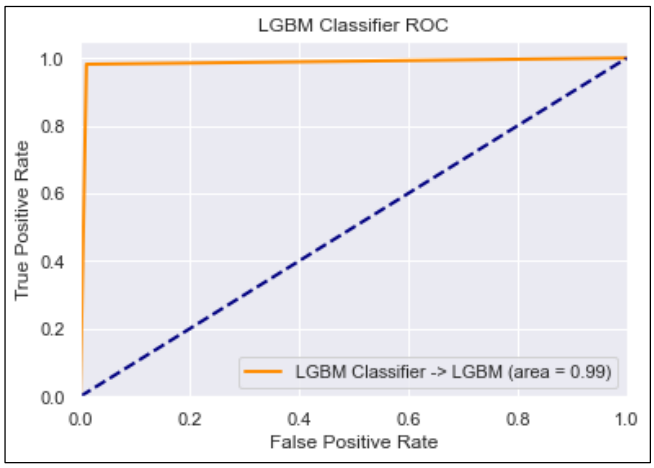
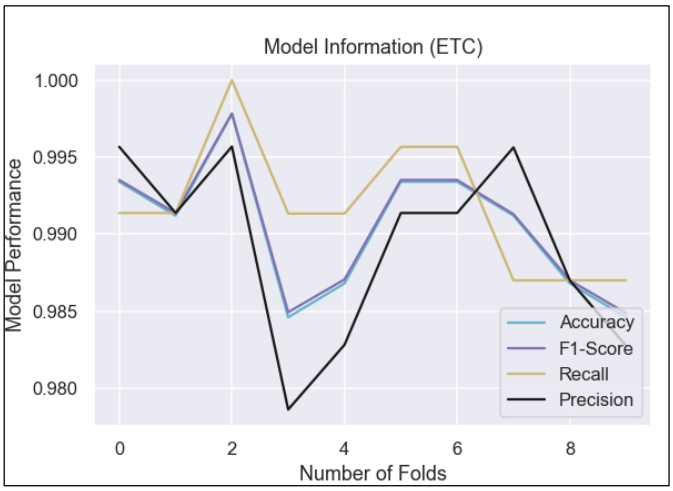
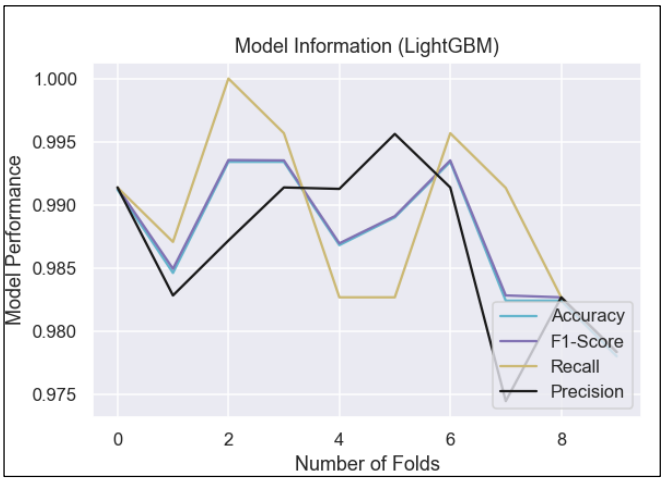
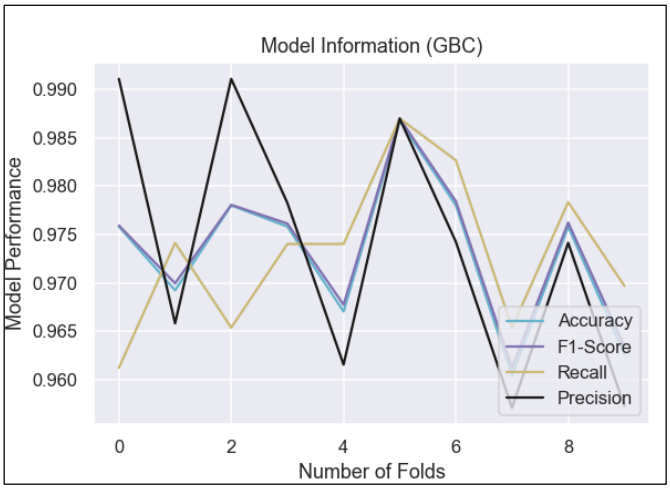
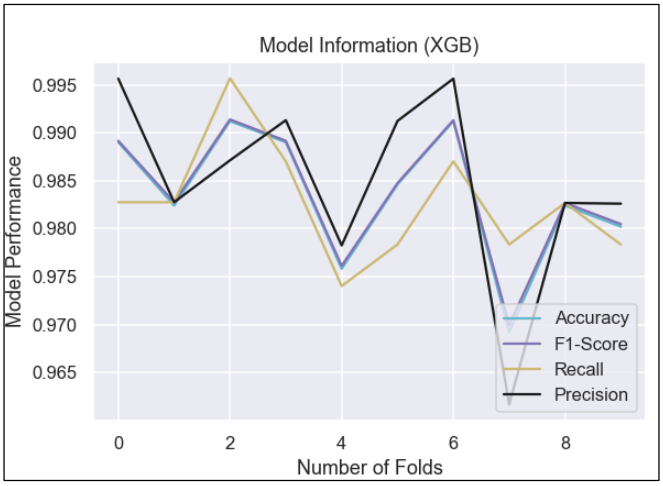


Fig. 6. The AUC values were calculated for all machine learning models with the Holdout cross-validation approach









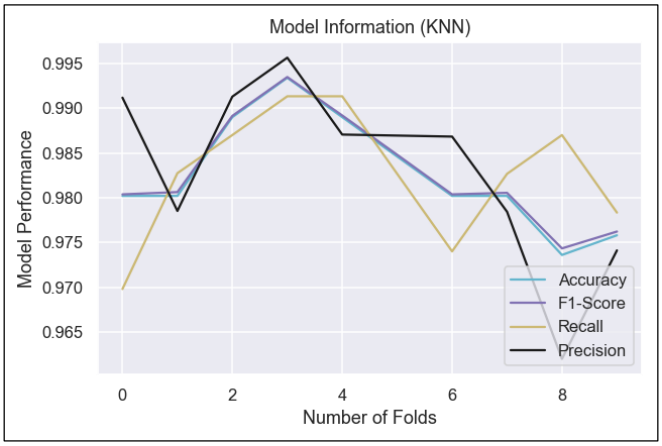


Fig. 7. All classification models were evaluated with a 10-fold cross-validation approach

TABLE IX. EVALUATION METRICS FOR ALL CLASSIFICATION MODELS IN THE HOLDOUT CROSS-VALIDATION APPROACH IN THE CASE OF THE ORIGINAL DATASET (WITHOUT SMOTE+ENN HYBRID TECHNIQUE).

Model	Accuracy	F1-Score	Precision	Recall	AUC	Time (s)
KNN	83.92	88.66	88.75	88.56	80.56	0.236
GBC	**85.71**	**90.16**	**88.19**	**92.21**	**81.01**	**0.653**
XGB	82.82	87.78	88.63	86.95	79.84	3.275
ETC	83.72	88.44	89.11	87.79	80.77	0.825
Light GBM	84.12	88.9	88.19	89.62	80.14	1.073

TABLE X. EVALUATION METRICS FOR ALL CLASSIFICATION MODELS IN A 10-FOLD CROSS-VALIDATION APPROACH IN THE CASE OF THE ORIGINAL DATASET (WITHOUT SMOTE+ENN HYBRID TECHNIQUE).

Model	Accuracy	F1-Score	Precision	Recall	AUC	Time (s)
KNN	82.71	88.07	87.01	89.18	77.69	0.595
GBC	**85.72**	**90.25**	**88.27**	**92.34**	**79.03**	**9.693**
XGB	83.13	88.17	88.54	87.81	76.36	67.377
ETC	84.16	88.89	89.26	88.55	78.97	12.956
LightGBM	84.65	89.37	88.6	90.16	76.02	18.003

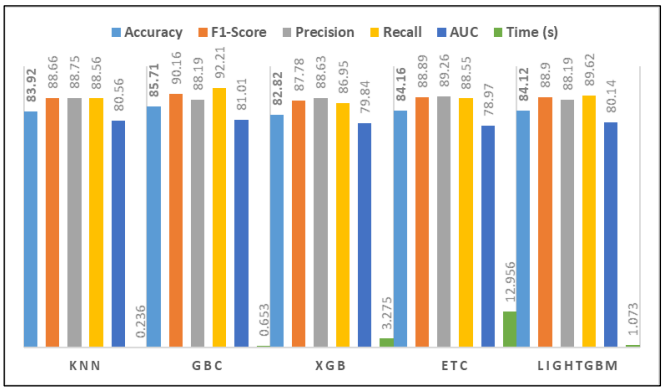


Fig. 8. Evaluation metrics for all classification models in the Holdout cross-validation approach in the case of the original dataset (without SMOTE+ENN hybrid technique).

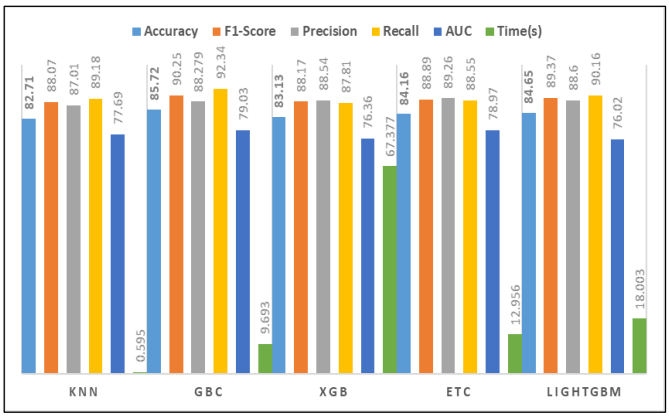


Fig. 9. Evaluation metrics for all classification models in a 10-fold cross-validation approach in case of the original dataset (without SMOTE+ENN hybrid technique).

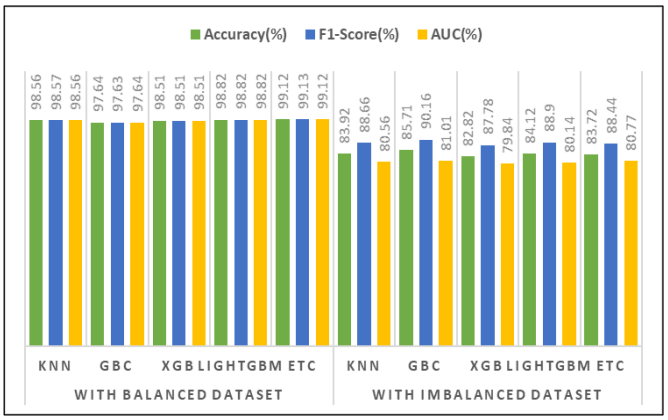


Fig. 10. Comparison of all machine learning models with balanced and Imbalanced datasets in a Holdout cross-validation approach.

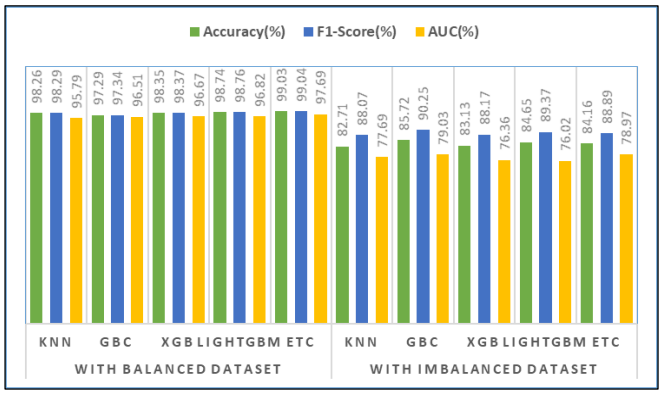


Fig. 11. Comparison of all machine learning models with balanced and Imbalanced datasets in a 10-fold cross-validation approach.

V. CONCLUSION AND FUTURE WORK

Dengue infection is a global problem today. The early detection and prevention of dengue can help to avoid complications and save human lives. In this paper, we proposed a framework for dengue prediction and evaluated the performance of five machine learning models for predicting dengue (KNN, GBC, XGB, ETC, and LighGBM).

In the initial stage of the work, namely the pre-processing stage, the missing values were processed by the mean method. The selection features technique was applied to select the important features. The data were normalized by Z-Score. For the imbalance problem of the dataset, we used the SMOTE+ENN hybrid technique. We applied cross-validation approaches such as 10-fold and Holdout cross-validation to split the dataset into training and testing sets. After that machine learning models were built, and their performance was evaluated using accuracy, F1-score, precision, recall, AUC scores, and operating time (s).

The experimental results show that the performance of the dengue prediction system is improved with the ETC model with 99.12% accuracy followed by 99.13% f1-score, 99.08% precision, 99.18% recall, AUC 99.12%, 0.637 (s) operating time in the Holdout Cross-Validation approach, and 99.03% accuracy followed by 99.04% f1-score, 98.92% precision, 99.17% recall, and AUC 97.69%, and 9.624 (s) operating time in a 10-fold Cross-Validation approach, we conclude that the ETC model achieved the best performance with holdout cross-validation approach. And in comparison, we conclude that machine learning techniques in the holdout cross-validation approach outperformed machine learning techniques in the 10-fold Cross-Validation approach for dengue disease prediction in the case of the proposed framework in this study.

The results of this work show that combining the SMOTE+ENN hybrid and selecting important feature methods improved the framework's accuracy in making clinical decisions in accurately predicting dengue and any other disease with different datasets.

In the future, we plan to extend this research by developing many deep learning techniques with large sizes of data that are likely to accurately predict dengue types.

ACKNOWLEDGMENT

The authors wish to thank the Epidemiological Monitoring Center (EMC) - Public Health and Population Office (PHPO)-Taiz city, YEMEN for giving us the dataset.

REFERENCES

[1] S. Bhatt, P. W. Gething, O. J. Brady, J. P. Messina, A. W. Farlow, C. L. Moyes, J. M. Drake, J. S. Brownstein, A. G. Hoen, O. Sankoh, M. F. Myers, D. B. George, T. Jaenisch, G. R. W. Wint, C. P. Simmons, T. W. Scott, J. J. Farrar, and S. I. Hay, "The global distribution and burden of dengue," Nature, vol. 496, no. 7446, pp. 504–507, Apr. 2013, https://doi.org/10.1038/nature12060.

[2] WHO. (2022, Mar. 11). Dengue and severe dengue Online].Available**:** https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue

[3] I. S. Abubakar, S. B. Abubakar, A. G. Habib, A. Nasidi, N. Durfa, P. O. Yusuf, S. Larnyang, J. Garnvwa, E. Sokomba, L. Salako, R. D. G. Theakston, E. Juszczak, N. Alder, and D. A. Warrell, "Randomised Controlled Double-Blind Non-Inferiority Trial of Two Antivenoms for Saw-Scaled or Carpet Viper (Echis ocellatus) Envenoming in Nigeria," PLoS Neglected Tropical Diseases, vol. 4, no. 7, p. e767, Jul. 2010. https://doi.org/10.1371/journal.pntd.0000767.

[4] A. A. H. Nassar, A. A. Torbosh, Y. A. Mahyoub, and M. A. A. Amad, "Risk Factors Associated With Dengue Fever Outbreak in Taiz Governorate, Yemen, 2018: Case-control Study," Jul. 2021, https://doi.org/10.21203/rs.3.rs-624873/v1.

[5] S. S. Nimmannitya, "Dengue and Dengue Haemorrhagic Fever," Manson's Tropical Diseases, pp. 753–761, 2009, https://doi.org/10.1016/b978-1-4160-4470-3.50045-8.

[6] D. J. GUBLER, "Dengue and Dengue Hemorrhagic Fever," Tropical Infectious Diseases, pp. 813–822, 1997., https://doi.org/10.1016/b978-0-443-06668-9.50077-6.

[7] Marimuthu, T., and V. Balamurugan. "A novel bio-computational model for mining the dengue gene sequences," International Journal of Computer Engineering & Technology, vol. 6, no. 10, pp. 17-33, Oct. 2015.

[8] Rao, NK Kameswara, GP Saradhi Varma, D. Rao, and P. Cse. "Classification rules using decision tree for dengue disease," International Journal of Research in Computer and Communication Technology, vol. 3, no. 3, pp. 340-343, Mar.2014.

[9] P. Manivannan and P. I. Devi, "Dengue fever prediction using K-means clustering algorithm," 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), Mar. 2017, https://doi.org/10.1109/itcosp.2017.8303126.

[10] K. S. Ahmed Bin and S. Kamran Jabbar, "Dengue Fever in Perspective of Clustering Algorithms," Journal of Data Mining in Genomics & Proteomics, vol. 06, no. 03, 2015, https://doi.org/10.4172/2153-0602.1000176.

[11] N. A. Husin, N. Salim, and A. R. Ahmad, "Modeling of dengue outbreak prediction in Malaysia: A comparison of Neural Network and Nonlinear Regression Model," 2008 International Symposium on Information Technology, Aug. 2008, https://doi.org/10.1109/itsim.2008.4632022.

[12] A. Padmapriya and N. Subitha, "Clustering Algorithm for Spatial Data Mining: An Overview," International Journal of Computer Applications, vol. 68, no. 10, pp. 28–33, Apr. 2013, https://doi.org/10.5120/11617-7014.

[13] Omkar, Buchade, Dalsania Preet, Deshpande Swarada, and Doddamani Poonam. "Dengue fever classification using smo optimization algorithm," Int. Res. J. Eng. Technol, vol. 4, no. 10, pp. 1683-1686, 2017.

[14] M. V. Martinez, C. Molinaro, J. Grant, and V. S. Subrahmanian, "Customized Policies for Handling Partial Information in Relational Databases," IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 6, pp. 1254–1271, Jun. 2013, https://doi.org/10.1109/tkde.2012.91.

[15] Bhavani, M., and S. Vinod Kumar. "A data mining approach for precise diagnosis of dengue fever," International journal of latest trends in engineering and technology, vol. 7, no. 4, 2016, https://doi.org/10.21172/1.74.048.

[16] P. H.M.NishanthiHerath, A. A. I. Perera, and H. P. Wijekoon, "Prediction of Dengue Outbreaks in Sri Lanka using Artificial Neural Networks," International Journal of Computer Applications, vol. 101, no. 15, pp. 1–5, Sep. 2014, https://doi.org/10.5120/17760-8862.

[17] Y. Mulyani, E. F. Rahman, Herbert, and L. S. Riza, "A new approach on prediction of fever disease by using a combination of Dempster Shafer and Naïve bayes," 2016 2nd International Conference on Science in Information Technology (ICSITech), Oct. 2016, https://doi.org/10.1109/icsitech.2016.7852664.

[18] K. Shaukat Dar and S. M. Ulya Azmeen, "Dengue Fever Prediction: A Data Mining Problem," Journal of Data Mining in Genomics & Proteomics, vol. 06, no. 03, 2015, https://doi.org/10.4172/2153-0602.1000181.

[19] Siriyasatien, Padet, Atchara Phumee, Phatsavee Ongruk, Katechan Jampachaisri, and Kraisak Kesorn. "Analysis of significant factors for dengue fever incidence prediction," BMC bioinformatics, vol. 17, no. 1, pp. 1-9, Dec. 2016, https://doi.org/10.1186/s12859-016-1034-5.

[20] Gambhir, Shalini, Sanjay Kumar Malik, and Yugal Kumar. "PSO-ANN based diagnostic model for the early detection of dengue disease." New Horizons in Translational Medicine, vol. 4, no.1-4, pp. 1-8, Nov. 2017, https://doi.org/10.1016/j.nhtm.2017.10.001.

[21] Sarma, Dhiman, Sohrab Hossain, Tanni Mittra, Md Abdul Motaleb Bhuiya, Ishita Saha, and Ravina Chakma. "Dengue Prediction using Machine Learning Algorithms." In IEEE 8th R10 Humanitarian Technology Conference (R10-HTC), Kuching, Malaysia, pp. 1-6. IEEE, Dec. 2020, https://doi.org/10.1109/r10-htc49770.2020.9357035.

[22] S. Gambhir, S. K. Malik, and Y. Kumar, "The Diagnosis of Dengue Disease," International Journal of Healthcare Information Systems and Informatics, vol. 13, no. 3, pp. 1–19, Jul. 2018, https://doi.org/10.4018/ijhisi.2018070101.

[23] N. Iqbal and M. Islam, "Machine learning for dengue outbreak prediction: A performance evaluation of different prominent classifiers," Informatica, vol. 43, no. 3, Sep. 2019, https://doi.org/10.31449/inf.v43i3.1548.

[24] Rajathi, N., S. Kanagaraj, R. Brahmanambika, and K. Manjubarkavi. "Early detection of dengue using machine learning algorithms," International Journal of Pure and Applied Mathematics, vol. 118, no. 18, pp. 3881-3887, 2018.

[25] S. A. alias Balamurugan, M. S. M. Mallick, and G. Chinthana, "Improved prediction of dengue outbreak using combinatorial feature selector and classifier based on entropy weighted score based optimal ranking," Informatics in Medicine Unlocked, vol. 20, p. 100400, 2020, https://doi.org/10.1016/j.imu.2020.100400.

[26] J. D. Mello-Román, J. C. Mello-Román, S. Gómez-Guerrero, and M. García-Torres, "Predictive Models for the Medical Diagnosis of Dengue: A Case Study in Paraguay," Computational and Mathematical Methods in Medicine, vol. 2019, pp. 1–7, Jul. 2019, https://doi.org/10.1155/2019/7307803.

[27] S. Malik, S. Harous, and H. El-Sayed, "Comparative Analysis of Machine Learning Algorithms for Early Prediction of Diabetes Mellitus in Women," Lecture Notes in Networks and Systems, pp. 95–106, Sep. 2020, https://doi.org/10.1007/978-3-030-58861-8_7.

[28] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 20–29, Jun. 2004, https://doi.org/10.1145/1007730.1007735.

[29] V. N. Vapnik, "The Nature of Statistical Learning Theory," 1995, https://doi.org/10.1007/978-1-4757-2440-0.

[30] M. H. Lino Ferreira da Silva Barros, G. Oliveira Alves, L. Morais Florêncio Souza, É. da Silva Rocha, J. F. Lorenzato de Oliveira, T. Lynn, V. Sampaio, and P. T. Endo, "Benchmarking of Machine Learning Models to Assist the Prognosis of Tuberculosis," Apr. 2021, https://doi.org/10.20944/preprints202103.0284.v2.

[31] T. Chen and C. Guestrin, "XGBoost," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug. 2016, https://doi.org/10.1145/2939672.2939785.

[32] A. Sharaff and H. Gupta, "Extra-Tree Classifier with Metaheuristics Approach for Email Classification," Advances in Computer Communication and Computational Sciences, pp. 189–197, 2019, https://doi.org/10.1007/978-981-13-6861-5_17.

[33] M. R. Machado, S. Karray, and I. T. de Sousa, "LightGBM: an Effective Decision Tree Gradient Boosting Method to Predict Customer Loyalty in the Finance Industry," 2019 14th International Conference on Computer Science & Education (ICCSE), Aug. 2019, https://doi.org/10.1109/iccse.2019.8845529.

[34] S. Gomathi and V. Narayani, "A proposed framework using CAC algorithm to predict systemic lupus erythematosus (SLE)," 2016 World Conference on Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave), Feb. 2016, https://doi.org/10.1109/startup.2016.7583974.

[35] B. Abdualgalil and S. Abraham, "Applications of Machine Learning Algorithms and Performance Comparison: A Review," 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), Feb. 2020, https://doi.org/10.1109/ic-etite47903.2020.490.

[36] T. B. Alakus and I. Turkoglu, "Comparison of deep learning approaches to predict COVID-19 infection," Chaos, Solitons & Fractals, vol. 140, p. 110120, Nov. 2020, https://doi.org/10.1016/j.chaos.2020.110120.

[37] F. Itoo, Meenakshi, and S. Singh, "Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection," International Journal of Information Technology, vol. 13, no. 4, pp. 1503–1511, Feb. 2020, https://doi.org/10.1007/s41870-020-00430-y.

[38] L. Akter, Ferdib-Al-Islam, M. M. Islam, M. S. Al-Rakhami, and M. R. Haque, "Prediction of Cervical Cancer from Behavior Risk Using Machine Learning Techniques," SN Computer Science, vol. 2, no. 3, Mar. 2021, https://doi.org/10.1007/s42979-021-00551-6.

[39] M. Bracher-Smith, K. Crawford, and V. Escott-Price, "Machine learning for genetic prediction of psychiatric disorders: a systematic review," Molecular Psychiatry, vol. 26, no. 1, pp. 70–79, Jun. 2020, https://doi.org/10.1038/s41380-020-0825-2.

[40] J. Xu, Y. Zhang, and D. Miao, "Three-way confusion matrix for classification: A measure driven view," Information Sciences, vol. 507, pp. 772–794, Jan. 2020, https://doi.org/10.1016/j.ins.2019.06.064.