



Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/combiomed

QCovSML: A reliable COVID-19 detection system using CBC biomarkers by a stacking machine learning model

Tawsifur Rahman^a, Amith Khandakar^a, Farhan Fuad Abir^b, Md Ahasan Atick Faisal^b,
Md Shafayet Hossain^c, Kanchon Kanti Podder^d, Tariq O. Abbas^e, Mohammed Fasihul Alam^f,
Saad Bin Kashem^g, Mohammad Tariqul Islam^c, Susu M. Zughailer^h, Muhammad E.
H. Chowdhury^{a,*}

^a Department of Electrical Engineering, Qatar University, Doha, 2713, Qatar

^b Department of Electrical and Electronics Engineering, University of Dhaka, Dhaka, 1000, Bangladesh

^c Dept. of Electrical, Electronics and Systems Engineering, Universiti Kebangsaan Malaysia, Bangi, Selangor, 43600, Malaysia

^d Department of Biomedical Physics & Technology, University of Dhaka, Dhaka, 1000, Bangladesh

^e Urology Division, Surgery Department, Sidra Medicine, Doha, 26999, Qatar

^f Department of Public Health, College of Health Sciences, QU Health, Qatar University, Doha, 2713, Qatar

^g Department of Computing Science, AFG College with the University of Aberdeen, Doha, Qatar

^h Department of Basic Medical Sciences, College of Medicine, QU Health, Qatar University, Doha, 2713, Qatar

ARTICLE INFO

Keywords:

COVID-19
Detection
Complete blood count (CBC)
Stacking machine learning
RT-PCR

ABSTRACT

The reverse transcription-polymerase chain reaction (RT-PCR) test is considered the current gold standard for the detection of coronavirus disease (COVID-19), although it suffers from some shortcomings, namely comparatively longer turnaround time, higher false-negative rates around 20–25%, and higher cost equipment. Therefore, finding an efficient, robust, accurate, and widely available, and accessible alternative to RT-PCR for COVID-19 diagnosis is a matter of utmost importance. This study proposes a complete blood count (CBC) biomarkers-based COVID-19 detection system using a stacking machine learning (SML) model, which could be a fast and less expensive alternative. This study used seven different publicly available datasets, where the largest one consisting of fifteen CBC biomarkers collected from 1624 patients (52% COVID-19 positive) admitted at San Raphael Hospital, Italy from February to May 2020 was used to train and validate the proposed model. White blood cell count, monocytes (%), lymphocyte (%), and age parameters collected from the patients during hospital admission were found to be important biomarkers for COVID-19 disease prediction using five different feature selection techniques. Our stacking model produced the best performance with weighted precision, sensitivity, specificity, overall accuracy, and F1-score of 91.44%, 91.44%, 91.44%, 91.45%, and 91.45%, respectively. The stacking machine learning model improved the performance in comparison to other state-of-the-art machine learning classifiers. Finally, a nomogram-based scoring system (QCovSML) was constructed using this stacking approach to predict the COVID-19 patients. The cut-off value of the QCovSML system for classifying COVID-19 and Non-COVID patients was 4.8. Six datasets from three different countries were used to externally validate the proposed model to evaluate its generalizability and robustness. The nomogram demonstrated good calibration and discrimination with the area under the curve (AUC) of 0.961 for the internal cohort and average AUC of 0.967 for all external validation cohort, respectively. The external validation shows an average weighted precision, sensitivity, F1-score, specificity, and overall accuracy of 92.02%, 95.59%, 93.73%, 90.54%, and 93.34%, respectively.

1. Introduction

After only one year and a half years' duration of the pandemic, the

SARS-CoV2 coronavirus, the pathogen responsible for COVID-19, has caused about 360 million infections with over five million casualties around the globe [1]. It is the worst pandemic that afflicted humanity

* Corresponding author.

E-mail address: mchowdhury@qu.edu.qa (M.E.H. Chowdhury).

<https://doi.org/10.1016/j.combiomed.2022.105284>

Received 21 December 2021; Received in revised form 26 January 2022; Accepted 2 February 2022

Available online 12 February 2022

0010-4825/© 2022 Published by Elsevier Ltd.

since the Spanish flu of 1918, which overwhelmed the global healthcare systems and caused a worldwide economic crisis. Unlike other coronaviruses (e.g., SARS, MERS), SARS-CoV-2 can infect an individual without causing any symptoms or very mild and non-characteristic ones for a prolonged duration. This increases the risk of spreading the disease and early identification of patients becomes crucial [2]. To detect SARS-CoV-2 infections, the current gold standard tool is the reverse transcription-polymerase chain reaction (RT-PCR). However, it requires specialized equipment and reagents (at least 4–5 h under optimal conditions), trained personnel for the samples collection, and proper genetic conservation of the Ribonucleic acid (RNA) sequences used for annealing the primers [3].

The data science community has proposed several machine learning (ML) based methods to assist the diagnosis and contain the transmission of COVID-19 [4]. In several studies, a number of biomarkers have been found to identify COVID-19 patients at the risk of severe infection and death by providing insight into their underlying health conditions. Zheng et al. [5] found that neutrophil, platelet, and white blood cell counts were typically normal in the COVID-19 patients at the time of admission (87.9%, 85.1%, and 88.7%, respectively). However, the majority of the patients developed lymphopenia right after the onset of symptoms, and it became more pronounced with the progression of the disease. They also showed that the neutrophil, lymphocyte, and platelet (NLP) counts are clinically useful in stratifying patients. However, the dataset was small, and they did not test the model on any external dataset. Moreover, Wang et al. [6] used 301 adult patient data to develop a new prediction score called ANDC for evaluating the mortality risk of COVID-19 patients. LASSO regression was used to identify age, D-dimer, neutrophil-to-lymphocyte ratio (NLR), and C-reactive protein as death predictors for COVID-19 patients [6]. Although several clinical studies [5,7] have shown blood test-based diagnosis as an effective and low-cost solution for the early COVID-19 detection, only a few ML models have been applied to hematological parameters [4,6,8–11]. Weng et al. [12] developed a nomogram using a logistic regression classifier on a single dataset collected from a hospital in Wuhan. Ramachandran et al. [8] used multivariable logistic regression analyses on a single dataset collected from a tertiary care academic medical center in New York City and found the deviations in the Red Blood Cell distribution width (RDW) among the hospitalized COVID-19 patients. A similar study was done by Foy et al. [10] using a dataset collected from Massachusetts General Hospital in Boston, Massachusetts, USA. Gong et al. [9] developed a nomogram using logistic regression on a dataset collected from Wuhan. A different study has developed several machine learning (ML) models to improve diagnostic capabilities to prevent the spread of the pandemic [4]. The majority of these models [4,13–16] employed computed tomography (CT) scans and chest X-rays. However, the CT scan-based diagnostic approaches showed higher false-negative results that caused concern for real-life applications [17]. Moreover, CT imaging solutions are expensive, time-consuming, and require specialized equipment; hence, they are not feasible for frequent testing.

To overcome the above-mentioned limitations, we develop a stacking machine learning approach to identify COVID-19 patients using blood-test biomarkers that are widely used in clinical practice and take a few minutes (under emergency conditions) and are much cheaper than RT-PCR test, chest X-ray, and CT imaging. Furthermore, we develop a nomogram-based scoring technique using the machine learning approach and the scoring technique uses complete blood count (CBC) parameters to diagnose COVID-19 patients in resource-constrained situations and countries, where the availability and application of RT-PCR are limited. Moreover, the developed method can be utilized in combination with the RT-PCR to improve its sensitivity. Different biomarkers have been used in several recent works on machine learning-based early mortality prediction systems [18–22]. However, to the best of the authors' knowledge, there is no prior work employing such biomarkers to diagnose COVID-19 patients that are trained with a large dataset and validated externally on datasets from different countries. Fast screening

of COVID-19 with high sensitivity is crucial for both resource and treatment planning [19,20,23–26]. Therefore, this study proposes a reliable COVID-19 diagnostic technique that can be generalized in other health care settings. It adds to the body of knowledge for developing a framework using a stacking machine learning technique and validates the performance with six completely unseen datasets.

The remaining article is organized as stated: Section II explains the detailed methodology of the study along with the datasets used for this study, the data preprocessing stages, and the nomogram-based scoring technique, while Section III shows the results. Moreover, Section IV discusses the results and validates the performance of the nomogram-based scoring technique. Finally, the article is concluded in Section V discussing the future implications of this study.

2. Methodology

The study has used seven publicly available clinical datasets [27] for model training, internal and external validation. Among them, one was used for training and internal validation whilst others were used for external validation. Feature Engineering was done extensively on getting the best-trained model which involved pre-processing the data, using different feature ranking techniques, applying different machine learning classifiers with the help of top-ranked features. We proposed a stacking-based ML model leveraging the top-performing classifiers and performance compared with other conventional ML classifiers to identify COVID-19 positive patients. The trained model is used to develop a nomogram-based scoring system (**Qu Covid Stacking-based ML (QCovSML)**) system, which is externally validated with six clinical datasets collected from different hospitals and countries to prove/validate its robustness. Fig. 1 illustrates the schematic overview of the methodology.

2.1. Database description/study population

In this study, we used seven different hospital clinical biomarker datasets [27] from three different countries: Italy, Brazil, and Ethiopia. While the largest dataset from Italy was kept for training, the remaining datasets were used for external validation. The training dataset had a total of 1624 instances and 21 features, collected upon admission at the emergency departments (ED) of two hospitals - IRCCS Hospital San Raaele (OSR) and the IRCCS Istituto Ortopedico Galeazzi (IOG), which were largely impacted during the first COVID19 pandemic in Milan, Italy. The dataset spanned from March 5, 2020 to May 26, 2020, and it was sufficiently balanced and heterogeneous.

We have validated the developed model with the following six external datasets.

- **The Italy-1 dataset:** This dataset contains 337 instances (163 positives, 174 negative) collected in March/April at the Desio Hospital.
- **The Italy-2 dataset:** It was collected at the "Papa Giovanni XXIII" Hospital of Bergamo, containing 249 patients' data with 104 COVID-19 positive, and 145 COVID-19 negative patients. The dataset was collected between March and April of 2020.
- **The Italy-3 dataset:** This dataset includes 224 patients' data with 118 COVID-19 positive and 106 COVID-19 negative patients collected in November 2020 at the IRCCS Hospital San Raaele.
- **The Brazil-1 dataset:** This dataset was collected in the Fleury private clinics from Sao Paulo of Brazil, containing 1301 patients' data with 352 COVID-19 positive, and 949 COVID-19 negative patients.
- **The Brazil-2 dataset:** It has 345 patient instances with 334 COVID-19 positive and 11 COVID-19 negative patients collected from the Hospital Srio-Libanes in Sao Paulo.

Both Brazil-1 & 2 were collected between February 2020 and June 2020.

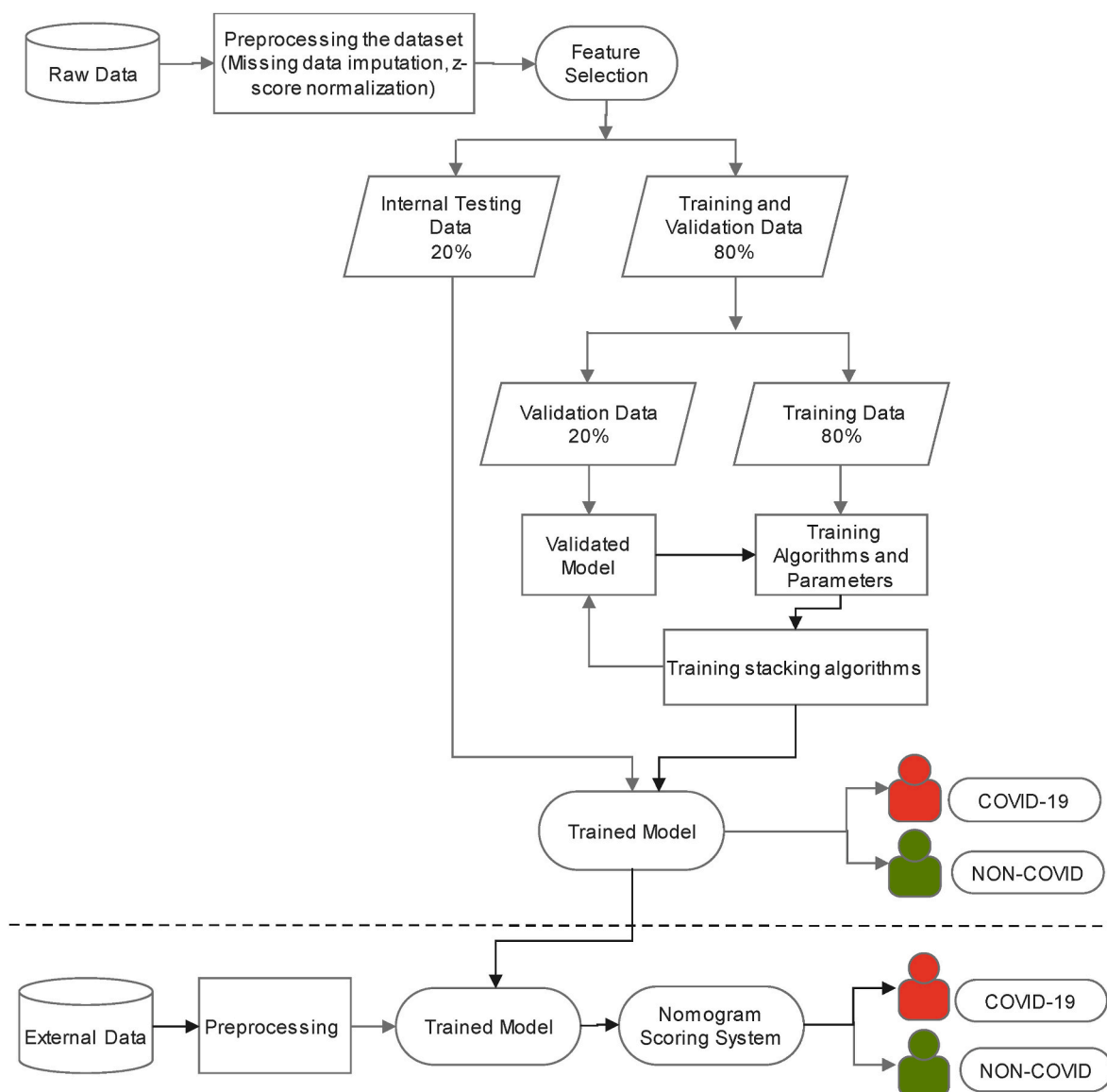


Fig. 1. Step-by-step overview of the methodology.

● **The Ethiopia dataset:** This dataset contains 200 COVID-19 patients' data, which were collected between January and March 2021 at the National Reference Laboratory for Clinical Chemistry, Millennium COVID-19 Treatment and Care Center, the Ethiopian Public Health Institute in Addis Ababa.

Table 1 shows the dataset information for COVID-19 and normal classes in detail.

2.2. Statistical analysis

We performed all statistical analysis using Python 3.7, where age, continuous variables, and other clinical data were reported (with a mean and standard deviation) for each biomarker for COVID-19 and Normal groups. The Chi-square univariate test and rank-sum test were employed to find the statistically significant features in both groups. We determined the significance with the P-value < 0.05. Gender, age, and thirteen different features were identified from the database. Table 2 summarizes 14 parameters (age, and 13 clinical biomarkers) and their statistical characteristics.

In the training data, there are 786 (48.4%) COVID-19 and 838

Table 1 Dataset description for Model development, internal and external validation.

	Dataset	COVID-19	Non-COVID	Total
^a Training (80% of OSR)	Italy	629	670	1299
^a Balanced Training Set (OSR)		670	670	1340
^a Internal validation (20% of OSR)		157	168	325
External validation	Italy-1	163	174	337
	Italy-2	104	145	249
	Italy-3	118	106	224
	Brazil-1	352	949	1301
	Brazil-2	334	11	345
	Ethiopia	200	-	200

^a Five-fold cross validation was used for performance evaluation. Number of samples per fold for training, augmented training, and internal validation is reported here.

(51.6%) Non -COVID patients. The average age of the subjects in the COVID-19 and Normal groups were 61.85 ± 16.3 and 59.3 ± 22.24 , respectively. 507 (53%) and 279 (41.2%) subjects were COVID-19

Table 2
Statistical analysis of the COVID-19 and Non -COVID groups' characteristics using the internal training dataset.

Features	Unit	Acronym	Missing rate (%)	COVID-19	Non-COVID	Overall	p-value
				mean ± std	mean ± std	mean ± std	
Age	years	Age	0	61.85 ± 16.3	59.3 ± 22.24	60.54 ± 19.61	<0.05
White blood cells	10 ⁹ /L	WBC	2.4	7.66 ± 3.88	9.72 ± 5.17	8.7 ± 4.69	<0.05
Red blood cells	10 ¹² /L	RBC	3.6	4.65 ± 0.68	4.44 ± 0.74	4.54 ± 0.72	0.112
Hemoglobin	g/dl	HGB	2.4	13.54 ± 1.89	12.86 ± 2.09	13.2 ± 2.02	0.545
Hematocrit	%	HCT	2.4	40.27 ± 5.3	38.55 ± 5.7	39.41 ± 5.56	<0.05
Mean corpuscular volume	fL	MCV	3.6	86.9 ± 6.67	87.5 ± 7.41	87.2 ± 7.06	0.288
Mean corpuscular hemoglobin	pg/Cell	MCH	3.6	29.23 ± 2.63	29.18 ± 2.8	29.2 ± 2.73	0.655
Mean corpuscular hemoglobin concentration	g Hb/dL	MCHC	2.4	33.62 ± 1.33	33.32 ± 1.36	33.47 ± 1.35	0.881
Platelets	10 ⁹ /L	PLT1	3.6	222.73 ± 90.78	246.16 ± 97.5	234.5 ± 94.8	<0.05
Neutrophils count	10 ⁹ /L	NET	18.9	5.88 ± 3.6	7.2 ± 5.4	6.47 ± 4.52	<0.05
Lymphocytes count	10 ⁹ /L	LYT	15.2	1.15 ± 0.84	1.64 ± 1.04	1.37 ± 0.96	<0.05
Monocytes count	10 ⁹ /L	MOT	15.2	0.54 ± 0.6	0.71 ± 0.39	0.61 ± 0.5	<0.05
Eosinophils count	10 ⁹ /L	EOT	15.2	0.023 ± 0.08	0.11 ± 0.18	0.064 ± 0.14	<0.05
Basophils count	10 ⁹ /L	BAT	15.2	0.005 ± 0.02	0.029 ± 0.052	0.016 ± 0.04	0.075

positive among males and females, respectively. The mean white blood cells (WBC) of the study population were 7.66 ± 3.88 ($10^9/L$) for the COVID-19 group and 9.72 ± 5.17 ($10^9/L$) for the Non -COVID group. For monocytes count (MOT), the mean of the study population was 0.54 ± 0.6 ($10^9/L$) for the COVID-19 group and 0.71 ± 0.39 ($10^9/L$) for the Non -COVID group. Moreover, the mean lymphocytes count (LYT) of the study population was 1.15 ± 0.84 ($10^9/L$) for the COVID-19 group whereas 1.64 ± 1.04 ($10^9/L$) for the Non -COVID group.

2.3. Data preprocessing

Before extracting features from the dataset and passing it to the training steps, we followed three preprocessing steps, namely data imputation, normalization, and balancing, to condition the dataset.

2.3.1. Missing data imputation

Identification and missing data imputation are the primary preprocessing steps in this study. For each patient, multiple blood biomarkers were acquired. However, some parameters are missing among the patients. The data collected during admission for the identification of the infection was used for training and validation. Data imputation techniques were used for the missing data for the key predictors instead of the simplest technique removing the instances, which can cause the loss of vital and contextual information and affect the generalized representation of the dataset [26].

In recent years, machine learning-based data imputation techniques are widely used for missing value imputation. However, this approach

requires a separate model for each column containing missing values. We used multivariate imputation by chained equations (MICE), which has become a popular method of handling missing data and has outperformed other imputation techniques especially for clinical data [28]. Fig. 2 illustrates the missing values for different features in the OSR (training) dataset. While the majority of the columns are well-populated, lymphocytes, monocyte, basophils, and neutrophils have some missing values. Moreover, the spark lines at the right side of the figure denote data completeness.

2.3.2. Normalization

The success of the machine learning models is highly dependent on the input data quality for achieving generalized performance. Data normalization involves scaling or transforming the data to make each feature contribute equally during the training process. Many studies testified performance enhancement of the machine learning models employing such normalization techniques [29]. In this study, we employed Z-score normalization using the formula:

$$v' = \frac{v - \mu_v}{\sigma_v} \tag{1}$$

where v' , v , μ_v , and σ_v denote the new value, original value, mean, and standard deviation of the variable values in the training samples, respectively. This method transforms the data with a mean of 0 and a standard deviation of 1. Moreover, this algorithm is sensitive to outliers.

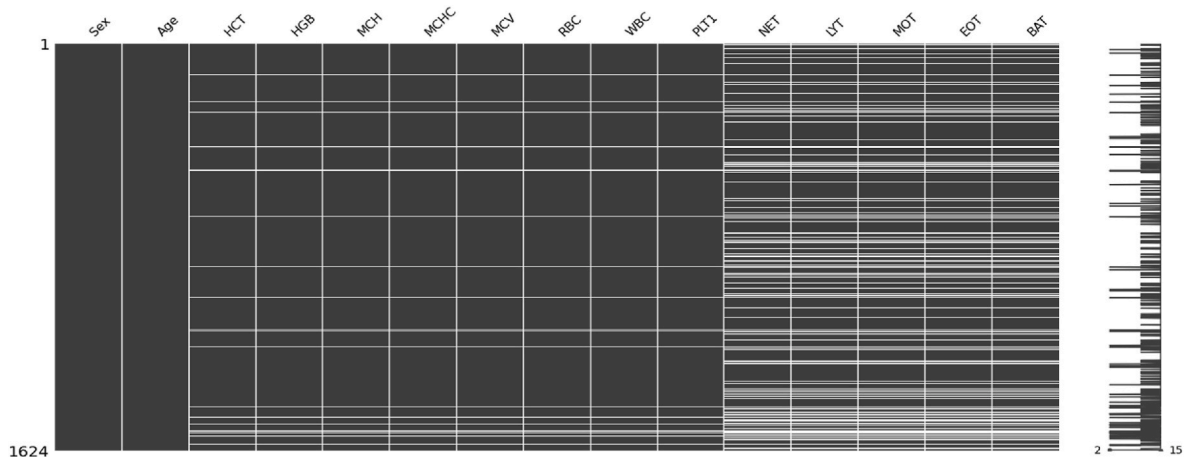


Fig. 2. The number of missing data for different features in the OSR dataset (Training data). The missing data are shown as spottier, and the spark-line at right shows the shape of the dataset.

2.3.3. Data balancing

Data imbalance can cause bias in the machine learning models. Therefore, dataset balancing is considered another important step to achieve model performance. The synthetic minority oversampling technique (SMOTE) is a frequently used method to resolve the data imbalance [30]. In this study, the number of Non-COVID patients for training was 670 while the number of COVID patients was 629, which is slightly unequal not entirely imbalance. However, we used SMOTE to up-sample the data of the COVID class; thus, balancing both classes of the training dataset.

Fifteen features were already present in the training dataset that were evaluated carefully to determine the correlation among them. Fig. 3 (A) shows the heatmap of correlation, and it was found that few features are not highly correlated. A maximum correlation of 0.97 was found between HCT and HGB. Moreover, four features had a high correlation ($r > 0.85$) with other features, which were removed from the dataset. Fig. 3(B) shows the correlation heatmap after removing highly correlated features. Fig. 4 shows the marginal distribution of the data in each feature.

2.4. Feature selection

In this study, five different feature selection techniques were used namely, chi-square test, Pearson correlation coefficient, recursive feature elimination (RFE), logistic regression, and random forest to find the best feature combination for detecting COVID-19 positive patients with high probability. Moreover, we calculated the feature importance score for each feature using five feature selection techniques, then an average of the feature importance score was used as a threshold to select the features for each technique. Finally, we have selected those features which are exceeded the threshold for all five feature selection techniques to develop the classification model.

Firstly, the chi-square metric between the target and the expected values was calculated and the features with the highest chi-squared values were selected. The intuition is that if a feature is independent of the target, it is uninformative for classifying observations. The chi-square test is represented by,

$$\text{Chi - Square test, } X^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i} \quad (2)$$

Where O_i is observation in a class and E_i is the expected observations in class.

We also used Pearson's correlation method for finding the association among the features. If X, Y are two features of length N; x_i, y_i are the values of X, Y, and \bar{x} and \bar{y} are the means of x_i 's and y_i 's then,

$$\text{Correlation (X, Y), } r = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^N (y_i - \bar{y})^2}} \quad (3)$$

Recursive feature elimination (RFE) is another technique, which is often used for feature selection, which selects features recursively considering smaller feature sets. At first, the estimator takes the whole feature set into account and the importance scores of individual features are evaluated. Afterward, the least important features are excluded from the feature set. After several recursive feature exclusion steps, the desired number of features is finally chosen [31]. Moreover, we used logistic regression and random forest for feature selection, where logistic regression is often used for binary classification. Assume that we have a data set $\{\vec{x}_i, y_i\}; 1 \leq \vec{x}_i \leq N$, where \vec{x}_i are the vectors of input features and $y_i \in \{0, 1\}$ are binary response values [32]. If p is the predicted probability $P(y = 1)$ and $\vec{\beta}$ is the vector of model parameters, logistic regression can represent the event's log odds as a linear model:

$$\log \frac{p}{1-p} = \vec{\beta} \cdot \vec{x} \quad (4)$$

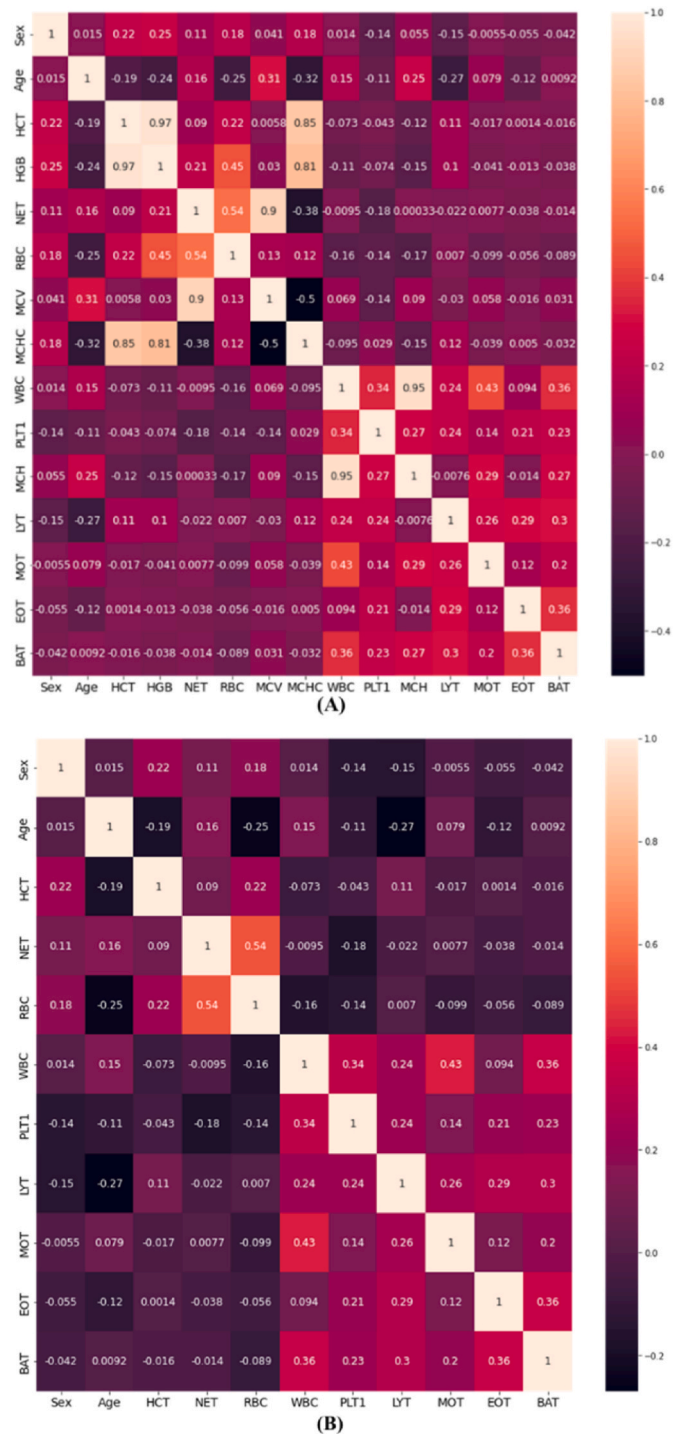


Fig. 3. Heatmap of correlation among different features (A) using all features, and (B) removing highly correlated features.

On the other hand, Random Forest (RF) combines several decision trees, which shows better performance compared to a single tree classifier. It makes the combination by the bootstrap aggregation method during a single decision tree construction along with random data node selection [31]. For example, a decision tree with N leaves clusters the feature space into N regions $R_n, 1 \leq n \leq N$. where N and R_n respectively denote the total region number and nth region in the feature space.

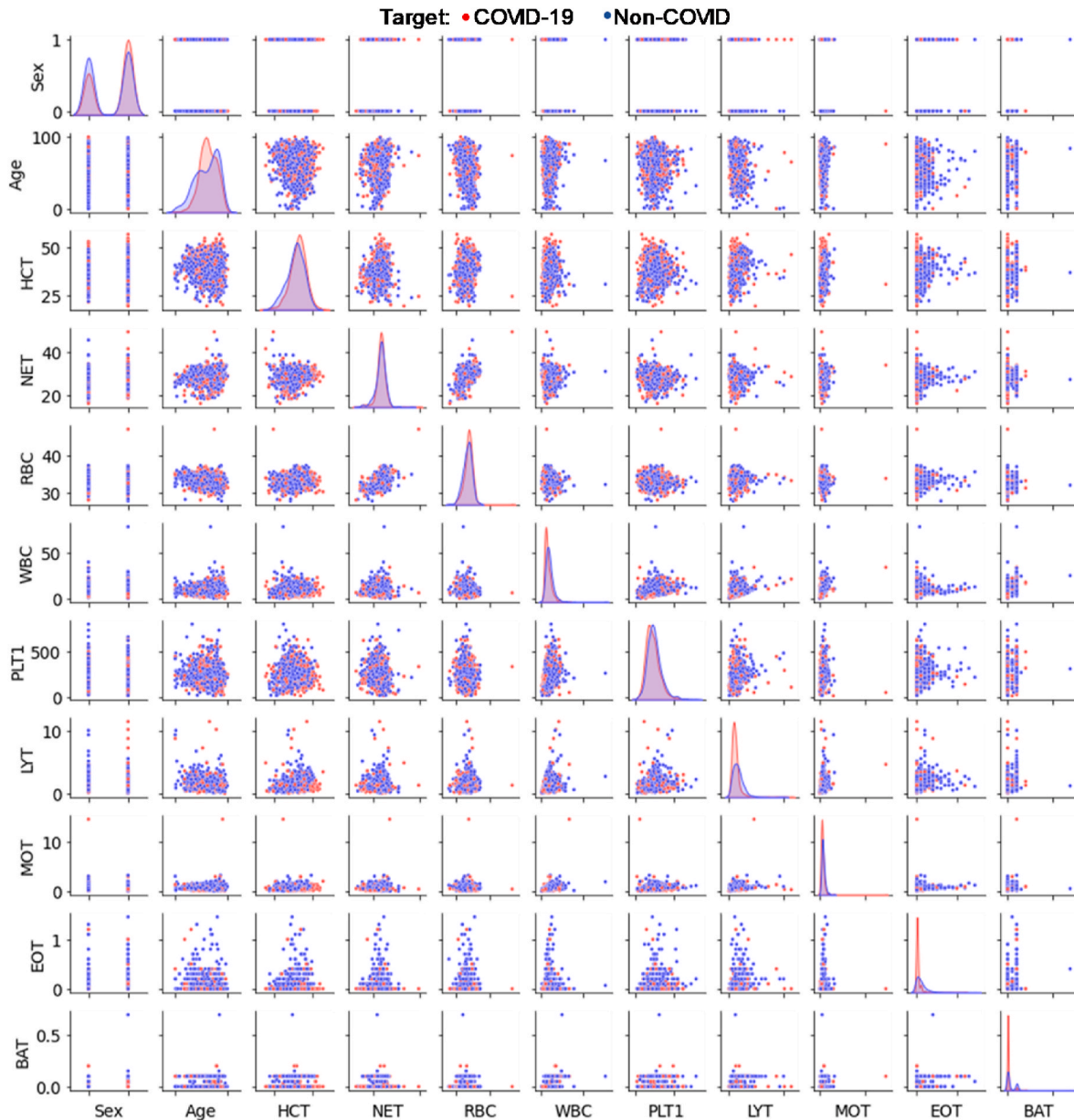


Fig. 4. Pair plot for the distribution of the dataset.

2.5. Stacking-based machine learning model

In this study, a stacking-based classifier is proposed which combines multiple best performing model on a single dataset S, which consists of

feature vectors (x_i) and their ground truth labels (y_i). Here p, q denotes the number of subjects and features, respectively. At first, N number of base-level classifiers M_1, \dots, M_N trained using the S (x_i, y_i) dataset; the predictions probabilities were then used to train the meta-level classifier

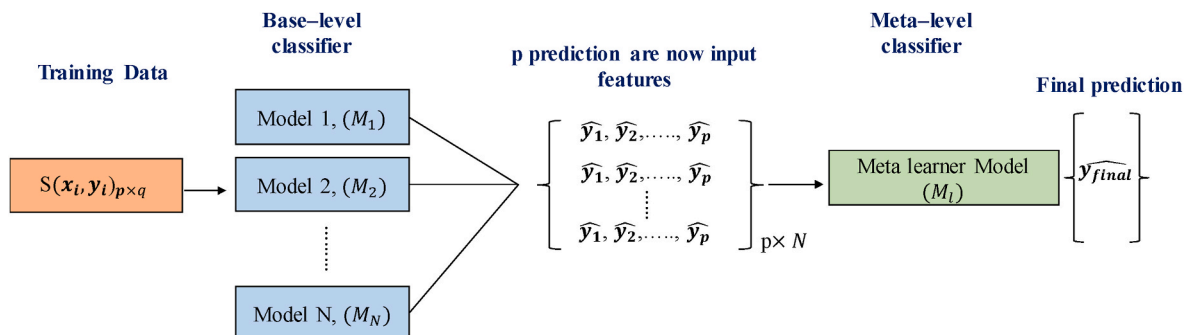


Fig. 5. Proposed stacking model architecture.

to produce final prediction, as illustrated in Fig. 5.

We used five-fold cross-validation to generate a training set for the meta-level classifier M_j . Among these folds, base-level classifiers were used on four-folds, leaving one-fold for validation. Each base-level classifier predicts a probability distribution over the possible class values. Thus, using input x_i , a probability distribution is created using the predictions of the base-level classifier set:

$$\mathbf{P}^M(\mathbf{x}) = (\mathbf{P}^M(c_1|\mathbf{x}), \mathbf{P}^M(c_2|\mathbf{x}), \dots, \mathbf{P}^M(c_m|\mathbf{x})) \tag{5}$$

where (c_1, c_2, \dots, c_m) is the set of possible class values m , and $\mathbf{P}^M(c_i|\mathbf{x})$ denotes the probability that example x belongs to a class c_i as estimated (and predicted) by classifier M_N . The meta-level attributes are thus the probabilities predicted for each possible class by each of the base-level classifiers, i.e., $\mathbf{P}^{M_j}(c_i|\mathbf{x})$ for $i = 1, \dots, m$ and $j = 1, \dots, N$.

2.6. Development and internal validation of classification model

We investigated eight machine learning classifiers in this study, namely Random Forest [33], Support Vector Machine (SVM) [34], K-nearest neighbor (KNN) [35], XGBoost [36], Adaboost [37], Gradient boosting, linear discriminant analysis (LDA) [38], and Logistic regression [39]. We used four top-ranked features to compare different classification models, which were selected by five feature selection techniques. The best three performing classifiers were selected as base learner models (M_1, M_2, M_3) in the stacking architecture and logistic regression was used for the meta learner model (M_j) in the second phase of the stacking model and finally produced different performance matrices based on the prediction.

2.7. Nomogram-based scoring system and external validation

Nomograms are a frequently used graphical estimation technique that reduces the statistical models into a single probability estimation of an event. In Zlotnik’s Nomolog [40], a diagnosis nomogram was constructed based on multivariate logistic regression analysis. In a binary classification scenario, logistic regression provides the probability of a class ranging from 0 to 1.

In this study, a nomogram-based scoring system (QCoVSMML) using logistic regression was created using the probability scores from the top base learner models. Six different hospital datasets from three different countries (Italy, Brazil, and Ethiopia) were utilized as the external validation sets to evaluate the nomogram-based stacking ML system. Moreover, to compare the predicted and actual probability of COVID-19 patients, calibration curves for internal and external validation were plotted. Moreover, we used decision curve analysis (DCA) in Stata software for finalizing the ranges of threshold of individual algorithms to identify the effectiveness of individual algorithms.

2.8. Performance metrics

To evaluate the performance of the classifiers, we used the receiver operating characteristic (ROC) curves and area under the curve (AUC) along with Precision, Sensitivity, Specificity, Accuracy, and F1-Score. Moreover, we used five-fold cross-validation, which results in an 80% and 20% split for train and test sets, respectively and according to the fold number, this process is repeated 5 times to validate the entire dataset.

Since different classes had a different number of instances, we used per class weighted metrics and overall accuracy. Additionally, we used the AUC value as another evaluation metric. The mathematical representation of five evaluation metrics (weighted sensitivity or recall, specificity, precision, overall accuracy, and F1 score) are shown in Equations 6–10:

$$Accuracy_{class-i} = \frac{TP_{class-i} + TN_{class-i}}{TP_{class-i} + TN_{class-i} + FP_{class-i} + FN_{class-i}} \tag{6}$$

$$Precision_{class-i} = \frac{TP_{class-i}}{TP_{class-i} + FP_{class-i}} \tag{7}$$

$$Recall / Sensitivity_{class_i} = \frac{TP_{class_i}}{TP_{class_i} + FN_{class_i}} \tag{8}$$

$$F1_score_{class_i} = 2 \frac{Precision_{class_i} \times Sensitivity_{class_i}}{Precision_{class_i} + Sensitivity_{class_i}} \tag{9}$$

$$Specificity_{class-i} = \frac{TN_{class-i}}{TN_{class-i} + FP_{class-i}} \tag{10}$$

where $class_i = Non - COVID$ and $COVID - 19$

Here, true positive, true negative, false positive, and false negative are represented as TP, TN, FP, and FN, respectively.

3. Results

3.1. Best feature combination for prediction of COVID-19 positive patients

As mentioned earlier, five different feature selection techniques were used to select a feature combination, where we calculated feature importance using these feature selection techniques to select the final best feature combination. We used an average of the feature importance as a threshold to select the features. All of the five feature selection techniques selected four features, such as white blood cell count (WBC), monocyte count (MOT), age, and lymphocyte count (LYT) to classify COVID-19 and Non-COVID patients while other features were not selected by all five feature selection approaches, as shown in Table 3.

3.2. Development and internal validation of the stacking model

The selected four features were tested on eight different ML classifiers using five-fold cross-validation to identify which models performed well to diagnose COVID-19 positive patients. Gradient boosting classifier outperformed others with weighted precision, sensitivity, F1-score, specificity, and overall accuracy of 89.86%, 89.88%, 89.88%, 89.87%, and 89.88%, respectively. The top three algorithms (Gradient Boosting, Random Forest, and XGBoost) were used for the next step. The accuracies of the three best algorithms are 89.88%, 82.91%, and 81.43% respectively. A logistic regression model was used as a meta learner in the stacking model. The stacking model provides the best performance with weighted precision, sensitivity, F1-score, specificity, and overall accuracy, of 91.44%, 91.45%, 91.45%, 91.44%, and 91.45%, respectively. Therefore, the proposed stacking model performance outweighs

Table 3
Feature ranked according to different feature selection algorithms.

Feature	Pearson	Chi-2	RFE	Logistics	Random Forest	Total
WBC	✓	✓	✓	✓	✓	5
MOT	✓	✓	✓	✓	✓	5
Age	✓	✓	✓	✓	✓	5
LYT	✓	✓	✓	✓	✓	5
EOT	✓	✓	✓		✓	4
NET	✓	✓	✓	✓		4
RBC	✓	✓	✓		✓	4
HCT	✓	✓	✓			3
PLT1	✓	✓	✓			3
Sex	✓	✓	✓			3
BAT	✓	✓	✓			3

other state-of-the-art machine learning classifiers. The performance of the stacking model showed about 2% improvement in terms of all the different performance matrices. Table 4 compares the overall accuracies and weighted average performance of all experimented classifiers with a 95% confidence interval to diagnose COVID-19 patients. Fig. 6 clearly shows ROC curves for the different ML classifiers and the proposed stacking model, where the proposed model produced 96% AUC and outperformed different state-of-the-art ML classifiers.

3.3. Nomogram-based stacking ML model and external validation

The study also developed a nomogram based on multivariate logistic regression analysis using the probability scores of the three best models (namely Gradient Boosting (M1), Random Forest (M2), and XGBoost (M3)) which identify the NON-COVID and COVID-19 subjects reliably. The relationship between these base learner models' probability scores and the probability of COVID-19 positive patients was evaluated by the multivariate logistic regression analysis (Table 5). The z-value is a widely used method for determining a data position in the distribution, which is calculated by dividing the regression coefficient by its standard error. If the z-value is too big, the corresponding true regression coefficient is not 0, and the corresponding independent variable is significant.

Therefore, out of 3 probability scores, for COVID-19 subjects, using the three best models, XGBoost (M3) is not a very strong predictor but Gradient Boosting (M1) and Random Forest (M2) are strong ones. P-value can determine if an X-variable has a significant relationship to the Y variable by testing the null hypothesis that the corresponding regression coefficient is 0. The X-variables with $p < 0.05$, have a significant relationship to Y-variables according to the null hypothesis. This is also reflecting that the probability scores of M3 are weakly related to Y-variable.

Thus, to build the nomogram prediction model, these 3-probability scores were used as predictors. As illustrated in Fig. 7, there were 6 rows in the nomogram, ranging from 1 to 3 representing the included variables. For each variable from COVID-19 or NON-COVID group, a score was derived by drawing a vertical line downward from the value on the variable axis to the "points" axis. The points of the three variables indicated the scores (row 4) and the scores were added to derive the total

Table 4

Comparison of the average performance metrics from five-fold cross-validation for different classifiers and the stacking classifier.

Classifier	Overall	Weighted with 95% CI			
	Accuracy	Precision	Recall	F1-score	Specificity
Linear Discriminant Analysis (LDA)	67.88 ± 2.27	67.69 ± 2.27	67.88 ± 2.27	67.88 ± 2.27	67.77 ± 2.27
XGBoost (XGB)	81.43 ± 1.89	81.37 ± 1.89	81.43 ± 1.89	81.43 ± 1.89	81.39 ± 1.89
Random Forest (RF)	82.91 ± 1.83	82.87 ± 1.83	82.91 ± 1.83	82.91 ± 1.83	82.74 ± 1.84
Logistic Regression (LR)	68.37 ± 2.26	68.63 ± 2.26	68.37 ± 2.26	68.37 ± 2.26	68.47 ± 2.26
Support Vector Machine (SVM)	62.28 ± 2.36	70.03 ± 2.23	62.28 ± 2.36	62.28 ± 2.36	61.53 ± 2.37
AdaBoost	74.66 ± 2.12	74.45 ± 2.12	74.66 ± 2.12	74.66 ± 2.12	74.22 ± 2.13
K-Nearest Neighbors (KNN)	79.17 ± 1.97	79.11 ± 1.98	79.17 ± 1.97	79.17 ± 1.97	79.13 ± 1.98
Gradient Boosting (GB)	89.88 ± 1.47	89.86 ± 1.47	89.88 ± 1.47	89.88 ± 1.47	89.87 ± 1.47
Stacking model (GB + RF + XGB)	91.45 ± 1.36	91.44 ± 1.36	91.45 ± 1.36	91.45 ± 1.36	91.44 ± 1.36

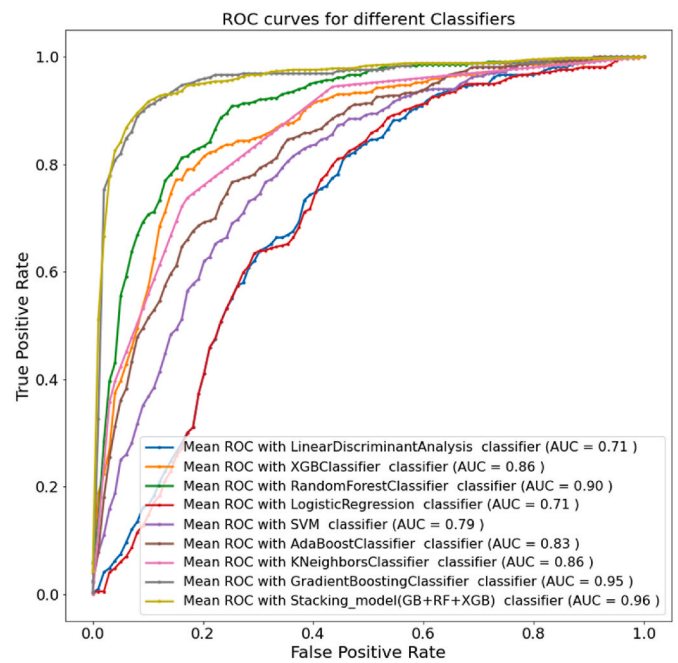


Fig. 6. ROC curves for different ML classifiers and the stacking ML model.

score, which was displayed in row 6. Afterward, to determine the probability of patients with COVID-19, a line could be drawn from the "Total Score" axis to the "Prob" axis (row 5).

Alternatively, the following formula can be used to calculate the nomogram score:

$$\text{Linear prediction} = -3.516128 + 6.685314 \times \text{Gradient Boosting (M1)} + 1.3158 \times \text{Random Forest (M2)} + 0.6338573 \times \text{XGBoost (M3)} \quad (11)$$

$$\text{Probability of COVID Infection (QCovSML)} = 1/(1 + \exp(-\text{Linear Prediction})) \quad (12)$$

The corresponding classification score (QCovSML) was calculated and listed in Fig. 8. For the development dataset (Training set), it was observed that a QCovSML score of 4.8, which denotes 40% cutoffs of classification probability provides best COVID-19 and NON-COVID stratification. Therefore, we used 4.8 as the cutoff value to classify a subject into NON-COVID and COVID-19 groups, where greater than 40% probability refers to COVID-19 infection.

3.4. Performance evaluation of the nomogram-based model

Six external datasets collected from six different hospitals from three different countries (Italy, Brazil, and Ethiopia) were used to validate the nomogram-based scoring system. According to Table 6, the multivariate logistic regression-based nomogram model performed well for internal and external validation data. For testing data (internal validation set), it produced weighted precision, sensitivity, specificity, F1-score, and overall accuracy of 89.5%, 91.12%, 90.3%, 89.71%, and 91%, respectively. For the six external datasets, the nomogram score produced weighted precision, sensitivity, specificity, F1-score, and overall accuracy of 92.02%, 95.59%, 93.73%, 90.54%, and 93.34% respectively.

Fig. 9 shows that the calibration plots were near to the diagonal lines for both internal and external validation, which is the indication of a reliable model. Moreover, Fig. 10 illustrates that the net benefit of every single predictor model was positive until the threshold of 0.95, which indicates that all predictors contributed to the outcome determination. In particular, the full model showed the best performance, and where, three predictors were combined in the model.

Table 5
The logistic regression analysis to construct the stacking ML based Nomogram.

Outcome	Coef.	Bootstrap Std. Err.	Z	P> z	[95% conf. Interval]	
Gradient Boosting (M1)	6.685314	0.7587142	8.81	0.000	5.198262	8.172367
Random Forest (M2)	1.3158	0.4752868	2.77	0.006	0.3842555	-2.247345
XGBoost (M3)	0.6338573	0.4685463	1.35	0.176	-0.2844766	-1.552191
cons	-3.516128	0.205643	-17.10	0.000	-3.919181	-3.113076

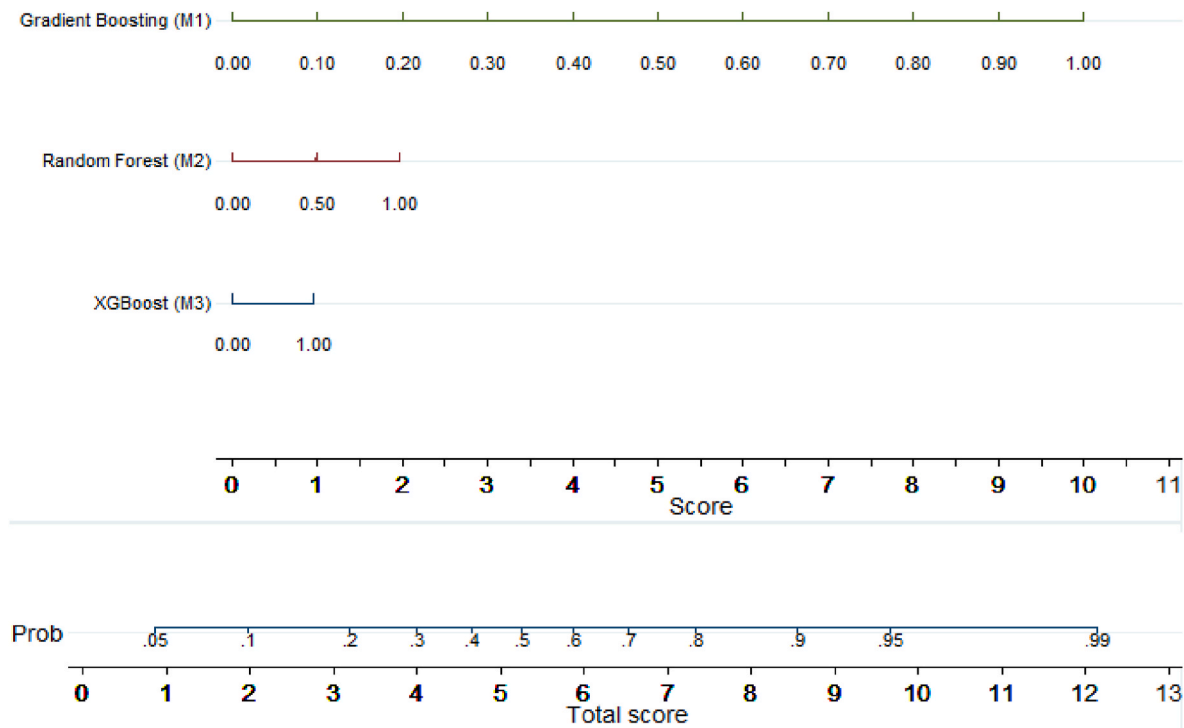


Fig. 7. Multivariate logistic regression-based Nomogram to detect COVID-19 patients. Nomogram for prediction of COVID-19 was created using Gradient boost (M1), Random Forest (M2), and XGBoost (M3).

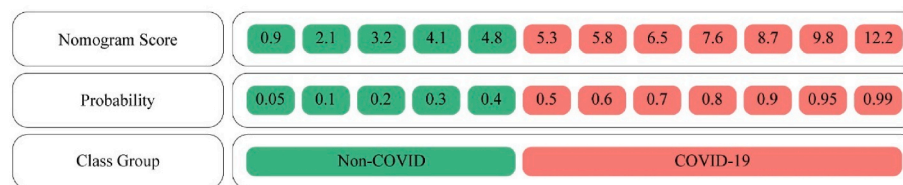


Fig. 8. Nomogram scores corresponding to the classification probabilities of COVID-19 and NON-COVID subjects.

Table 6
Comparison of the average and weighted performance metrics using logistic regression-based nomogram for the external datasets.

Dataset	Country	Overall	Weighted with 95% CI			
		Accuracy	Precision	Recall	F1-score	Specificity
Internal validation	Italy (OSR)	91 ± 1.41	89.5 ± 1.49	91.12 ± 1.38	90.3 ± 1.44	89.71 ± 1.54
External validation	Italy-1	91.69 ± 1.93	88.02 ± 2.27	97.13 ± 1.17	92.35 ± 1.86	85.89 ± 2.43
	Italy-2	95.18 ± 1.98	94.63 ± 2.08	97.24 ± 1.51	95.92 ± 1.83	92.31 ± 2.46
	Italy-3	92.86 ± 1.97	88.79 ± 2.42	97.17 ± 1.27	92.79 ± 1.98	88.98 ± 2.4
	Brazil-1	92.7 ± 2.75	94.25 ± 2.46	95.26 ± 2.24	93.25 ± 2.65	87.27 ± 3.52
	Brazil-2	95.11 ± 1.22	94.42 ± 2.06	94.25 ± 2.12	94.33 ± 2.14	95.67 ± 2.11
	Ethiopia ^a	92.5 ± 3.65	-	92.5 ± 3.65	-	-
	Average		93.34 ± 2.19	92.02 ± 2.26	95.59 ± 2.18	93.73 ± 2.33

^a Ethiopia dataset has only COVID patients.

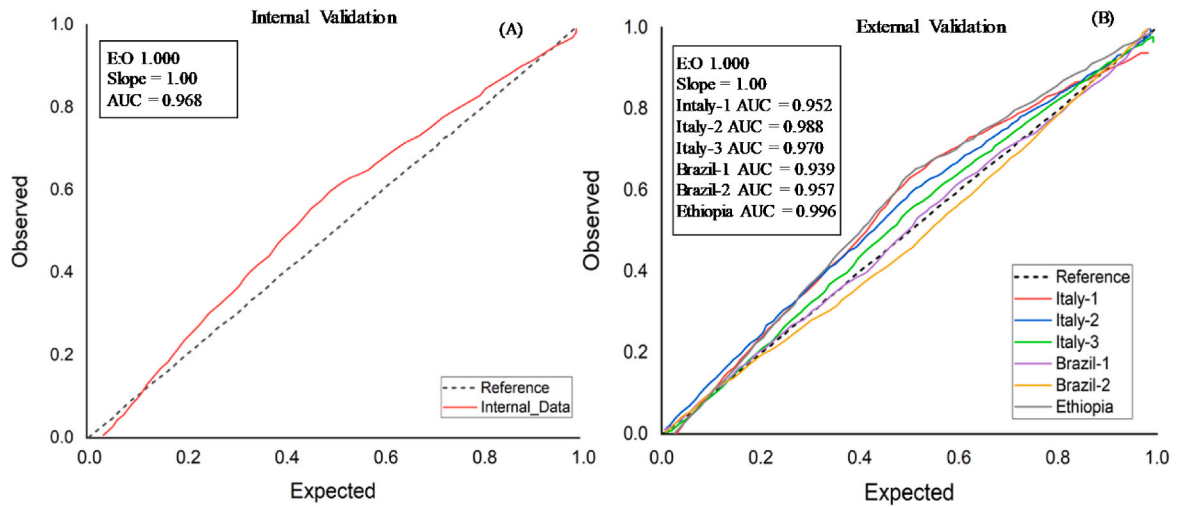


Fig. 9. Comparison of predicted and actual probability of COVID-19 patients. (A) Internal validation representation, (B) external validation representation.

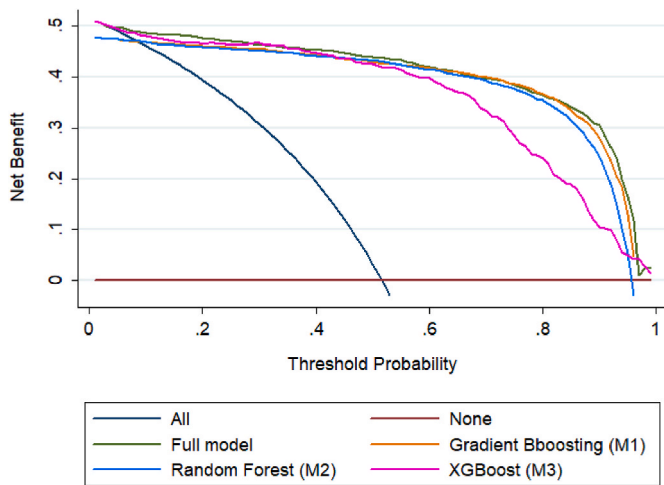


Fig. 10. Comparison of decision curves analysis of different models. The net benefit balances the probability scores for COVID-19 patients.

3.5. Web application for QCovSML

As an extension of this study, we have developed a web application that allows the doctors to put some demographic and clinical data (white blood cell count (WBC), monocyte count (MOT), age, and lymphocyte count (LYT)) to the web interface and our AI-based application analyzes the data to predict whether the user is COVID-19 positive or not. This application was developed using Flutter, a cross-platform app (android and iOS) development framework using the Dart programming language maintained by Google. This will in essence provide us with the maximum coverage or users, quicker development and continuous integration, seamless deployment and maintenance, easier cloud integration and increase stability.

In the prototype system, the user will fill in some demographic data, as well as RT-PCR test results (if available). The RT-PCR test result will not be used for model but it can be used to enhance our model in the future with a larger dataset collected through this online framework. Next, the app will ask to input only four parameters: Age, WBC, MOT, and LYT. After collecting the data from the user, it will be transferred to our server which performs pre-processing and passes through the proposed trained model to determine whether the user is COVID-19 positive or not (Fig. 11). The AI backend will analyze the data and return a response to the screen. The application will display the results as well as store them in a local database based on SQLite. In brief, the application can help in screening the COVID patients quickly with few blood

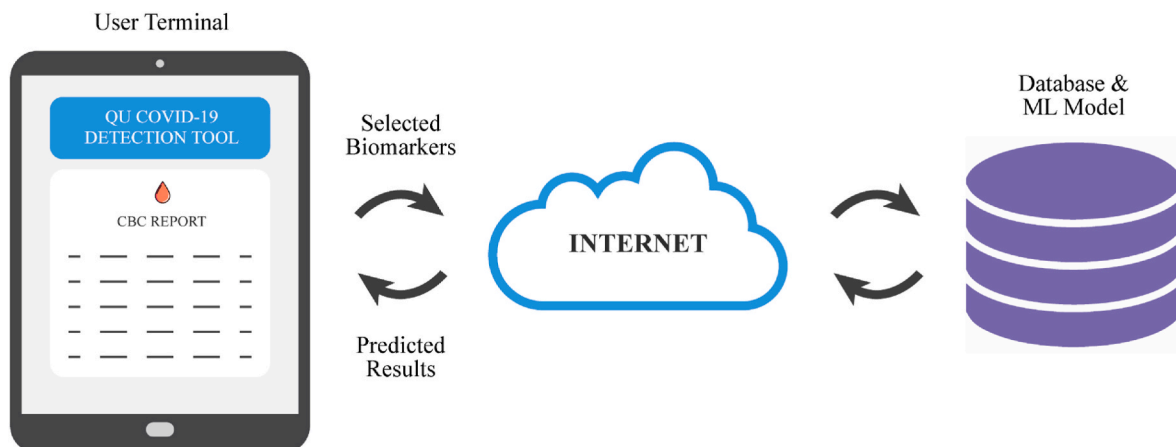


Fig. 11. Illustration of a generic framework for the COVID-19 detection tool.

biomarkers and can help to reduce the load on the healthcare system.

4. Discussion

Our study showed that CBC biomarkers can provide useful information about a probable SARS-CoV2 infection, allowing clinicians to make better judgments during hospital admission. To diagnose the COVID-19 positive patients, four impactful features: Age, WBC, MOT, and LYT were identified by the ML model. These top-ranked features were used for training and validation using different ML models, and finally, the three top-ranked models were used to train a stacking model. The stacking model produced weighted precision, sensitivity, F1-score, specificity, and overall accuracy of 91.44%, 91.45%, 91.45%, 91.44%, and 91.45%, respectively, which are comparatively better than the best performing Gradient Boosting classifier with overall accuracy, and weighted precision, sensitivity, specificity, and F1-score of 89.88%, 89.86%, 89.88%, 89.88%, and 89.87%, respectively. Moreover, a nomogram-based scoring system (QCovSML) using this stacking technique was developed, which provides 91% overall accuracy. The nomogram-based scoring system was validated with six external datasets. The overall average accuracy for the six datasets is 93.34%. The QCovSML technique requires only easy-to-obtain biomarkers, which makes it useful for the diverse environments for both in and out-patient settings. This can be particularly useful for resource-constrained countries where RT-PCR sometimes becomes scarce, and the laboratory facility becomes unavailable as well due to reagent and expert limitations.

Several studies have shown that COVID-19 patients had a less total count of leukocytes and particular of its subpopulations than healthy people and Non-COVID-19 patients with various infectious illnesses [41–43]. As a result, the most common CBC biomarkers for identifying COVID-19 individuals, and potentially categorizing its severity are reduced neutrophil count, lymphopenia, and eosinopenia. Liu et al. [44] recently suggested that decreasing lymphocyte counts can help for predicting the severity of COVID-19 infection early. Both lymphocytes and neutrophils are essential components of our immune system since they play a vital role in infection clearance and host defense. Lymphopenia is a common symptom among COVID-19 patients, which is caused by the reduction in the number of lymphocytes in the blood [45]. We used the percentage of lymphocytes and neutrophils in this study, and similar to the previous research [46,47], found a correlation between a lower percentage of these two concentrations with extremely ill COVID-19 patients. Furthermore, bone marrow increases neutrophil production, as cytokines induce immunosuppression and lymphocyte apoptosis, which results in a rise in NLR [48].

Limited studies have shown an investigation of ML models on routine blood exams. Formica et al. [13] proposed an ML model based on CBC, with 83% sensitivity and 82% specificity for COVID-19 detection. However, the study was built on a small sample (171 patients) gathered in a short time frame. Similarly, Avila et al. [49] developed a Bayesian model, reporting 76.7% sensitivity and specificity. In our study, we have used a large CBC biomarkers dataset to develop our proposed model and used multicenter and multicountry datasets for external validation. Firstly, WBC count, monocytes (%), age, and lymphocyte (%) parameters were selected as important biomarkers out of eleven different biomarkers for COVID-19 disease prediction using five different state-of-the-art feature selection techniques. Moreover, we developed a stacking-based ML model which produced the best performance with weighted precision, sensitivity, F1-score, specificity, and overall accuracy of 91.44%, 91.45%, 91.45%, 91.44%, and 91.45%, respectively. The stacking model provides a 2% performance gain over the other state-of-the-art machine learning classifiers. Finally, developed a nomogram-based scoring system using this stacking approach and the study validated the QCovSML scoring system with six other external CBC datasets from three different countries and the average weighted precision, sensitivity, F1-score, specificity, and overall accuracy for the six external datasets are 92.02%, 95.59%, 93.73%, 90.54%, and 93.34%,

respectively. This scoring system can be used for COVID-19 detection, which can be used as an alternative method to RT-PCR for the fast and cost-effective identification of COVID-19-positive patients.

Several predictive models have been published in the literature reported age, and hematologic measures, among the main variables elucidating the predictive model to predict critically ill people [50,51]. Contrary to that, our model provides a valid utilization of such biomarkers as an early predictive tool for the timely and cost-effective diagnosis of COVID-19 disease. This could provide a significantly accessible patient triaging technique, which can ease the optimum staff allocation.

5. Conclusion

In conclusion, the proposed tool using the stacking ML model shows promising results for identifying COVID-19 positive patients with very high sensitivity using a few CBC biomarkers (WBC, Monocytes, Age, Lymphocyte count). The performance of the model reported in this study is comparable to the current gold standard for COVID-19 diagnosis, the RT-PCR test [52]. Moreover, the proposed model has a high degree of sensitivity to predict the COVID-19 positive patients in both internal and external datasets. The model was tested on completely unseen six external clinical datasets from three different countries - making this model helpful for clinicians to effectively optimize patients' stratification, devise efficient management strategies, and better resource utilization. Finally, this ML-based solution can be applied in both in-patients and out-patients as an alternative to the RT-PCR test for fast and cost-effective identification of COVID-19-positive patients. This is especially useful in the low-resource environment, where screening of patients can be done using the proposed method and selective RT-PCR tests can be carried out to minimize RT-PCR test and selective RT-PCR tests.

Funding

This work was supported by the Qatar National Research Fund (QNRF) Grant: UREP28-144-3-046. The statements made herein are solely the responsibility of the authors.

Dataset and code availability

The dataset used for the development and validation of this study is available at [27] and the code and models are available at <https://github.com/tawsifur/QCovSML-COVID-19-detection-system-using-CBC-bio-markers>.

Ethical approval

This study used seven different hospital clinical biomarker datasets [27] from three different countries: Italy, Brazil, and Ethiopia with de-identified patients' information. For all datasets, research involving human subjects complied with all relevant national and international regulations, institutional policies and was in accordance with the tenets of the Helsinki Declaration (as revised in 2013) and was approved by the authors Institutional Review Board (70/INT/2020).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This study acknowledges all the datasets [27] which were used for model development and external validation.

References

- [1] COVID-19 Coronavirus Pandemic [Online]. Available: <https://www.worldometers.info/coronavirus/>. [Accessed on 20-06-2021].
- [2] D.P. Oran, E.J. Topol, Prevalence of asymptomatic SARS-CoV-2 infection: a narrative review, *Ann. Intern. Med.* 173 (5) (2020) 362–367.
- [3] C.B. Vogels, A.F. Brito, A.L. Wyllie, J.R. Fauver, I.M. Ott, C.C. Kalinich, et al., Analytical sensitivity and efficiency comparisons of SARS-CoV-2 RT-qPCR primer–probe sets, *Nat. Microbiol.* 5 (10) (2020) 1299–1305.
- [4] L. Wynants, B. Van Calster, G.S. Collins, R.D. Riley, G. Heinze, E. Schuit, et al., Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal, *BMJ* 369 (2020).
- [5] Y. Zheng, Y. Zhang, H. Chi, S. Chen, M. Peng, L. Luo, et al., The hemocyte counts as a potential biomarker for predicting disease progression in COVID-19: a retrospective study, *Clin. Chem. Lab. Med.* 58 (7) (2020) 1106–1115.
- [6] Z. Weng, Q. Chen, S. Li, H. Li, Q. Zhang, S. Lu, et al., ANDC: an early warning score to predict mortality risk for patients with coronavirus disease 2019, *J. Transl. Med.* 18 (1) (2020) 1–10.
- [7] S. Al Youha, S.A. Doi, M.H. Jamal, S. Almazzeedi, M. Al Haddad, M. AlSeaidan, et al., Validation of the Kuwait Progression Indicator Score for Predicting Progression of Severity in COVID19, medRxiv, 2020.
- [8] P. Ramachandran, M. Gajendran, A. Perisetti, K.O. Elkholy, A. Chakraborti, G. Lippi, et al., Red Blood Cell Distribution Width (RDW) in Hospitalized COVID-19 Patients, medRxiv, 2020.
- [9] J. Gong, J. Ou, X. Qiu, Y. Jie, Y. Chen, L. Yuan, et al., A tool for early prediction of severe coronavirus disease 2019 (COVID-19): a multicenter study using the risk nomogram in Wuhan and Guangdong, China, *Clin. Infect. Dis.* 71 (15) (2020) 833–840.
- [10] B.H. Foy, J.C. Carlson, E. Reinertsen, R.P. Valls, R.P. Lopez, E. Palanques-Tost, et al., Elevated RDW Is Associated with Increased Mortality Risk in COVID-19, medRxiv, 2020.
- [11] X. Jianfeng, D. Hungerford, H. Chen, S. Abrams, S. Li, G. Wang, Development and External Validation of a Prognostic Multivariable Model on Admission for Hospitalized Patients with Covid-19, medRxiv, 2020.
- [12] Z. Weng, Q. Chen, S. Li, H. Li, Q. Zhang, S. Lu, et al., ANDC: An Early Warning Score to Predict Mortality Risk for Patients with Coronavirus Disease 2019, 2020.
- [13] V. Formica, M. Minieri, S. Bernardini, M. Ciotti, C. D'Agostini, M. Roselli, et al., Complete blood count might help to identify subjects with high probability of testing positive to SARS-CoV-2, *Clin. Med.* 20 (4) (2020) e114.
- [14] O. Gozes, M. Frid-Adar, H. Greenspan, P.D. Browning, H. Zhang, W. Ji, et al., Rapid AI Development Cycle for the Coronavirus (Covid-19) Pandemic: Initial Results for Automated Detection & Patient Monitoring Using Deep Learning Ct Image Analysis, 2020 *arXiv preprint arXiv:2003.05037*.
- [15] T. Ozturk, M. Talo, E.A. Yildirim, U.B. Baloglu, O. Yildirim, U.R. Acharya, Automated detection of COVID-19 cases using deep neural networks with X-ray images, *Comput. Biol. Med.* 121 (2020), 103792.
- [16] X. Mei, H.-C. Lee, K.-y. Diao, M. Huang, B. Lin, C. Liu, et al., Artificial intelligence-enabled rapid diagnosis of patients with COVID-19, *Nat. Med.* 26 (8) (2020) 1224–1228.
- [17] M.B. Weinstock, A. Echenique, J. Russell, A. Leib, J. Miller, D. Cohen, et al., Chest x-ray findings in 636 ambulatory patients with COVID-19 presenting to an urgent care center: a normal chest x-ray is no guarantee, *J. Urgent Care Med.* 14 (7) (2020) 13–18.
- [18] D. Huang, T. Wang, Z. Chen, H. Yang, R. Yao, Z. Liang, A novel risk score to predict diagnosis with coronavirus disease 2019 (COVID-19) in suspected patients: a retrospective, multicenter, and observational study, *J. Med. Virol.* 92 (11) (2020) 2709–2717.
- [19] Y.-Q. Cai, H.-Q. Zeng, X.-B. Zhang, X.-J. Wei, L. Hu, Z.-Y. Zhang, et al., Prognostic Value of Neutrophil-To-Lymphocyte Ratio, Lactate Dehydrogenase, D-Dimer and CT Score in Patients with COVID-19, 2020.
- [20] L. Zhang, X. Yan, Q. Fan, H. Liu, X. Liu, Z. Liu, et al., D-dimer levels on admission to predict in-hospital mortality in patients with Covid-19, *J. Thromb. Haemostasis* 18 (6) (2020) 1324–1329.
- [21] C. Zhang, L. Qin, K. Li, Q. Wang, Y. Zhao, B. Xu, et al., A novel scoring system for prediction of disease severity in COVID-19, *Front. Cell. Infect. Microbiol.* 10 (2020) 318.
- [22] Y. Shang, T. Liu, Y. Wei, J. Li, L. Shao, M. Liu, et al., Scoring systems for predicting mortality for severe patients with COVID-19, *EClinicalMedicine* 24 (2020), 100426.
- [23] W. Liang, J. Yao, A. Chen, Q. Lv, M. Zanin, J. Liu, et al., Early triage of critically ill COVID-19 patients using deep learning, *Nat. Commun.* 11 (1) (2020) 1–7.
- [24] C. Wang, R. Deng, L. Gou, Z. Fu, X. Zhang, F. Shao, et al., Preliminary study to identify severe from moderate cases of COVID-19 using combined hematology parameters, *Ann. Transl. Med.* 8 (9) (2020).
- [25] M.P. McRae, G.W. Simmons, N.J. Christodoulides, Z. Lu, S.K. Kang, D. Fenyo, et al., Clinical decision support tool and rapid point-of-care platform for determining disease severity in patients with COVID-19, *Lab Chip* 20 (12) (2020) 2075–2085.
- [26] H. Hegde, N. Shimpi, A. Panny, I. Glurich, P. Christie, A. Acharya, MICE vs PPCA: missing data imputation in healthcare, *Informat. Med. Unlocked* 17 (2019), 100275.
- [27] F. Cabitza, A. Campagner, F. Soares, L.G. de Guadiana-Romualdo, F. Challa, A. Sulejmani, et al., The importance of being external. methodological insights for the external validation of machine learning models in medicine, *Comput. Methods Progr. Biomed.* 208 (2021), 106288.
- [28] J.R. Stevens, A. Suyundikov, M.L. Slattery, Accounting for missing data in clinical research, *JAMA* 315 (5) (2016) 517–518.
- [29] D. Singh, B. Singh, Investigating the impact of data normalization on classification performance, *Appl. Soft Comput.* 97 (2020), 105524.
- [30] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, *J. Artif. Intell. Res.* 16 (2002) 321–357.
- [31] R.-C. Chen, C. Dewi, S.-W. Huang, R.E. Caraka, Selecting critical features for data classification based on machine learning methods, *J. Big Data* 7 (2020) 1–26.
- [32] J. Kim, J. Lee, C. Lee, E. Park, J. Kim, H. Kim, et al., Optimal feature selection for pedestrian detection based on logistic regression analysis, in: 2013 IEEE International Conference on Systems, Man, and Cybernetics, 2013, pp. 239–242.
- [33] M. Pal, Random forest classifier for remote sensing classification, *Int. J. Rem. Sens.* 26 (1) (2005) 217–222.
- [34] S.S. Keerthi, S.K. Shevade, C. Bhattacharyya, K.R.K. Murthy, Improvements to Platt's SMO algorithm for SVM classifier design, *Neural Comput.* 13 (3) (2001) 637–649.
- [35] G. Guo, H. Wang, D. Bell, Y. Bi, K. Greer, KNN model-based approach in classification, in: OTM Confederated International Conferences" on the Move to Meaningful Internet Systems, 2003, pp. 986–996.
- [36] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, Xgboost: extreme gradient boosting, R package version 0 1 (4) (2015) 1–4. 4-2.
- [37] A. Khandakar, M.E. Chowdhury, M.B.I. Reaz, S.H.M. Ali, M.A. Hasan, S. Kiranyaz, et al., A machine learning model for early detection of diabetic foot using thermogram images, *Comput. Biol. Med.* 137 (2021), 104838.
- [38] Q. Gu, Z. Li, J. Han, Linear discriminant dimensionality reduction, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2011, pp. 549–564.
- [39] C. Subasi, Logistic Regression Classifier, 2019.
- [40] A. Zlotnik, V. Abraira, A general-purpose nomogram generator for predictive logistic regression models, *STATA J.* 15 (2) (2015) 537–546.
- [41] G. Lu, J. Wang, Dynamic changes in routine blood parameters of a severe COVID-19 case, *Clin. Chim. Acta* 508 (2020) 98–102.
- [42] A. Anurag, P.K. Jha, A. Kumar, Differential white blood cell count in the COVID-19: a cross-sectional study of 148 patients, *Diabetes; Metabol. Syndr.: Clin. Res. Rev.* 14 (6) (2020) 2099–2102.
- [43] Y. Sun, J. Zhou, K. Ye, White blood cells and severe COVID-19: a Mendelian randomization study, *J. Personalized Med.* 11 (3) (2021) 195.
- [44] J. Liu, Y. Liu, P. Xiang, L. Pu, H. Xiong, C. Li, et al., Neutrophil-to-lymphocyte Ratio Predicts Severe Illness Patients with 2019 Novel Coronavirus in the Early Stage, medRxiv, 2020.
- [45] I. Huang, R. Pranata, Lymphopenia in severe coronavirus disease-2019 (COVID-19): systematic review and meta-analysis, *J. Intens. Care* 8 (2020) 1–10.
- [46] Z. Illg, G. Muller, M. Mueller, J. Nippert, B. Allen, Analysis of absolute lymphocyte count in patients with COVID-19, *Am. J. Emerg. Med.* 46 (2021) 16–19.
- [47] S. Selim, Leukocyte count in COVID-19: an important consideration, *Egypt. J. Bronchol.* 14 (1) (2020) 1–2.
- [48] M. Adamzik, J. Broll, J. Steinmann, A.M. Westendorf, I. Rehfeld, C. Kreissig, et al., An increased alveolar CD4+ CD25+ Foxp3+ T-regulatory cell ratio in acute respiratory distress syndrome is associated with increased 30-day mortality, *Intensive Care Med.* 39 (10) (2013) 1743–1751.
- [49] E. Avila, A. Kahmann, C. Alho, M. Dorn, Hemogram data as a tool for decision-making in COVID-19 management: applications to resource scarcity scenarios, *PeerJ* 8 (2020), e9482.
- [50] V.M. Castro, T.H. McCoy, R.H. Perlis, Laboratory findings associated with severe illness and mortality among hospitalized individuals with coronavirus disease 2019 in eastern Massachusetts, *JAMA Netw. Open* 3 (10) (2020) e2023934-e2023934.
- [51] J. Berenguer, P. Ryan, J. Rodríguez-Baño, I. Jarrín, J. Carratalà, J. Pachón, et al., Characteristics and predictors of death among 4035 consecutively hospitalized patients with COVID-19 in Spain, *Clin. Microbiol. Infect.* 26 (11) (2020) 1525–1536.
- [52] S. Woloshin, N. Patel, A.S. Kesselheim, False negative tests for SARS-CoV-2 infection—challenges and implications, *N. Engl. J. Med.* 383 (6) (2020) e38.