Light Gradient Boosting with Hyper Parameter Tuning Optimization for COVID-19 Prediction

Ferda Ernawan[1]
Faculty of Computing
Universiti Malaysia Pahang
Pekan, Malaysia

Kartika Handayani[2]
Faculty of Engineering and Informatics
Universitas Bina Sarana Informatika
Jakarta, Indonesia

Mohammad Fakhreldin[3]
Faculty of Computer Science and Information Technology
Jazan University
Jazan, Saudi Arabia

Yagoub Abbker[4]
Faculty of Computer Science & Information Technology
Jazan University
Jazan, Saudi Arabia

***Abstract*—The 2019 coronavirus disease (COVID-19) caused pandemic and a huge number of deaths in the world. COVID-19 screening is needed to identify suspected positive COVID-19 or not and it can reduce the spread of COVID-19. The polymerase chain reaction (PCR) test for COVID-19 is a test that analyzes the respiratory specimen. The blood test also can be used to show people who have been infected with SARS-CoV-2. In addition, age parameters also contribute to the susceptibility of COVID-19 transmission. This paper presents the extra trees classification with random over-sampling by considering blood and age parameters for COVID-19 screening. This research proposes enhanced preprocessing data by using KNN Imputer to handle large missing values. The experiments evaluated the existing classification methods such as Random Forest, Extra Trees, Ada Boost, Gradient Boosting, and the proposed Light Gradient Boosting with hyperparameter tuning to measure the predictions of patients infected with SARS-CoV-2. The experiments used Albert Einstein Hospital test data in Brazil that consisted of 5,644 sample data from 559 patients with infected SARS-CoV-2. The experimental results show that the proposed scheme achieves an accuracy of about 98,58%, recall of 98,58%, the precision of 98,61%, F1-Score of 98,61%, and AUC of 0,9682.**

***Keywords*—ROS; light gradient boosting; hyper parameter tuning; COVID-19 screening; blood and age based**

I. Introduction

Coronavirus 19 (COVID-19) is a highly contagious viral infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [1]. SARS-CoV-2 can cause tissue damage and cause acute respiratory distress syndrome. It is rapidly increasing transmission rate which demands an early response to diagnose and prevent the rapid spread of this disease [2]. Currently, COVID-19 is being transmitted by human-to-human through air transmission that cause a wide spread of the disease [3]. One way to detect COVID-19 is through the Reverse-Transcriptase Polymerase Chain Reaction, also known as RT-PCR [4]. RT-PCR has limited resources, it has high specificity and high sensitivity [5]. However, according to the study of validation of the SARS-CoV-2 RT-PCR test [6], blood or hematological parameters showed high sensitivity and specificity as well as intra and inter-test precision and efficiency.

Machine learning can become an alternative for diagnosing and analyzing COVID-19 infection [7]. Machine Learning has been widely used to investigate and help in screening with suspected COVID-19 infection [8]. The implementation of machine learning in RT-PCR with blood assessments has a critical function for diagnosing COVID-19 and different respiration diseases. The parameters are involved white blood cells, C-reactive protein, neutrophils, lymphocytes, monocytes, eosinophils, basophils, aspartate and alanine, lactate dehydrogenase, and others. Those parameters have proven an excessive correlation in sufferers identified with COVID [9]. In addition, age parameters [10] also affect the susceptibility of COVID-19 transmission. Therefore, it motivates researchers to investigate parameters that significantly effect for covid-19 prediction.

This research presents a predictive model for diagnosing COVID-19 by considering C-reactive protein, neutrophils, lymphocytes, monocytes, eosinophils, basophils, aspartate and alanine, lactate dehydrogenase, including blood and age parameters. This research proposes a predictive model by using ensemble learning which involved Random Forest, Extra Trees, AdaBoost, Gradient Boosting and Light Gradient Boosting, then optimizes the best model with hyperparameter tuning. The experiments also investigate the best solution for imbalance data by implementing the existing sampling methods such as Random Under Sampling (RUS), Random Over Sampling (ROS) and Synthetic Minority Over Sampling TEchnique (SMOTE). The sampling class imbalance approaches is used to overcome imbalance data that has been carried out in the research related to Covid-19 [11]. This research is expected to obtain the best predictive model that can achieve high accuracy, recall, precision, f-score and AUC compared to the existing schemes.

II. Related Work

Several researches have proven the significant of blood exams for the diagnosis of Covid-19 [12] analyzing the blood

index of 69 COVID-19 sufferers. All have been dealt with on the National Center for Infectious Diseases (NCID) placed in Singapore. Among those sufferers, sixty-five underwent whole blood assume the day of admission. In addition, demographic facts inclusive of age, gender, ethnicity, and region have been furnished for this study. Around 13,4% of sufferers require in-depth care unit (ICU) care, specifically the elderly. During the primary examination, 19 sufferers had leukopenia (low white blood cells) and 24 had lymphopenia (low lymphocyte stage with inside the blood), with five instances categorized as severe (Absolute lymphocyte count (ALC).

The application of a Covid-19 diagnosis based on blood tests has previously been carried out to provide comprehensible answers primarily based totally on device studying techniques using public data from the Albert Einstein Hospital. Previously, data preprocessing was carried out for selection of blood features. Then normalization of features with z-score and use of iterative imputer method to fill in missing values is done. The remaining 608 patients, 84 of whom have been high-quality for COVID-19 showed with the aid of using RT-PCR [13]. In order to apprehend the decisions, a neighborhood Decision Tree Explainer (DTX) approach is performed to obtain the results.

Data from the Israel Albert Einstein Hospital located in São Paulo, Brazil are also used in the application of machine learning in the diagnosis of COVID-19 with hematological parameters. Pre-processing is done by selecting features using particle swarm optimization (PSO) and evolutionary search (ES). Furthermore, experiments were carried out with different machine learning techniques. The experimental results show that Bayesian networks [7] have superior performance compared to other techniques with an overall accuracy of 95,159%, kappa index 0,903, sensitivity 0,968, precision 0.938, and specificity 0,936.

A study was also conducted to identify SARS-CoV-2 positive patients from a total of 598 complete data and 5046 were not used because they were incomplete. A machine learning model, ANN was carried out to test based on the dataset obtained from the Israelta Albert Einstein Hospital, in São Paulo, Brazil by testing various hematological parameters. As a result, the flexible ANN model [14] predicts COVID-19 patients with high accuracy between the population in the regular ward AUC 94-95% and those not hospitalized or in the community AUC 80-86%.

Other research was conducted by building a two-stage test; in level one, no preprocessing technique is carried out even as in level preprocessing is emphasized to attain higher predictive effects. Blood samples from sufferers from Einstein Hospital in Brazil were amassed and used for prediction of the severity of COVID-19 with studying algorithms. The Tuned Random Forest algorithm [15] produced an accuracy of 0,98 with numerous preprocessing methods.

Based on the description of the related research above, the existing considers few parameters to diagnose COVID-19. There are a quite few research studies on blood exams for the diagnosis of COVID-19. However, studies on eosinophils, age and blood parameters are rare to find in literature. This study proposes a pre-processing KNN imputer data to overcome the large missing values. Then various data sampling class with imbalance approaches methods is used to find out the best sampling class for imbalance datasets. Whereas the prediction model generated from the data classification process using an ensemble, namely Extra Trees, Bagging Decision Tree, Random Forest, Ada Boost, Gradient Boosting and Light Gradient Boosting.

III. Preliminaries

A. Ensemble Learning Classification Model

Ensembles learning classification model can increase the computational costs [16], as it is necessary to train several individual classifiers, and their computational requirements can grow exponentially when dealing with large scales.

B. Extra Trees

The extra tree classifier creates a gaggle of unpruned decision trees in step with the standard top-down method. The predictions of all trees were combined to determine the ultimate prediction, through the majority alternative [17]. The extra tree classifier generates a random multiple of the choice tree with completely different sub-samples while not bootstrapping. The extra trees can avoid over-fitting issues and improves accuracy [18]. Efficiency is also the main strength of this study.

C. AdaBoost

AdaBoost is an iterative algorithm, in each iteration, instances that were wrongly classified in the previous iteration are given more weight. Sequentially apply the learning algorithm to reweighted the sample from the original training data. Initially, each instance is ssigned the same weight and iteration as the iteration, the weight of all misclassified instances is increased and the correctly classified instances are reduced [17]. The AdaBost algorithms [19] are defined by:

1) Minimize the error function with the formula

$$\mathrm{w_e} = \sum_{yi \neq km(xi)} w_i^{(m)} \exp(a_m) \tag{1}$$

2) Set the value a with the formula

$$a_m = \frac{1}{2} n\left(\frac{1-e_m}{e_m}\right), e_m = \frac{w_e}{w} \tag{2}$$

3) Update values if observing misclassification by formula

$$w_i^{(m+1)} = w_i^{(m)} \exp(a_m) = w_i^{(m)} \sqrt{\frac{1-e_m}{e_m}} \tag{3}$$

4) For other values using the formula

$$w_i^{(m+1)} = w_i^{(m)} \exp(a_m) = w_i^{(m)} \sqrt{\frac{e_m}{1-e_m}} \tag{4}$$

D. Gradient Boosting

Gradient Boosting is a machine learning algorithm that can solve regression and classification problems. Gradient Boosting generates a prediction model consisting of an ensemble of weak prediction models in the decision tree [20]. The construct of a gradient boosting call tree is to mix a series of weak base classifiers into one sturdy one. a conventional boosting methodology that weighs positive and negative samples, GBDT builds a world convergence rule by following the direction of the negative gradient [21]. The GBDT measures GBDT [21] are presented as follows.

1) Step 1: The values for the initial constants of the model β are given:

$$f0(x) = \arg\min \beta \sum_{i=1}^{n} L(yi, \beta) \tag{5}$$

2) Step 2: For the number of iterations m = 1: M (M is the iteration time), the residual gradient direction is calculated

$$y_i - \left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \right] f(x) - f_{m-1(x)} \tag{6}$$

3) Step 3: Base classifiers are used to adjust the sample data and obtain the initial model. According to the least squares approach, the parameters of the model are obtained and the model *h (xi; am*) is installed

$$a_m = \arg\min_{a\beta} \sum_{i=1}^{N} \left[y_i^* - \beta h(x_i; a) \right]^2 \tag{7}$$

4) Step 4: Function loss is minimized. According to Eq. (4), the new step size of the model, i.e. the weight of the current model, is calculated.

$$\beta_m = \arg\min_{a\beta} \sum_{i=1}^{N} L\left[y_i^*, f_{m-}(x) = +\beta h(x_i; a) \right] \tag{8}$$

5) Step 5: the model is updated as follows

$$F_m(x) = F_{m-1}(x) + \beta_m h(x_i; a) \tag{9}$$

E. Light Gradient Boosting

Light Gradient Boosting Machine or LightGBM uses gradient enhancement in its construction, but light GBM does not divide the eigenvalues one by one, so it is necessary to calculate the splitting benefit of each eigenvalue. LightGBM algorithm on the model to improve forecasting accuracy and robustness [22]. It can indeed find the optimal split value, but it costs a lot, and may not be good for generalizing information when the amount of data is large [23]. Remembering the supervised training $setX = \{(xi, yi)\}_i^n$ LightGBM's target is to find approximation for a particular function $f(x)$ to a certain function $\hat{f}(x)$ which reduces the expected loss function value , $(y, f(x))$ as follows [24] :

$$\hat{f} = \arg\min Ey, xL(y, f(x)) \tag{10}$$

LightGBM integrates a number of T regression trees to approach the final model, which is.

$$f_T(x) \sum_{t=2}^{T} f_t(x) \tag{11}$$

$q(x), q \in \{1,2...., J\}$, where J denotes the number of leaves, represents the guideline of thumb of the choice tree and is the leaf node weight vector. Therefore, LightGBM could be educated additively inside the following steps:

$$T_t = \sum_{i=1}^{n} L(yi, f_{t-1}(xi) + f_t(xi)) \tag{12}$$

In LightGBM, Newton' technique simply approximates the target function. Where *gi* and *hi* indicate the first- and second-order gradient statistics of the loss function, let *Ij* show the instance set of leaf*j*.

$$T_t = \sum_{i=1}^{n} \left(\left(\sum_{i \in Ij} g_i \right) + \frac{1}{2} \left(\sum_{i \in Ij} h_i + \lambda \right) w_j^2 \right) \tag{13}$$

For the tree structure$q(x)$, the optimum leaf weight score of every leaf node $w*$ and therefore the extreme worth of T_t may be solved as follows:

$$w_j^* = -\frac{\sum_{i \in Ijgi}}{\sum_{i \in Ij} (h_i + \lambda) w_j^2} \tag{14}$$

F. Random Forest

Random Forest is an integrated learning method based on bagging. The essence is to apply the bootstrap method to the CART algorithm. Random Forest samples were taken using the bootstrap method, and then an independent decision tree model was built using the CART algorithm [25]. Random forest algorithm (for each type and regression) [26] are discussed as follows:

1) From Training n samples draw n_{tree} bootstrap samples.

2) For every of the bootstrap samples, develop classification or regression tree with the subsequent modification: at every node, in place of selecting the excellent break up amongst all predictors, randomly pattern m_{trv} of the predictors and select the excellent break up amongst the ones variables. The tree is grown to the most length and not pruned back. Bagging may be concept of because the unique case of random forests received while $m_{try} = p$, the wide variety of predictors.

3) Predict new facts by combining n_{tree} tree predictions (i.e., majority vote for type, common for regression.

G. Random Over Sampling (ROS)

ROS algorithm randomly replicates samples from the minority classes [27]. Oversampling [28] can be done by

increasing number of instances or minority class samples by production new instance or repeated multiple instances.

H. Random Under Sampling (RUS)

RUS technique at random eliminates samples from the bulk categories, till achieving a relative categories balance [27]. For the under-sampling approach, most of the category instances are discarded till additional a balanced distribution of information is achieved. This data merchandising method is completed every which way. Considering an information set with a hundred minority class instances and 2,000 magnitude class instances, a complete of 1800 categories that majority are going to be deleted randomly within the RUS technique. The dataset will be balanced with two hundred instances, it will be delineating with 200 instances, whereas minority also have 200.

I. Synthetic Minority Over-Sampling Technique (SMOTE)

SMOTE produces artificial samples from the minority class by interpolating existing instances that are terribly near to every other [27]. For the minority category within the information set, SMOTE initial selects the minority class data instance randomly. The distance from the sample set to several classes is calculated by the Euclidean distance *D*, and K-nearest neighbors are obtained. The Euclidean distance *D* is defined by:

$$D = \sqrt{\sum_{k-1}^{n} (X_k - Y_k)^2} \tag{15}$$

According to the proportion of the unbalanced data set, the sampling rate N is set. The six samples closest to D were selected as one group. Each sample group is connected to each other to generate several new samples at random, which are added to the data set and recycled [29]. This results in a new formula:

$$X_{new} = x_i + rand(0,1) * |x_i - x_j| \tag{16}$$

IV. Experimental Setup

Images are divided by 70% for training, 20% for validation, and 10% for testing. Then the YOLO architectural model is used from training and validation and then a data test is carried out with data testing and detecting disease. After that, a performance evaluation's carried out for the architectural model used. The block diagram of the proposed covid-19 classification is depicted in Fig. 1.

This study uses machine learning techniques to predict negative and positive cases using RT-PCR data with blood parameters. Before applying the machine learning classification method, data preparation was carried out by using several methods, namely, Remove non-blood parameter, Imputation Missing Values, Label Encoding Class and Normalization with Z-Score. The processed data was tested using several machine learning classification methods using an ensemble, namely Extra Trees, Bagging Decision Tree, Random Forest, Ada Boost, Gradient Boosting and Light Gradient Boosting. In testing the machine learning classification method, the best method was chosen based on the evaluation of the results in terms of accuracy, precision, recall, F-1score and AUC. The best method is optimized by searching for the best parameters by using hyper parameter tuning. Then, the results were compared before using hyper parameter tuning and after using hyper parameter tuning. The results of the best methods can be used for prediction of COVID-19.

A. Data Collection

The dataset is collected from the existing benchmark [30]. The dataset consists of 5644 patients treated at the Albert Einstein Israelta Hospital located in Saulo Paulo, Brazil. Kaggle makes data sets available for public access. Data was collected from 28 March 2020 to 3 April 2020, with more than 100 laboratory tests including blood test, urine test, SARS-CoV-2 test, RT-PCR test, presence of influenza virus [30]. The dataset consists of 89% missing values, so the missing value is handled by filling in the missing value using the KNN Imputer method using K = 5 [31]. Label encoding is done which aims to perform coding on the class label. Label Encoding serves to change the data format of numbers 0 to n_classes-1, this is intended to make data training easier. Normalization of the data was performed using Z-Score [32]. Then the best method is to optimize hyper parameter tuning using GridSearchCV. GridSearchCV taken from Scikit learn [33]. This study considers several features for classification as shown in Table I.

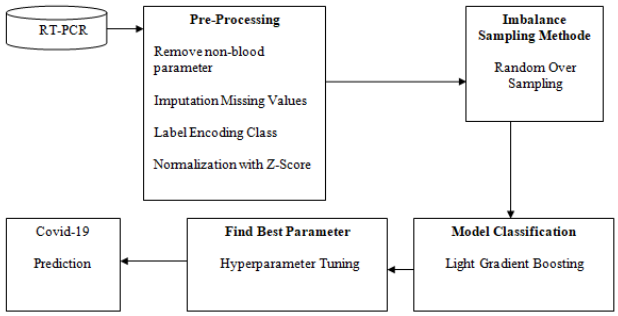

Fig. 1. Block Diagram of Covid-19 Prediction.

TABLE I. Selection of Features

No.	Features	No.	Features
1	Hematocrit	13	Red blood cell distribution width
2	Hemoglobin	14	Monocytes
3	Platelets	15	Mean platelet volume
4	Red blood Cells	16	Neutrophils
5	Lymphocytes	17	C-reactive protein
6	Mean corpuscular hemoglobin	18	Creatinine
7	MCH concentration	19	Urea
8	Leukocytes	20	Potassium
9	Basophils	21	Sodium
10	Eosinophils	22	Aspartate transaminase
11	Lactate dehydrogenase	23	Alanine transaminase
12	Mean corpuscular volume	24	Age

B. Split Validation

In this study, the experiments divide the data based on the ratio entered, for example the percentage of 80:20 [34]. There are 80% of the total amount for training set and 20% for test set.

C. Evaluation

To compare the overall performance of the proposed scheme, we decided on five metrics: accuracy, recall, precision, F1-Score and receiver running characteristic (ROC) curves, and the cost of the vicinity below the ROC curve (ROC AUC). Accuracy is the maximum generally used assessment metric for type. However, for imbalance facts type problems, accuracy won't be a great preference due to the fact accuracy regularly has a bias closer to the bulk class [35][36]. The accuracy can be defined by:

$$Accuracy = \frac{TP + TN}{TotalSample} \tag{17}$$

Recall is the collection of data that has been successfully taken from the part of the data relevant to the query [37]. The Recall is defined by:

$$\mathrm{Re}call = \frac{TP}{TP + FN} \tag{18}$$

Precision is part of the data taken in accordance with the required information [38]-[40]. The precision is defined by:

$$\mathrm{Pr}ecision = \frac{TP}{TP + FP} \tag{19}$$

The F1 score is the Harmonic Mean between precision and Recall [41]. The F-Score indicates how precise the classifier is (how many instances are correctly classified), as well as how strong it is (it doesn't miss a large number of instances). The F1-Score formula is defined by:

$$F1 - Score = 2\frac{\mathrm{Re}call * \mathrm{Pr}ecision}{\mathrm{Re}call + \mathrm{Pr}ecision} \tag{20}$$

The ROC curve represents the genuine advantageous rate (TPR) and fake advantageous rate (FPR). TPR represents the ratio of advantageous samples that have been successfully detected through the algorithm, and FPR represents the ratio of terrible samples that have been incorrectly labeled as advantageous. The expressions for TPR and FPR are as follows:

$$TPR = \frac{TP}{TP + FN} \tag{21}$$

$$FPR = \frac{FP}{TN + FP} \tag{22}$$

where TP is the number of true positives, TN is the number of true negatives, FN is the number of false negatives, FP is number of false positives.

V. Experimental Results

After pre-processing the data to overcome the missing value, performing a Z-Score then encoding the dataset class, testing the specified model without using the sampling class imbalance approaches method. Testing the model without sampling class imbalance approaches method is carried out first for further comparison with various sampling class imbalance approaches methods to be tested. The test results are listed in Table II.

The best accuracy was obtained by using extra trees method with an average accuracy of 98.40% for imbalance sampling method. While, the light gradient boosting achieved high accuracy with random under sampling than extra trees, AdaBoost, Gradient Boosting, and Random Forest methods. Overall, the extra trees method performs better than other method for different types of sampling method except random under sampling. The experimental results in terms of recall, precision, F1-Score, and AUC are listed in Tables III, IV, V and VI.

The classification of light gradient boosting method achieved recall value of 91.96%. The best recall result was obtained from sampling technique of without imbalance sampling method, random under sampling, SMOTE and SMOTE-Tomek. The experiments also evaluate the precision of the classification method; classification by using extra tree produced a high precision result except sampling technique of random under sampling. The classification of light gradient boosting method can achieve a good F1-Score and AUC score under various sampling techniques. The visual comparison of the accuracy, recall, precision, F1-score and AUC is shown in Fig. 2, 3, 4, 5 and 6.

TABLE II. Selection of Accuracy Result for 5644 RT-PCR Data

Sampling Technique	Extra Trees	Light Gradient Boosting	AdaBoost	Gradient Boosting	Random Forest
Without imbalance sampling method	**98,4**	98,22	96,89	97,69	98,22
Random Under Sampling	96,27	**96,63**	95,3	96,27	96,63
Random Over Sampling	**98,4**	98,22	96,89	97,69	98,22
SMOTE	**98,4**	98,22	97,34	97,6	97,96
SMOTE- Tomek	**98,49**	98,22	97,34	97,69	98,05

TABLE III. SELECTION OF RECALL RESULT FOR 5644 RT-PCR DATA

Sampling Technique	Extra Trees	Light Gradient Boosting	AdaBoost	Gradient Boosting	Random Forest
Without imbalance sampling method	88,39	**91,96**	86,60	90,17	89,28
Random Under Sampling	96,42	**97,32**	94,64	93,75	96,42
Random Over Sampling	88,39	91,96	**94,64**	93,75	90,17
SMOTE	90,17	**91,96**	90,17	89,28	88,39
SMOTE- Tomek	90,17	**91,96**	90,17	89,28	88,39

TABLE IV. SELECTION OF PRECISION RESULT FOR 5644 RT-PCR DATA

Sampling Technique	Extra Trees	Light Gradient Boosting	AdaBoost	Gradient Boosting	Random Forest
Without imbalance sampling method	**95,19**	94,49	90,65	92,66	93,45
Random Under Sampling	72,97	75,69	69,28	75	**76,05**
Random Over Sampling	**95,19**	90,35	78,51	84,67	91,81
SMOTE	**91,81**	91,15	87,06	88,49	89,59
SMOTE- Tomek	**94,39**	90,35	84,16	87,71	91,66

TABLE V. SELECTION OF F1-SCORE RESULT FOR 5644 RT-PCR DATA

Sampling Technique	Extra Trees	Light Gradient Boosting	AdaBoost	Gradient Boosting	Random Forest
Without imbalance sampling method	91,66	**93,21**	88,58	91,4	91,32
Random Under Sampling	83,07	**85,15**	80	83,33	85,03
Random Over Sampling	**91,66**	91,15	85,82	88,98	90,99
SMOTE	**93,51**	90,35	84,16	87,71	90,82
SMOTE- Tomek	90,41	**91,15**	87,06	88,49	89,99

TABLE VI. SELECTION OF AUC RESULT FOR 5644 RT-PCR DATA

Sampling Technique	Extra Trees	Light Gradient Boosting	AdaBoost	Gradient Boosting	Random Forest
Without imbalance sampling method	0,9395	**0,9568**	0,9281	0,9469	0,9429
Random Under Sampling	0,9634	**0,9693**	0,9501	0,9515	0,9654
Random Over Sampling	0,9395	0,9544	0,9589	**0,9594**	0,9464
SMOTE	0,9474	**0,9544**	0,9415	0,9395	0,937
SMOTE- *Tomek*	0,9479	**0,9544**	0,9415	0,9395	0,9375

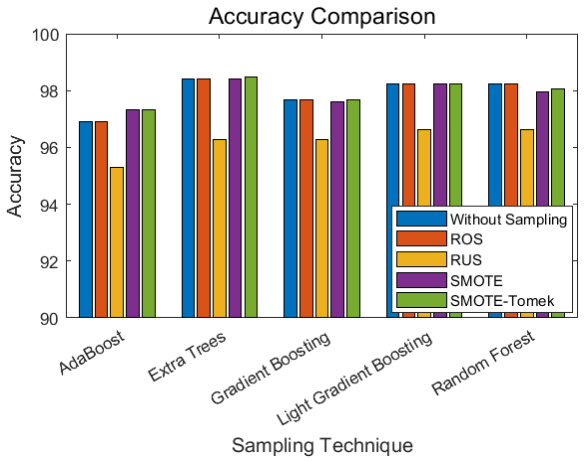

Fig. 2. The Accuracy of the Existing Classification Methods for Covid-19 Prediction.

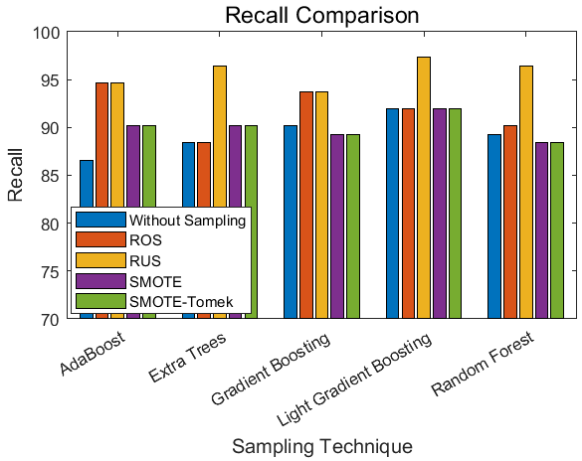

Fig. 3. The Recall of the Existing Classification Methods for Covid-19 Prediction.

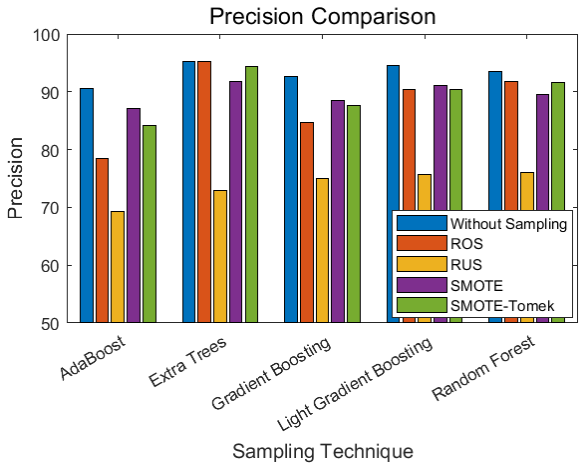

Fig. 4. The Precision of the Existing Classification Methods for Covid-19 Prediction.

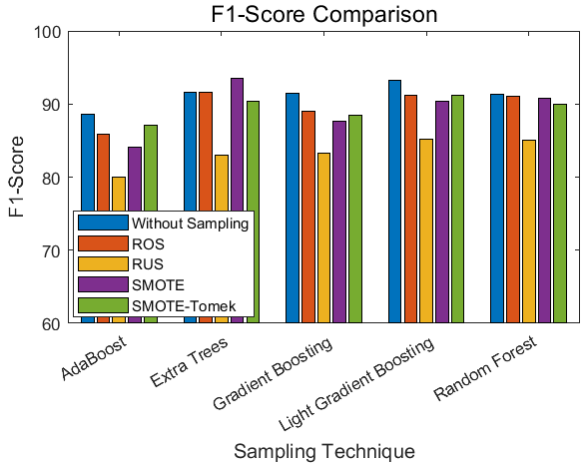

Fig. 5. The F1-score of the Existing Classification Methods for Covid-19 Prediction.

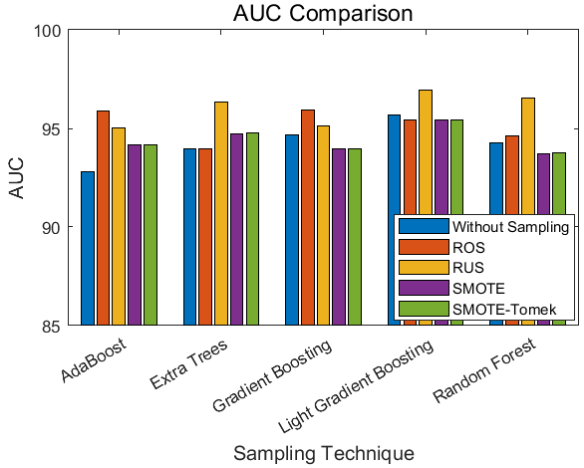

Fig. 6. The AUC of the Existing Classification Methods for Covid-19 Prediction.

The best AUC was produced by light gradient boosting with RUS sampling technique. Light gradient boosting with RUS sampling technique produces AUC score of 0.9693. It can be concluded that the best model that has improved majority of performance in terms of accuracy, precision, recall, f1-score and AUC is light gradient boosting. Light Gradient Boosting produces the best accuracy of 98.49%, recall on the RUS sampling technique is 97.32% and AUC is 0.9693. Furthermore, hyperparamater tuning tests were carried out to optimize the results of Light Gradient Boosting. The parameters used in the Hyperparameter tuning are listed in Table VII.

TABLE VII. SELECTION OF LIGHT GRADIENT BOOSTING PARAMETER ON HYPERPARAMETER TUNING GRID SEARCH

Parameters	**Value**
n_estimators	100, 400, 10
min_child_weight	3, 20, 2
colsample_bytree	0.4, 1.0
max_deph	5, 15, 2
num_leaves	8, 40
min_child_weight	10,30
Learning_rate	0.01,1

After going through the Grid Search process, the best parameters were found that could be tested on the Light Gradient Boosting model. These parameters can be seen in Table VIII.

TABLE VIII. SELECTION OF LIGHT GRADIENT BOOSTING PARAMETERS

Parameters	**Value**	**Parameters**	**Value**
boosting_type	'gbdt',	*n_jobs*	-1,
class_weight	None,	*num_leaves*	40,
colsample_bytree	0.4,	*objective*	None,
importance_type	'split',	*random_state*	None,
learning_rate	0.01,	*reg_alpha*	0.0,
'max_depth	15,	*reg_lambda*	0.0,
min_child_samples	10,	*silent*	True,
min_child_weight	3,	*subsample*	1.0,
min_split_gain	0.0,	*subsample_for_bin*	200000,
n_estimators	400,	*subsample_freq*	0

Table IX is a comparison of light gradient boosting before optimization of hyper parameter tuning and after optimization of hyper parameter tuning.

TABLE IX. SELECTION OF COMPARISON OF LIGHT GRADIENT BOOSTING

Evaluation	***without sampling***	***ROS***	***RUS***	**SMOTE**	**SMOTE-*Tomek***
Accuracy	97.78	98.58	97.25	98.31	98.31
Recall	97.78	98.58	97.25	98.31	98.31
Precision	97.83	98.61	97.65	98.34	98.34
F1-Score	97.83	98.61	97.65	98.34	98.34
AUC	95.68	96.82	96.88	95.88	95.88

The hyper parameter tuning has increased the accuracy of light gradient boosting with an accuracy of 98.58%. The comparison of recall light gradient boosting has increased in almost all tests using sampling techniques. Random forest before the sampling technique was 92.59%. The comparison of F1-score light gradient boosting after hyperparamerer tuning achieved 98.61% on the ROS sampling technique. Based on the results, it can be concluded that light gradient boosting with hyperameter tuning can improve the accuracy, recall, precision, F1-score and AUC. The use of the ROS sampling technique has some advantages in terms of accuracy, recall, precision, f1-score. With the conclusion that the results are 98.58% accuracy, 98.58% recall, 98.61% precision, f1-Score 98.61% and AUC 0.9682%. Based on the results obtained, the results of feature importance are shown in Fig. 7.

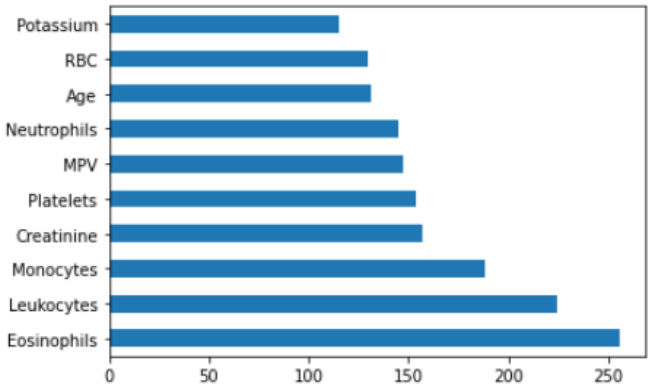

Fig. 7. Importance Features.

Based on Fig. 7, it shows that the first order important features in eosinophiles are. Followed by leukocytes, monocytes, creatinine, platelets, MPV, neutrophils, age, RBC and potassium. The addition of age in the proposed test becomes the seventh most important feature of the best model. The comparison with related research was conducted to assess the performance of the proposed research, the comparison results is listed in Table X.

TABLE X. SELECTION OF COMPARISON WITH RELATED RESEARCH

Model	Accuracy	Recall	Precision	F1 Score	AUC
proposed method	98,58	98,58	98,61	98,61	96,82
Alves *et al* [13]	88	66	91	76	86
Barbosa *et al* [7]	95.16	96.80	93.80	-	-

Based on the table above, it shows that the proposed model produces the best results for all evaluation matrices compare to the previous related studies. With the results of accuracy 98.58%, recall 98.58%, precision 98.61%, F1-Score 98.61% and AUC 0.9682. The visual comparison with related research studies is shown in Fig. 8.

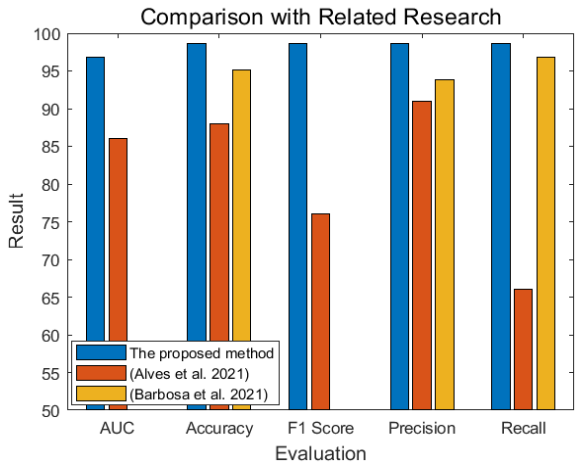

Fig. 8. Comparison with the Existing Studies.

VI. CONCLUSION

This paper has presented various classification methods for COVID-19 prediction. The classification method of light gradient boosting with hyper parameter tuning using ROS sampling technique perform better than the existing the classification methods such as extra trees, random forest, adaboost and gradient boosting for predicting the COVID-19 data. Eosinophils, blood and age parameters has potential become important parameters for COVID-19 prediction. The data was taken from kaggle.com with 5644 data, it shows a classification improvement based on the majority of performance in terms of recall, precision, f1-score and AUC score due to eosinophils, blood and age parameters. Hyper parameter tuning using ROS sampling technique achieved an accuracy of 98.58%, recall of 98.58%, precision of 98.61%, f1-score of 98.61% and AUC of 0.9682. The first important feature in these experiments is eosinophils; it can significantly influence the classification results, while age feature is in the seventh order of important features. In the future research, the proposed model has potential to predict monkey pox disease by identifying important features.

ACKNOWLEDGMENT

This work was supported by Universiti Malaysia Pahang, Universitas Bina Sarana Informatika and Jazan University.

CONFLICT OF INTEREST

On behalf of all authors, the corresponding author states that there is no conflict of interest.

REFERENCES

[1] M. A. Shereen, S. Khan, A. Kazmi, N. Bashir, and R. Siddique, "COVID-19 infection: Origin, transmission, and characteristics of human coronaviruses," J. Adv. Res., vol. 24, pp. 91–98, 2020, doi: 10.1016/j.jare.2020.03.005.

[2] H. A. Rothan and S. N. Byrareddy, "The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak," J. Autoimmun., vol. 109, no. February, p. 102433, 2020, doi: 10.1016/j.jaut.2020.102433.

[3] Q. Li et al., "Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia," N. Engl. J. Med., vol. 382, no. 13, pp. 1199–1207, 2020, doi: 10.1056/nejmoa2001316.

[4] Z. Zhang et al., "Insight into the practical performance of RT-PCR testing for SARS-CoV-2 using serological data: a cohort study," The Lancet Microbe, vol. 2, no. 2, pp. e79–e87, 2021, doi: 10.1016/S2666-5247(20)30200-7.

[5] T. Ai et al., "Correlation of Chest CT and RT-PCR Testing for Coronavirus Disease 2019 (COVID-19) in China: A Report of 1014 Cases," Radiology, vol. 296, no. 2, pp. E32–E40, 2020, doi: 10.1148/radiol.2020200642.

[6] E. F. Strasser et al., "Validation of a SARS-CoV-2 RNA RT-PCR assay for high-throughput testing in blood of COVID-19 convalescent plasma donors and patients," Transfusion, vol. 61, no. 2, pp. 368–374, 2021, doi: 10.1111/trf.16178.

[7] V. A. de F. Barbosa et al., "Heg.IA: an intelligent system to support diagnosis of Covid-19 based on blood tests," Res. Biomed. Eng., no. December 2019, 2021, doi: 10.1007/s42600-020-00112-5.

[8] A. Imran et al., "AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app," Informatics Med. Unlocked, vol. 20, p. 100378, 2020, doi: 10.1016/j.imu.2020.100378.

[9] D. Ferrari, A. Motta, M. Strollo, G. Banfi, and M. Locatelli, "Routine blood tests as a potential diagnostic tool for COVID-19," Clin. Chem. Lab. Med., vol. 58, no. 7, pp. 1095–1099, 2020, doi: 10.1515/cclm-2020-0398.

[10] C. M. Goldstein E, Lipsitch M, "On the effect of age on the transmission of SARS-CoV-2 in households, schools and the community," J. Infect. Dis., 2020, doi: https://doi.org/10.1101/2020.07.19.20157362.

[11] M. Dorn et al., "Comparison of machine learning techniques to handle imbalanced COVID-19 CBC datasets," PeerJ Comput. Sci., vol. 7, p. e670, 2021, doi: 10.7717/peerj-cs.670.

[12] B. E. Fan et al., "Hematologic parameters in patients with COVID-19 infection," Am. J. Hematol., vol. 95, no. 6, pp. E131–E134, 2020, doi: 10.1002/ajh.25774.

[13] M. A. Alves et al., "Explaining machine learning based diagnosis of COVID-19 from routine blood tests with decision trees and criteria graphs," Comput. Biol. Med., vol. 132, no. March, 2021, doi: 10.1016/j.compbiomed.2021.104335.

[14] A. Banerjee et al., "Use of Machine Learning and Artificial Intelligence to predict SARS-CoV-2 infection from Full Blood Counts in a population," Int. Immunopharmacol., vol. 86, no. June, p. 106705, 2020, doi: 10.1016/j.intimp.2020.106705.

[15] E. C. Gök and M. O. Olgun, "SMOTE-NC and gradient boosting imputation based random forest classifier for predicting severity level of covid-19 patients with blood samples," Neural Comput. Appl., vol. 0123456789, 2021, doi: 10.1007/s00521-021-06189-y.

[16] J. Large, J. Lines, and A. Bagnall, "A probabilistic classifier ensemble weighting scheme based on cross-validated accuracy estimates," Data Min. Knowl. Discov., vol. 33, no. 6, pp. 1674–1709, 2019, doi: 10.1007/s10618-019-00638-y.

[17] E. K. Ampomah, Z. Qin, and G. Nyame, "Evaluation of tree-based ensemble machine learning models in predicting stock price direction of movement," Inf., vol. 11, no. 6, 2020, doi: 10.3390/info11060332.

[18] A. Zafari, R. Zurita-Milla, and E. Izquierdo-Verdiguier, "Land Cover Classification Using Extremely Randomized Trees: A Kernel Perspective," IEEE Geosci. Remote Sens. Lett., vol. 17, no. 10, pp. 1702–1706, 2020, doi: 10.1109/LGRS.2019.2953778.

[19] K. Nugroho et al., "Improving random forest method to detect hatespeech and offensive word," 2019 Int. Conf. Inf. Commun. Technol. ICOIACT 2019, pp. 514–518, 2019, doi: 10.1109/ICOIACT46704.2019.8938451.

[20] S. B. Koduri, L. Gunisetti, C. R. Ramesh, K. Mutyalu, and D. Ganesh, "Prediction of crop production using adaboost regression method Prediction of crop production using adaboost regression method," J. Phys. Conf. Ser., 2019, doi: 10.1088/1742-6596/1228/1/012005.

[21] H. Rao et al., "Feature selection based on artificial bee colony and gradient boosting decision tree," Appl. Soft Comput. J., 2019, doi: 10.1016/j.asoc.2018.10.036.

[22] Y. Ju, G. Sun, Q. Chen, M. Zhang, H. Zhu, and M. U. Rehman, "A model combining convolutional neural network and lightgbm algorithm for ultra-short-term wind power forecasting," IEEE Access, vol. 7, no. c, pp. 28309–28318, 2019, doi: 10.1109/ACCESS.2019.2901920.

[23] Y. Su, "Prediction of air quality based on Gradient Boosting Machine Method," Proc. - 2020 Int. Conf. Big Data Informatiz. Educ. ICBDIE 2020, pp. 395–397, 2020, doi: 10.1109/ICBDIE50010.2020.00099.

[24] H. Khafajeh, "An efficient intrusion detection approach using light gradient boosting," J. Theor. Appl. Inf. Technol., vol. 98, no. 5, pp. 825–835, 2020.

[25] Y. Guo, Y. Zhou, X. Hu, and W. Cheng, "Research on recommendation of insurance products based on random forest," Proc. - 2019 Int. Conf. Mach. Learn. Big Data Bus. Intell. MLBDBI 2019, pp. 308–311, 2019, doi: 10.1109/MLBDBI48998.2019.00069.

[26] M. A. M. Hasan, M. Nasser, S. Ahmad, and K. I. Molla, "Feature Selection for Intrusion Detection Using Random Forest," J. Inf. Secur., vol. 07, no. 03, pp. 129–140, 2016, doi: 10.4236/jis.2016.73009.

[27] E. Rendón, R. Alejo, C. Castorena, F. J. Isidro-Ortega, and E. E. Granda-Gutiérrez, "Data sampling methods to dealwith the big data multi-class imbalance problem," Appl. Sci., vol. 10, no. 4, 2020, doi: 10.3390/app10041276.

[28] R. Mohammed, J. Rawashdeh, and M. Abdullah, "Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results," 2020 11th Int. Conf. Inf. Commun. Syst. ICICS 2020, no. May, pp. 243–248, 2020, doi: 10.1109/ICICS49469.2020.239556.

[29] X. Li and Q. Zhou, "Research on improving SMOTE algorithms for unbalanced data set classification," Proc. - 2019 Int. Conf. Electron. Eng. Informatics, EEI 2019, pp. 476–480, 2019, doi: 10.1109/EEI48997.2019.00109.

[30] "No Hospital Israelita Albert Einstein, Diagnosis of Covid-19 and its Clinical Spectrum - Ai and Data Science Supporting Clinical Decisions (From 28th Mar to 3st Apr).," Accessed on 15/09/2021., [Online]. Available: https://www.kaggle.com/einsteindata4u/covid19?select=dataset.xlsx (Ac.

[31] M. AlJame, I. Ahmad, A. Imtiaz, and A. Mohammed, "Ensemble learning model for diagnosing COVID-19 from routine blood tests," Informatics Med. Unlocked, vol. 21, p. 100449, 2020, doi: 10.1016/j.imu.2020.100449.

[32] H. A. Prihanditya, "The Implementation of Z-Score Normalization and Boosting Techniques to Increase Accuracy of C4.5 Algorithm in Diagnosing Chronic Kidney Disease," vol. 5, no. 1, pp. 63–69, 2020.

[33] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.

[34] I. H. Witten and E. Frank, Practical Machine Learning Tools and Techniques, Second. San Francisco: Diane Cerra, 2005.

[35] G. Haixiang, L. Yijing, L. Yanan, L. Xiao, and L. Jinling, "BPSO-Adaboost-KNN ensemble learning algorithm for multi-class imbalanced data classification," Eng. Appl. Artif. Intell., vol. 49, pp. 176–193, 2016, doi: 10.1016/j.engappai.2015.09.011.

[36] M. R. Camana Acosta, S. Ahmed, C. E. Garcia, and I. Koo, "Extremely randomized trees-based scheme for stealthy cyber-attack detection in smart grid networks," IEEE Access, vol. 8, no. Ml, pp. 19921–19933, 2020, doi: 10.1109/ACCESS.2020.2968934.

[37] L. J. Halawa, A. Wibowo, and F. Ernawan, "Face Recognition Using Faster R-CNN with Inception-V2 Architecture for CCTV Camera," ICICOS 2019 - 3rd Int. Conf. Informatics Comput. Sci. Accel. Informatics Comput. Res. Smarter Soc. Era Ind. 4.0, Proc., Oct. 2019, doi: 10.1109/ICICOS48119.2019.8982383.

[38] M. S. Bin Othman Mustafa, M. Nomani Kabir, F. Ernawan, and W. Jing, "An Enhanced Model for Increasing Awareness of Vocational Students Against Phishing Attacks," 2019 IEEE Int. Conf. Autom. Control Intell. Syst. I2CACIS 2019 - Proc., pp. 10–14, Jun. 2019, doi: 10.1109/I2CACIS.2019.8825070.

[39] I. Khandokar, M. Hasan, F. Ernawan, S. Islam, and M. N. Kabir, "Handwritten character recognition using convolutional neural network," J. Phys. Conf. Ser., vol. 1918, no. 4, Jun. 2021, doi: 10.1088/1742-6596/1918/4/042152.

[40] M.A.T. Mohammed, A.S. Sadiq, R.A. Arshah, F. Ernawan, and S. Mirjalili, “Soft set decision/forecasting system based on hybrid parameter reduction algorithm,” Journal of Telecommunication, Electronic and Computer Engineering, vol. 9, no. 2-7, pp. 143-148, 2017.

[41] A. P. Puspaningrum et al., "Waste Classification Using Support Vector Machine with SIFT-PCA Feature Extraction," 2020 4th International Conference on Informatics and Computational Sciences (ICICoS), 2020, pp. 1-6, doi: 10.1109/ICICoS51170.2020.9298982.