Journal of Information Systems Engineering and Management
2025, 10(18s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Prediction of COVID-19: WAOptimizer Ensemble Classification Model with Clinical Parameters

L.William Mary[1], Dr. S. Albert Antony Raj[2]
[1]*Research Scholar, Department of Computer Applications, SRM Institute of Science and Technology, Kattankulathur, Chennai.*
[2]*Professor, Deputy Dean, Department of Computer Applications SRM Institute of Science and Technology, Kattankulathur, Chennai.*
wl6649@srmist.edu.in , alberts@srmist.edu.in
Corresponding Author: Dr. S. Albert Antony Raj

ARTICLE INFO

Received: 18 Dec 2024

Revised: 28 Jan 2025

Accepted: 12 Feb 2025

ABSTRACT

Introduction:

Machine learning-based prediction systems have enormous potential to enhance clinical utility in diagnosing COVID-19. Machine learning can address prediction challenges in the healthcare sector by improving diagnostic efficiency and accuracy. Delivering high-quality services is challenging. Effective illness management and precise diagnosis are critical elements of healthcare. Machine learning is transforming better progression toward prediction. The healthcare industry is adopting machine learning to improve efficiency. This technique is essential in healthcare to detect patterns within large datasets and diagnose the disease. Previous studies have predicted COVID-19 mortality using blood biomarkers and machine-learning approaches. The outcome of the prediction method effectively predicted the non-linear relationships among blood biomarkers. In addition, the prediction includes traditional assessment techniques for monitoring pulmonary diseases, such as X-rays and CT scans. Prompt detection and virus diagnosis are essential for infection control and reducing mortality rates.

Objectives: The primary goal of this research is to predict COVID-19 positivity and negativity based on blood test data by developing a stacking ensemble classifier algorithm WA-COVID Optimizer.

Methods: This research focuses on predicting the positive and negative statuses of the disease using a blood count dataset. To achieve better prediction performance, the study aims to develop an ML-driven diagnostic framework for early-stage COVID-19 diagnosis utilizing an ensemble stacking classification method. Several supervised machine learning methods are commonly used for predictions. These include Random Forest, LightGBM, Support Vector Machine, Logistic Regression with Lasso and Ridge regularization, XGBoost, AdaBoost, Gradient Boosting Machine, Multilayer Perceptron, Deep Neural Networks, and K-Nearest Neighbors. These models are combined to construct a stacking ensemble classification model that acts as a meta-model, leveraging the strengths of the base models.

Results: The performance metrics accuracy, precision, recall, F1-score, and the area under the ROC curve (AUC-ROC) are used to evaluate the prediction model. The Matthews correlation coefficient (MCC) assesses the ROC performance metrics. The proposed stacking ensemble classifier achieved an accuracy of 85%, an AUC-ROC of 90%, an MCC of 0.66, a precision of 81%, a recall of 85%, and an F1-score of 83%.

Conclusions: We developed a new data-driven strategy, the WA-COVID Optimizer, which synergizes multiple base models with a boosting mechanism. The proposed stacking classifier, WA-COVID Optimizer, predicted the best accuracy of 84% and a ROC AUC -90% for COVID-19-positive cases. The MCC validates the classifier performance, and the evaluation score is 66%.

Keywords: COVID-19 diagnosis, ensemble stacking classification, machine learning models, clinical decision support systems, predictive modeling

INTRODUCTION

The lethal infectious virus SARS-CoV-2 caused COVID-19 contagious disease. In 2019, an infected virus was discovered for the first time in Wuhan, China. The WHO declared it a pandemic due to its fast spread [1]. The outbreak has affected millions in 229 countries, severely disrupting life [2]. Research indicates that individuals aged 40 to 69 are severely affected due to COVID-19 and are hospitalized at the highest rate [3]. This virus presents clinical features similar to those of viral pneumonia. The viral pathogens cause mild to severe respiratory infections and lead to mortality. Symptoms can range significantly in intensity from highly minor to severe. In addition, less common symptoms include skin conditions and diarrhea, fever, dry cough, and headache. It designated the associated illness as coronavirus disease 2019 (COVID-19) [4]. The virus causes mild to severe respiratory diseases, some of which can be fatal. Symptoms range significantly in intensity, from highly light to severe. In addition to less common symptoms such as skin conditions and diarrhea, common symptoms include fever, dry cough, headache, and exhaustion [5].
The SARS-CoV-2 genome size is 30 kb with two encodes of proteins: the structural and non-structural genome. The spike protein has 1,160 to 1,542 amino acids. It binds the ACE2 receptor, facilitating the virus's entry into human cells [6]. COVID-19 can be confirmed through laboratory testing. A healthcare provider may collect either a saliva sample or use a swab to obtain a specimen from the nose or throat for analysis. Diagnosis is primarily made using RT-PCR testing, which involves examining nasopharyngeal swabs or other specimens from the upper respiratory tract [7,8]. The typical turnaround time for RT-PCR testing is 48 hours [9].

RELATED WORK

Numerous research studies have utilized machine learning classifier techniques to predict COVID-19 infection. The study used data from 375 patients at a Wuhan, China, hospital to develop an ML model based on features such as lactic dehydrogenase, lymphocytes, and high-sensitivity C-reactive protein. The model validated the performance in an independent dataset of 110 patients [10]. Research highlights the feasibility and the role of routine blood tests in the early detection of COVID-19 [11]. It represents a faster and more efficient option for PCR testing with comparable performance when integrated with various ML techniques [12]. Booth et al. developed a panel consisting of five laboratory biomarkers: C-reactive protein, blood urea nitrogen, serum calcium, serum albumin, and lactic acid, based on data from 398 patients from the USA to predict the COVID-19 mortality ranges [13].

The wrapper [14], filter, and embedding are the three important strategies for feature selection. Wrapper approaches train the model by consuming diverse subsets of characteristics. Filter methods, on the other hand, choose features on their own by considering the statistical correlations between the traits and the variable of interest. Although embedded approaches are practical and consider the model when selecting features, they are designed to work with particular algorithms. Fernandes et al. developed a model using a database of 1,040 Brazilian patients. This model incorporated routine biomarkers, such as ferritin, CRP, and lymphocytes, with the intensive care unit (ICU) score to predict ICU admission, mechanical ventilation, and mortality [15].

METHODS

The COVID-19 infection prediction was constructed using the Ensemble Stacking Classifier algorithm to train and test datasets for accurate prediction. Normalized data is used throughout the training and testing to develop reliable and robust prediction models for disease prognosis.

Data Collections

Data collection is the primary step in the model development process. A publicly accessible dataset from the Zendo platform, comprising 1,724 cases with 35 attributes, was used in this investigation. Information about each patient's COVID-19 illness status, including disease instances, is provided by the dataset. The datasheet includes 814 positive data and 910 negative data related to the blood biomarkers. The target variable is the numerical values of the confirmed cases.

Preprocessing Process

Preprocessing is the basic step for converting raw data into information. The raw data is incomplete, and missing information. It combines approaches for data integration, cleaning, transformation, and reduction. The data is

integrated, enhanced, and structured for the prediction. Data is an important part of the analysis process. Data reduction, transformation, cleaning, and integration are just a few of the methods it includes. Missing values in the blood test results were managed using K-Nearest Neighbors (KNN) imputation algorithms based on the statistical mean—furthermore, the process involved detecting and removing outliers, scaling, normalization, and feature selection. Among statistical methods, the parametric standard scaler Z-score normalization proved highly effective. This normalizing technique transforms the data as the format of the mean value is 0 and a standard deviation of 1. The categorical variable value in the COVID-19 blood test dataset is gender, encoded using 0 for Males and 1 for Females.

Handling Missing Values

Initially, samples with over 75% of their features missing were discarded. Then, to address data incompleteness, we employed the k-nearest Neighbors (KNN) for missing data imputation and used the mean value derived from the nearest neighbors. The KNN identifies a data point with its closest k neighbors within a multi-dimensional distance.

Mitigating Class Imbalance with the SMOTE Technique

SMOTE is a new method for balancing datasets that generates synthetic samples for the minority class. It has emerged as one of the most widely adopted oversampling techniques for addressing challenges associated with imbalanced classification [16]. In the initial evaluation of the COVID-19 blood test dataset, we observed an imbalance in the class distribution. The analysis utilized the SMOTE technique to rectify this imbalance and generate synthetic samples for the minority class. Before applying SMOTE, the distribution was 637 Negative cases and 569 Positive cases. After using the SMOTE technique, the dataset achieved a balanced distribution, maintaining Negative cases at 637 and increasing Positive cases to match that number. The figures below illustrate the class distribution of the blood test dataset before and after the application of SMOTE.

Implementation of Machine Learning Models

Machine learning techniques are increasingly widely used in healthcare, particularly as data becomes more accessible. The study uses multiple algorithms to assess a dataset of blood tests for COVID-19 positive and negative cases.

Random Forest - Random Forest algorithm generates decision trees and chooses the result through a voting procedure [17]. It also reduces correlation among trees by randomly selecting features at each split.

LightGBM - Microsoft developed LightGBM in 2017, which is based on Gradient-Boosting Decision Trees [18]. While traditional GBDT iterates over the entire training dataset multiple times, LightGBM leverages a histogram-based technique and a leaf-wise growth strategy with a depth limit. This design enhances training efficiency and reduces memory usage.

Support Vector Machine - The SVM algorithm used for both classification and regression tasks and outlier detections. It discovers a hyperplane in an n-dimensional space that separates data points [19].

XGBoost - XGBoost created efficiency and excellent performance with large datasets [20]. This algorithm was introduced by Tianqi Chen and Carlos Guestrin in 2011 for classification and regression.

Logistic Regression - Logistic regression uses a statistical method sigmoid function to convert the input variables into probabilities. The binary variables are generally used to build a model [21].

Naive Bayes - This algorithm is designed to handle binary and multi-class problems using Bayes' theorem. It determines the likelihood that a sample is associated with a specific class [22].

RESULTS

Analyzing COVID-19 Blood Test Data with Machine Learning Algorithms

This analysis compares various traditional algorithms for blood biomarker data. The model's performance is calculated based on Accuracy, Precision, Recall, F1-score, and ROC AUC. The model achieved the highest scores. Random Forest, LightGBM, and Support Vector Machine, each with an accuracy of 0.82 achieved. The accuracy of predictive models depends on the volume and high-quality data [23]. Both regularized versions of Logistic Regression (Lasso and Ridge) performed similarly, achieving an accuracy of 0.82. These models are known for their

interpretability, which makes them useful in therapeutic contexts where understanding feature contributions is critical.

Performance Metrics - Receiver Operating Characteristic

The ROC curve helps assess performance. They provide insights into sensitivity (recall) and specificity across various thresholds. The random forest method predicts 0.89% of ROC AUC. Similarly, LightGBM achieves performance with an ROC AUC of 0.89, indicating its dependability in categorizing samples and selecting the dataset. XGBoost also performs well, with an ROC AUC of 0.89, proving its ability to predict class associations, primarily in complex datasets. The Lasso and Ridge regularized versions of Logistic Regression demonstrate commendable ROC AUC scores of 0.88. This determines its ability to manage feature contributions while efficiently retaining high classification accuracy. AdaBoost closely follows with an ROC AUC of 0.86, indicating high classification performance. The Multilayer Perceptron (MLP) shows an ROC AUC of 0.87, underscoring its effectiveness in handling complex data patterns, even though its recall is lower than that of a few other techniques. Meanwhile, the SVM achieves an ROC AUC of 0.86, indicating its competence in class separation. Although its performance is slightly lower than that of the top algorithms, it still demonstrates strong capability in binary classification tasks.

Construction of Stacking Ensemble Classifier

The stacking Ensemble learning technique enhances prediction accuracy by integrating outputs from multiple base models. This method employs a meta-classifier, trained on the predictions of base models, to combine effectively. The study utilizes a COVID-19 blood test dataset comprising 1,724 records and 35 features, including age, gender, WBC, RBC, ALT, and other clinical parameters. Various classification techniques are employed in this process.

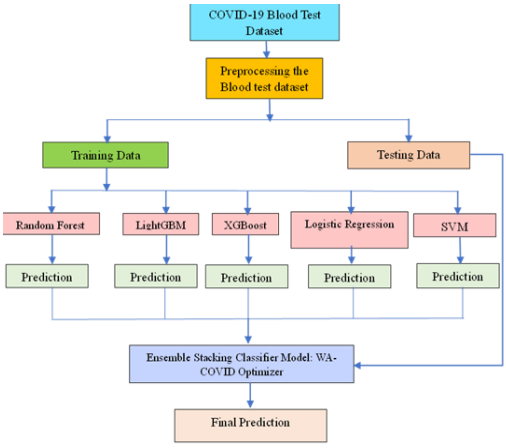

Fig.1 Workflow of proposed Stacking Ensemble Classifier

The second phase of the work introduces the WA-COVID Optimizer, a stacking-based ensemble that aims to improve classification model performance. This technique uses a meta-classifier to aggregate predictions from many base models efficiently. The optimizer uses 5-fold cross-validation to increase forecast accuracy and overall model reliability. The WA-COVID Optimizer achieved an accuracy of 0.84 and a ROC AUC of 0.90, showing superior classification. A Recall of 0.85 indicates the ability to recognize affirmative cases, while a Precision of 0.81 indicates consistent forecast reliability. An F1-Score of 0.83 suggests an appropriate balance of precision and recall. The optimizer outperforms other algorithms, particularly in ROC AUC, showcasing its robustness and effectiveness in differentiating between classes.

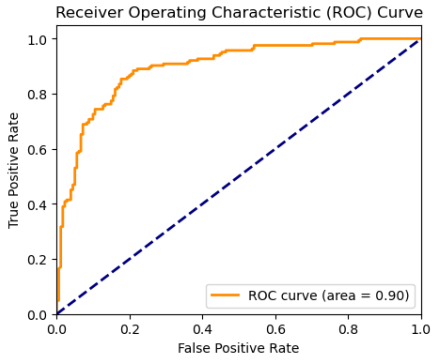

Fig.2 WA-COVID Optimizer ROC AUC value 0.90%

Matthews Correlation Coefficient (MCC)

The evaluation model was initially developed by B.W. Matthews in 1975 and later reintroduced by Baldi et al. in 2000 as a standard metric for evaluating machine learning performance. This coefficient extends to multiclass scenarios [24]. The ϕ coefficient is designed for 2×2 tables and is defined as:

$$MCC = \frac{TP.TN - FP.FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad [25]$$

In this analysis, the classification threshold is adjusted to 0.6 for categorizing samples. This means that if the predicted probability for a sample is equal to or greater than 0.6, it will be classified as a positive value (1); otherwise, it will be a negative value (0). The MCC is calculated using the actual labels and the original predictions. It is measured from -1 to 1, where 1 indicates perfect predictions, 0 indicates random predictions, and -1 signifies total disagreement between predictions and actual outcomes. The Matthews Correlation Coefficient (MCC) is calculated at 0.66 for the data. Based on the initial predictions, this score reflects a strong positive connection between projected and actual classes.

Pseudocode Implementation of WA-COVID Optimizer

Step 1: COVID-19 Blood Test Data Preprocessing
- Scale data using StandardScaler.
- Apply SMOTE to handle class imbalance and balance the dataset.

Step 2: Define Base Models
- Specify the base models for the first layer of the ensemble.

Step 3: Optimize Hyperparameters
- Optimize the hyperparameters of each base model before incorporating them into the WA-COVID Optimizer.

Step 4: WA-COVID Optimizer ensemble construction
- Define the base models as the first layer.
- Use a meta-model for final predictions in the second layer.

Step 5: Train the WA-COVID Optimizer
- Use the 5-fold cross-validation strategy.

Step 6: Threshold Adjustment
- Use a trained model to predict probabilities for the test set X_test.
- Adjust the classification threshold to optimize recall.

Step 7: Model Evaluation
- Evaluate the WA-COVID Optimizer using metrics.
- Generate visualizations

Step 8: Interpretation
- Compare metrics and visualizations to desired thresholds and interpret the results

Table:1 WA-COVID Optimizer Performance Metrics

Algorithm	Accuracy	Precision	Recall	F1-score	ROC AUC
WA-COVID Optimizer	0.84	0.81	0.85	0.83	0.90
Random Forest	0.82	0.83	0.79	0.81	0.89
LightGBM	0.82	0.84	0.77	0.81	0.89
Support Vector Machine	0.82	0.82	0.80	0.81	0.86
Logistic Regression	0.81	0.79	0.80	0.80	0.86

This section presents a performance comparisons of different algorithms, including a stacking classifier approach WA-COVID Optimizer, using classification metrics. The stacking classifier WA-COVID Optimizer demonstrates performance metrics, achieving 84% as accuracy, a precision - 81%, the recall value is 85%, F1-score 83%, and ROC AUC - 90% for the COVID-19 blood test dataset.

DISCUSSION

This study investigated a dataset of blood test results from 814 COVID-19-positive patients by developing an ensemble stacking classifier. Several important prognostic markers were identified by our feature analysis, including LDH (Lactate Dehydrogenase), WBC (White Blood Cell Count), AST (Aspartate Aminotransferase), CA (Calcium), EOT (Eosinophil Count), RBC (Red Blood Cell Count), ALT (Alanine Aminotransferase), ALP (Alkaline Phosphatase), and HCT (Hematocrit). This study developed a novel approach for predicting outcomes in COVID-19 patients through the integration of various classification techniques: Random Forest, LightGBM, Support Vector Machine, Logistic Regression (with Lasso and Ridge regularization), XGBoost, AdaBoost, Multilayer Perceptron, Gradient Boosting Machine, K-Nearest Neighbors, Naive Bayes, DNN, and ANN.

We developed a new data-driven strategy, the WA-COVID Optimizer, which synergizes multiple base models with a boosting mechanism. The proposed stacking classifier, WA-COVID Optimizer, predicted the best accuracy of 84% and a ROC AUC -90% for COVID-19-positive cases. The MCC validates the classifier performance, and the evaluation score is 66%. The findings suggest that COVID-19 patients continue with appropriate treatment to reduce the risk of severe illness and prevent the spread within the community. Our study does have some limitations. The number of patients analyzed was relatively small, and future research would benefit from a larger, preferably multinational dataset to enhance the findings. Despite its limitations, the Stacking Ensemble Classifier can reliably predict outcomes in COVID-19 patients and is a valuable tool for identifying significant drivers and prognostic signals.

REFERENCES

[1] Sheetal Rajpal, Manoj Agarwal, Ankit Rajpal, et al. (2022), Cov-elm classifier: an extreme learning machine based identification of covid-19 using chest x-ray images. arXiv,doi.org/10.48550/arXiv.2007.08637

[2] Countries where Coronavirus has spread Worldometer. https://www.worldometers.info/coronavirus/countries-where-coronavirus-has-spread/. Accessed 05 Oct, 2024.

[3] Kenneth McIntosh et al (2020). Coronavirus disease 2019 (COVID-19). *UpToDate Hirsch MS Bloom.* (2020) 5:1–1.

[4] WHO Announces Simple Easy-to-Say Labels for SARS-CoV-2 Variants of Interest and Concern. Available online at: https://www.who.int/news/item/31-05-2021-who-announces-simple-easy-to-say-labels-for-sars-cov-2-variants-of-interest-and-concern (accessed October 07, 2024).

[5] Gan, J.M., Kho, J., Akhunbay-Fudge, M. *et al.* (2021). Atypical presentation of COVID-19 in hospitalised older adults. *Ir J Med Sci* 190, 469–474. doi:10.1007/s11845-020-02372-7

[6] Sarwan Ali, Bikram Sahoo, Naimat Ullah, et al. (2021) A k-mer Based Approach for SARS-CoV-2 Variant Identification, Bioinformatics Research and Applications, doi: 10.1007/978-3-030-91415-8_14.

[7] Sethuraman N, Jeremiah SS, Ryo A. (2020). Interpreting Diagnostic Tests for SARS-CoV-2. JAMA. 323(22):2249-2251. doi: 10.1001/jama.2020.8259. PMID: 32374370.

[8] CDC:Interim Guidelines for Collecting and Handling of Clinical Specimens for COVID-19 Testing. 2024.

[9] Mei XY, Lee HC, Diao KY, et al. (2020). Artificial intelligence-enabled rapid diagnosis of patients with COVID-19. Nat Med 2020; 26:1224–8. doi: 10.1038/s41591-020-0931-3.

[10] Yan, L., Zhang, HT., Goncalves, J. *et al.*(2020). An interpretable mortality prediction model for COVID-19 patients. *Nat Mach Intell* 2, 283–288. doi:10.1038/s42256-020-0180-7.

[11] J. Bao, C. Li, K. Zhang, H. Kang, et al. (2020). Comparative analysis of laboratory indexes of severe and non-severe patients infected with COVID-19. Clin Chim Acta. 509:180-194. doi: 10.1016/j.cca.2020.06.009.

[12] D. Brinati, A. Campagner, D. Ferrari, et al. (2020). Detection of COVID-19 Infection from Routine Blood Exams with Machine Learning: A Feasibility Study. J Med Syst.44(8):135. doi: 10.1007/s10916-020-01597-4.

[13] Booth, A.L., Abels, E. & McCaffrey, P. (2021). Development of a prognostic model for mortality in COVID-19 infection using machine learning. Mod Pathol 34, 522–531. doi.org/10.1038/s41379-020-00700-x

[14] Karlupia N, Abrol P (2023). Wrapper-based optimized feature selection using nature-inspired algorithms. Neural Comput Appl 35:12675–12689

[15] Fernandes, F.T., de Oliveira, T.A., Teixeira, C.E. *et al.*(2021). A multipurpose machine learning approach to predict COVID-19 negative prognosis in São Paulo, Brazil. *Sci Rep* **11**, 3343. doi.org/10.1038/s41598-021-82885-y

[16] He, Y., Lu, X., Fournier-Viger, P. *et al.*(2024). A novel overlapping minimization SMOTE algorithm for imbalanced classification. *Front Inform Technol Electron Eng* 25, 1266–1281 doi:10.1631/FITEE.2300278

[17] Cornelius E, Akman O, Hrozencik D. (2021). COVID-19 mortality prediction using machine learning-integrated Random Forest Algorithm under varying patient Frailty. Mathematics;9(17):2043.

[18] Ke, G., Meng, Q., Finley, T., et al. (2017). Lightgbm: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30.

[19] Singh V, Poonia RC, Kumar S, et al. (2020). Prediction of COVID-19 corona virus pandemic based on time series data using support vector machine. J Discrete Math Sci Crypt.;23(8):1583–97.

[20] Luo J, Zhang Z, Fu Y, Rao F. (2021). Time series prediction of COVID-19 transmission in America using LSTM and XGBoost algorithms. Results in Physics. 27:104462.

[21] Josephus BO, Nawir AH, Wijaya E, (2021). Predict mortality in patients infected with COVID-19 Virus based on observed characteristics of the patient using logistic regression. Procedia Comput Sci. 179:871–7.

[22] Karaismailoglu E, Karaismailoglu S. (2021). Two novel nomograms for predicting the risk of hospitalization or mortality due to COVID-19 by the naïve bayesian classifier method. J Med Virol. 93(5):3194–201.

[23] Yury V. Kistenev, Denis A. Vrazhnov, et al (2022) Zuhayri,Predictive models for COVID-19 detection using routine blood tests and machine learning,Heliyon,Volume 8, Issue 10, e11185,doi:10.1016/j.heliyon.2022.e11185

[24] Chicco, D., Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 21. doi:10.1186/s12864-019-6413-7

[25] D. Chicco, V. Starovoitov and G. Jurman, (2021). The Benefits of the Matthews Correlation Coefficient (MCC) Over the Diagnostic Odds Ratio (DOR) in Binary Classification Assessment, in *IEEE Access*, vol. 9, pp. 47112-47124, doi: 10.1109/ACCESS.2021.3068614.