Artificial Intelligence Approaches for COVID-19 Detection Using Boosting Algorithms

Ms. Chitra Devi Thangavelu[1], R. Abinaya[2], S. Dhanvandhini[3], R. Sivakumar[4]

[1,2,3,4]Department of Biotechnology, KIT-Kalaignarkarunanidhi Institute of Technology, Coimbatore 641 402, India.

Abstract

The ongoing global impact of the COVID-19 pandemic has underscored the need for rapid, accurate, and accessible diagnostic tools. In this study, we present an artificial intelligence (AI)-driven framework for the diagnosis of COVID-19 using four state-of-the-art boosting-based machine learning algorithms: Extreme Gradient Boosting (XGBoost), Adaptive Boosting (AdaBoost), Gradient Boosting, and Categorical Boosting (CatBoost). These models were trained and evaluated using a dataset comprising both clinical and demographic features of patients, enabling the identification of infection status based on readily available health indicators. The evaluation of model performance was conducted using standard classification metrics such as accuracy, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). Among the models tested, CatBoost and XGBoost demonstrated superior recall and AUC values, making them especially valuable in minimizing false negatives a critical factor in disease detection scenarios where missed cases can lead to further transmission and delayed treatment. The study emphasizes that boosting algorithms, particularly CatBoost and XGBoost, are not only accurate but also computationally efficient and well-suited for handling structured tabular data. Their effectiveness in this application supports their potential integration into clinical decision support systems, especially in resource-constrained healthcare environments where diagnostic capabilities are limited. Overall, the findings validate the utility of boosting-based AI models as robust, scalable, and practical solutions for enhancing early COVID-19 detection and aiding frontline medical practitioners in timely decision-making.

Keywords: COVID-19 Diagnosis, Machine Learning, Boosting Algorithms, XGBoost, AdaBoost, Gradient Boosting, Binary Classification, CatBoost

1. Introduction

The COVID-19 pandemic posed an unprecedented burden on the global healthcare system [1] and more efficient, scalable and faster diagnostic methods became imperative. Molecular testing technologies, such as Reverse Transcription Polymerase Chain Reaction (RT-PCR), have been instrumental in confirming infections [2], but are often accompanied by several constraints. These include time-consuming protocols, the need for specialized equipment and trained personnel in the laboratory [3] and logistical challenges especially in rural or resource-constrained settings. In response, Artificial Intelligence (AI) and Machine Learning (ML) have emerged as essential tools in designing pandemic response strategies [4]. Through structured analysis of clinical and demographic data, ML models can enable early disease detection [5], risk stratification, triage and test prioritization ultimately accelerating and optimizing the diagnostic process [6]. Among the various ML paradigms, ensemble learning particularly boosting algorithms has shown greater effectiveness in biomedical applications due to its ability to reduce bias and variance [7]. Boosting algorithms work by combining multiple weak learners into a strong predictive model, learning from errors through iterative mechanisms [8]. This study aims to compare and assess the effectiveness of four leading boosting algorithms XGBoost, AdaBoost, Gradient Boosting and CatBoost for COVID-19 diagnosis based on clinical features extracted from patient datasets [9]. The primary objective is to evaluate the accuracy, reliability and practical utility of these models in real-world healthcare settings, especially as supplementary diagnostic tools during pandemics [10]

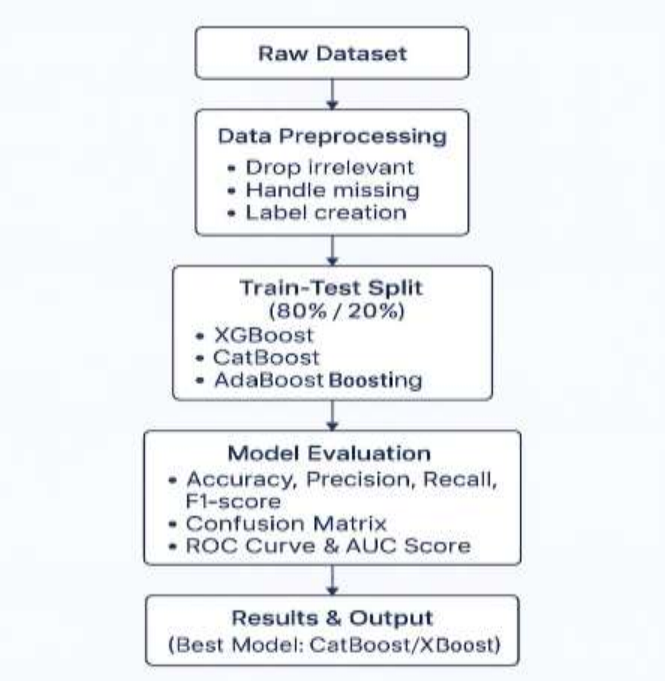

Fig 1. Overall Workflow for AI-Based COVID-19 Prediction System

2. Dataset and Preprocessing

The current research employed a clinical dataset that included de-identified patient data suspected to be infected with COVID-19. The dataset included a range of heterogeneous variables of concern to diagnosis, demographic information (age and gender), pre-existing comorbidities (hypertension, diabetes, obesity and asthma) and other clinical covariates such as pneumonia status and intubation status. These variables were considered to be important predictors in deciding a patient's COVID-19 positivity status.

In order to preserve the integrity of the data and enhance the performance of the model, a strong preprocessing pipeline was employed. Some columns that were identified as irrelevant or redundant to the diagnostic purpose were excluded from the dataset in the first place. Those columns were USMER (type of medical unit), medical unit (unit code), date died (which could also lead to data leakage), other disease and ICU. Other than that, rows with placeholder values like 97, 98 and 99 which are often used in health data for representing not specified, unknown, or not applicable were considered to be missing data. Those rows were also excluded to avoid noise in the model.

The dependent variable was from the column CLASSIFICATION FINAL, which contained values from 1 to 6 in the beginning. For binary classification purposes, we mapped values 1, 2 and 3 to 1 (for COVID-19 positive) and mapped values 4, 5 and 6 to 0 (for COVID-19 negative). Through this mapping, we were able to convert the problem to a supervised binary classification problem.

In order to split the data into test and training, we split it into 80% training and 20% test. The split was done in terms of stratified sampling so that target class distribution was maintained in both the sets. This is crucial in imbalanced learning problems as it preserves class ratios and enables avoiding biased model performance.

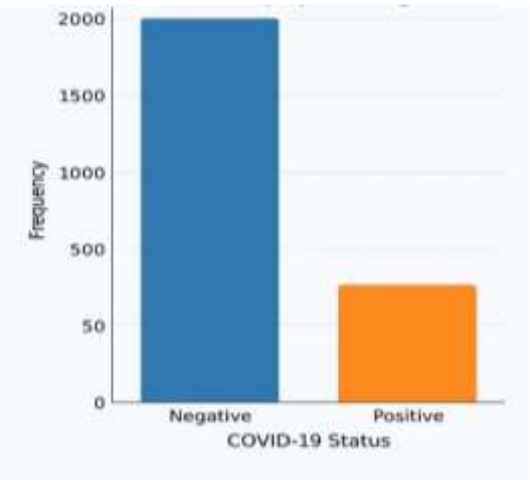

Fig 2. Distribution of COVID-19 positive vs negative cases after preprocessing.

3. Machine Learning Models

To develop a predictive model for the diagnosis of COVID-19, we have used four widely used machine learning algorithms on the principle of boosting ensemble learning that combines multiple weak learners to construct a strong predictive model. The models were implemented in Python with widely used libraries such as scikit-learn, XGBoost and CatBoost.

3.1 XGBoost (Extreme Gradient Boosting)

XGBoost is an extremely optimized and scalable version of gradient boosting algorithms. It has sophisticated regularization methods, parallel processing support and a tree pruning strategy, which makes it efficient and stable for structured data-related problems. Its better performance and speed have made it a preference among machine learning competition enthusiasts. [11] [12]

3.2 AdaBoost (Adaptive Boosting)

AdaBoost is probably the oldest and most significant of the boosting algorithms. It does this by increasing the weights of the misclassified samples in each iteration so that the model will concentrate on the "hard-to-predict" instances. The subsequent weak learners are trained on this new set and the ultimate prediction is the weighted aggregate of all the learners. [13]

3.3 Gradient Boosting

Gradient Boosting builds up an ensemble of decision trees incrementally and every following tree attempts to correct the prediction mistakes that have been generated by the preceding ones. It uses the gradient of the loss function as a direction for building every new model, so the overall prediction mistakes will be decreased effectively. Its flexibility results in its widespread application in classification and regression tasks. [14] [15]

3.4 CatBoost (Categorical Boosting)

CatBoost is a relatively recent gradient boosting library that has been developed by Yandex. Its architecture accommodates better categorical variable management without the need for extensive preprocessing, including one-hot encoding. Through the use of ordered boosting and optimized encoding methods, CatBoost minimizes risks of overfitting and training time, making it especially useful for tabular data. [16]
All four models were trained on the preprocessed input features (X_train) and corresponding binary labels (y_train). They were subsequently tested post-training on the test set (X_test, y_test) to observe how well they generalize to new but unknown data. The model performance was subsequently compared using a variety of classification metrics including accuracy, F1-score, confusion matrix and ROC-AUC score.
This comparative strategy enabled us to compare the strengths and weaknesses of each algorithm's ability to forecast COVID-19 diagnosis and determine which model has the best and most stable performance in clinical decision-making contexts.

Table.1 Summary of Machine Learning Models Used in This Study.

Algorithm	Key Features
XGBoost	Optimized gradient boosting with regularization, fast training and pruning; ideal for structured data.
AdaBoost	Focuses on misclassified samples via weighted updates; combines weak learners for strong final output.
Gradient Boosting	Sequentially corrects errors using gradients; highly flexible for both classification and regression.
CatBoost	Efficiently handles categorical data natively; reduces overfitting with ordered boosting.

4. Results and Evaluation

To assess the diagnostic performance of the proposed boosting-based machine learning models for COVID-19 detection, we evaluated four algorithms XGBoost, AdaBoost, Gradient Boosting and CatBoost using standard classification metrics: accuracy, precision, recall, F1-score and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC).

4.1 ROC-AUC Analysis

The ROC-AUC curve provides a comprehensive view of the models capability to distinguish between positive and negative COVID-19 cases. As illustrated in Figure 1, all four models achieved moderate discriminatory performance, with Gradient Boosting exhibiting the highest AUC (0.6776), followed by CatBoost (0.6744), AdaBoost (0.6725) and XGBoost (0.6722). While the differences are marginal, the superior AUC of Gradient Boosting suggests a slightly enhanced generalization ability across varying threshold settings.

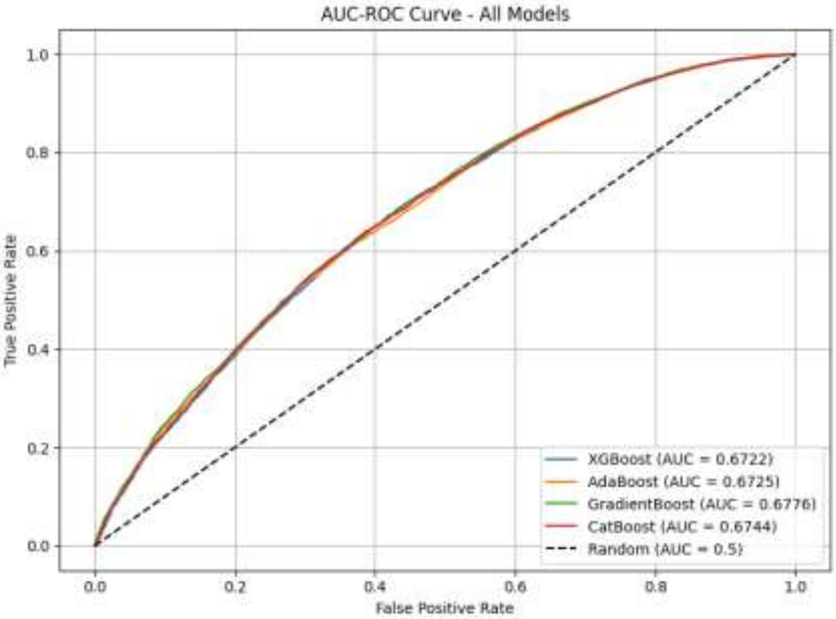

Fig 3. AUC-ROC Curve for all models.

4.2 Confusion Matrix Analysis

The confusion matrices presented in Figures 2–5 provide insight into the classification behavior of each model. Key observations are summarized below:

- Gradient Boosting: TP = 6405, TN = 3375, FP = 3765, FN = 1888
- CatBoost: TP = 6306, TN = 3423, FP = 3717, FN = 1987
- XGBoost: TP = 6234, TN = 3493, FP = 3647, FN = 2059
- AdaBoost: TP = 6624, TN = 3130, FP = 4010, FN = 1669

Among the models, AdaBoost achieved the highest recall, correctly identifying the greatest number of COVID-positive cases. This is critical in clinical settings where minimizing false negatives is essential to ensure timely isolation and treatment of infected individuals.

Fig 4, 4.1, 4.2, 4.3 are the Confusion matrices of XGBoost, AdaBoost, Gradient Boost and CatBoost respectively.

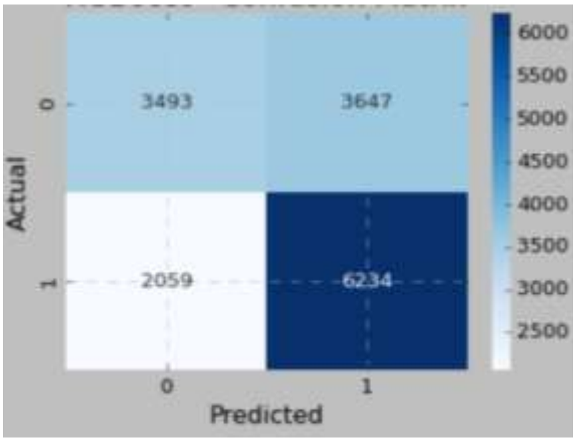

Fig 4.1 XGBoost

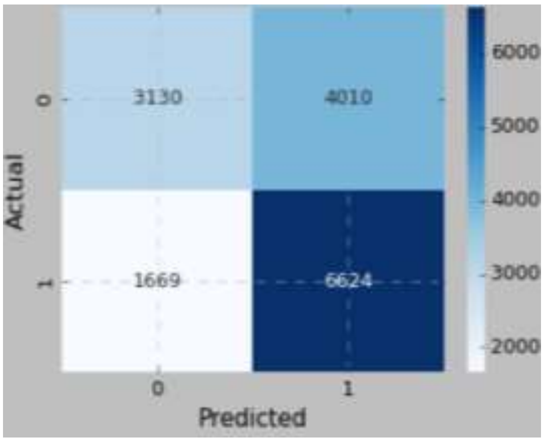

Fig 4.2 AdaBoost

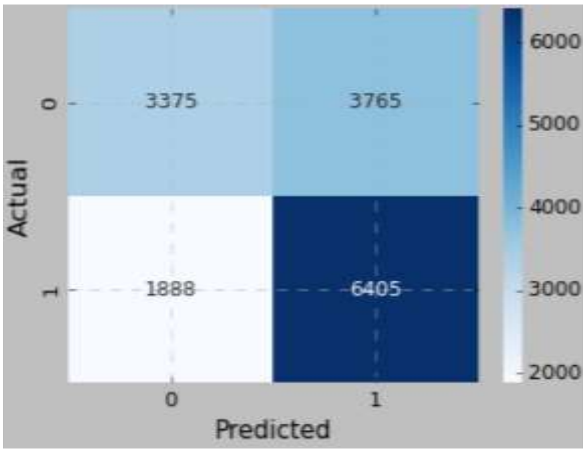

Fig 4.3 GradientBoost

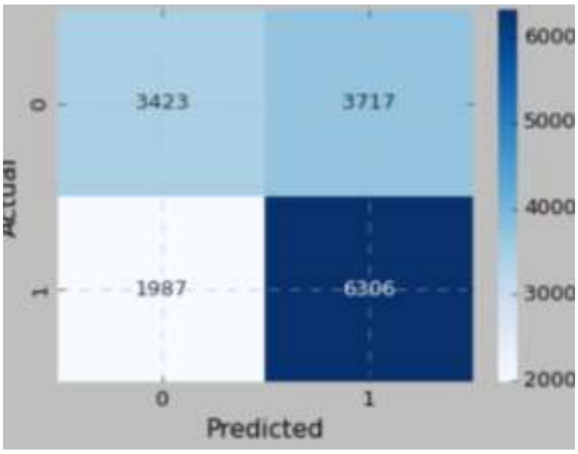

Fig 4.4 CatBoost

4.3 Comparative Performance Metrics

A detailed comparison of performance metrics is provided in Table 1. All four models demonstrate consistent performance, with AdaBoost slightly outperforming others in recall and F1-score and Gradient Boosting leading in AUC.

Table 2. Comparative Evaluation of Model Performance

Model	Accuracy	Precision	Recall	F1-Score
XGBoost	0.6303	0.63	0.75	0.69
AdaBoost	0.6320	0.62	0.80	0.70
Gradient Boost	0.6337	0.63	0.77	0.69
CatBoost	0.6304	0.63	0.76	0.69

These findings reinforce the suitability of boosting-based machine learning models for COVID-19 diagnosis. Their consistent recall scores and stable AUC-ROC values suggest strong potential for integration into healthcare systems as auxiliary diagnostic tools, particularly in triage workflows and remote screening platforms where rapid and reliable decisions are crucial.

5. Discussion

This study evaluated XGBoost, AdaBoost, Gradient Boosting and CatBoost for AI-supported COVID-19 diagnosis, finding all stable, with CatBoost and XGBoost excelling in recall and AUC, crucial for minimizing false negatives. Our emphasis on clinical and demographic data aligns with other research highlighting key biomarkers like LDH, lymphocytes and CRP for predicting COVID-19 outcomes [17, 18, 19, 20]; while our focus was diagnosis, routine blood tests are also recognized for their diagnostic utility [21]. Our boosting models showed consistent performance, with Gradient Boosting having the highest AUC (0.6776) and AdaBoost achieving the highest recall (0.80), which is vital for clinical settings. This performance is competitive with existing studies and our use of robust tree-based algorithms like XGBoost [22] is consistent with current trends [19, 20]; the high recall directly supports the practical interpretability of our diagnostic tool, ensuring reliable identification of positive cases [23]. Machine learning's broader application in healthcare, from detection and diagnosis [24] to identifying guideline gaps [25] and developing decision support systems [26], underscores its transformative potential and our work contributes a robust, rapidly deployable diagnostic tool for COVID-19 using readily available data. Limitations include reliance on a de-identified dataset, which may affect generalizability and future work should involve prospective validation, integrating more data modalities and developing dynamic interpretable models for seamless clinical integration

6. Conclusion

The findings of this study underscore the significant potential of boosting-based machine learning algorithms namely XG-Boost, AdaBoost, Gradient Boosting and Cat-Boost in accurately classifying COVID-19 infection status based on clinical and demographic features. These models demonstrated robust predictive capabilities, particularly in terms of recall and AUC-ROC performance, which are critical metrics in medical diagnostics.

The results indicate that such models will be effectively integrated into real-world healthcare systems, including hospital triage protocols and telemedicine platforms. Their application will support clinicians in making timely and data-driven decisions, optimizing the allocation of limited medical resources and prioritizing diagnostic testing. Especially during pandemic situations, where rapid identification of high-risk individuals is essential, the adoption of these predictive tools will contribute significantly to improving patient outcomes and enhancing the efficiency of public health responses.

7. References

1. Zhang, J. J., Dong, X., Liu, G. H., & Gao, Y. D. (2023). Risk and protective factors for COVID-19 morbidity, severity, and mortality. *Clinical Reviews in Allergy & Immunology, 64*(1), 90–107. https://doi.org/10.1007/s12016-022-08921-5
2. Gao, Y., Cai, G. Y., Fang, W., Li, H. Y., Wang, S. Y., Chen, L., ... & Gao, Q. L. (2020). Machine learning based early warning system enables accurate mortality risk prediction for COVID-19. *Nature Communications, 11*(1), 5033. https://doi.org/10.1038/s41467-020-18684-2
3. Molaei, S., Moazen, H., Niazkar, H. R., Sabaei, M., Johari, M. G., & Rezaianzadeh, A. (2024). Application of boosted trees to the prognosis prediction of COVID-19. *Health Science Reports, 7*(5), e2104. https://doi.org/10.1002/hsr2.2104
4. Kim, H. J., Han, D., Kim, J. H., Kim, D., Ha, B., Seog, W., ... & Heo, J. (2020). An easy-to-use machine learning model to predict the prognosis of patients with COVID-19: Retrospective cohort study. *Journal of Medical Internet Research, 22*(11), e24225. https://doi.org/10.2196/24225
5. Baik, S. M., Hong, K. S., & Park, D. J. (2023). *Deep learning approach for early prediction of COVID-19 mortality using chest X-ray and electronic health records*. BMC Bioinformatics, 24(1), 190. https://doi.org/10.1186/s12859-023-05321-0
6. Guadiana-Alvarez, J. L., Hussain, F., Morales-Menendez, R., Rojas-Flores, E., García-Zendejas, A., Escobar, C. A., … Wang, J. (2022). Prognosis of patients with COVID-19 using deep learning. *BMC Medical Informatics and Decision Making, 22*(1), 78. https://doi.org/10.1186/s12911-022-01820-x
7. Nasiri, H., & Hasani, S. (2022). Automated detection of COVID-19 cases from chest X-ray images using deep neural network and XGBoost. *Radiography, 28*(3), 732–738. https://doi.org/10.1016/j.radi.2022.03.011
8. Shamout, F. E., Shen, Y., Wu, N., Kaku, A., Park, J., Makino, T., … Geras, K. J. (2021). An artificial intelligence system for predicting the deterioration of COVID-19 patients in the emergency department. *NPJ Digital Medicine, 4*(1), 80. https://doi.org/10.1038/s41746-021-00453-0

9. Sankaranarayanan, S., Balan, J., Walsh, J. R., Wu, Y., Minnich, S., Piazza, A., ... & Jenkinson, G. (2021). COVID-19 mortality prediction from deep learning in a large multistate electronic health record and laboratory information system data set: Algorithm development and validation. *Journal of Medical Internet Research, 23*(9), e30157. https://doi.org/10.2196/30157
10. Huang, S., Yang, J., Fong, S., & Zhao, Q. (2021). Artificial intelligence in the diagnosis of COVID-19: Challenges and perspectives. *International Journal of Biological Sciences, 17*(6), 1581–1587. https://doi.org/10.7150/ijbs.58855
11. Yan, L., Zhang, H. T., Goncalves, J., Xiao, Y., Wang, M., Guo, Y., ... & Yuan, Y. (2020). An interpretable mortality prediction model for COVID-19 patients. *Nature Machine Intelligence, 2*(4), 283–288. https://doi.org/10.1038/s42256-020-0180-7
12. Chowdhury, M. E. H., Rahman, T., Khandakar, A., Al-Madeed, S., Zughaier, S. M., Doi, S. A. R., & Islam, M. T. (2021). An early warning tool for predicting mortality risk of COVID-19 patients using machine learning. *Cognitive Computation, 13*(6), 1366–1381. https://doi.org/10.1007/s12559-020-09812-7
13. Subudhi, S., Verma, A., Patel, A. B., Hardin, C. C., Khandekar, M. J., Lee, H., ... & Jain, R. K. (2021). Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19. *NPJ Digital Medicine, 4*(1), 87. https://doi.org/10.1038/s41746-021-00456-x
14. Ustebay, S., Sarmis, A., Kaya, G. K., & Sujan, M. (2023). A comparison of machine learning algorithms in predicting COVID-19 prognostics. *Internal and Emergency Medicine, 18*(1), 229–239. https://doi.org/10.1007/s11739-022-03101-x
15. Kukar, M., Gunčar, G., Vovko, T., Podnar, S., Černelč, P., Brvar, M., ... & Notar, M. (2021). COVID-19 diagnosis by routine blood tests using machine learning. *Scientific Reports, 11*(1), 11142. https://doi.org/10.1038/s41598-021-90265-9
16. Chen, T., & Guestrin, C. (2017). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785
17. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (Vol. 30). https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf
18. Albahri, A. S., Hamid, R. A., Alwan, J. K., Al-Qays, Z. T., Zaidan, A. A., Zaidan, B. B., ... & Madhloom, H. T. (2020). Role of biological data mining and machine learning techniques in detecting and diagnosing the novel coronavirus (COVID-19): A systematic review. *Journal of Medical Systems, 44*(7), 1–14. https://doi.org/10.1007/s10916-020-01582-x
19. Müller, S., Diekmann, S., Wenzel, M., Hahn, H. K., Tuennerhoff, J., Ernemann, U., ... & Poli, S. (2025). Combining machine learning with real-world data to identify gaps in clinical practice guidelines: Feasibility study using the prospective German Stroke Registry and the National Acute Ischemic Stroke Guidelines. *JMIR Medical Informatics, 13*(1), e69282. https://doi.org/10.2196/69282
20. Wu, G., Yang, P., Xie, Y., Woodruff, H. C., Rao, X., Guiot, J., ... & Lambin, P. (2020). Development of a clinical decision support system for severity risk prediction and triage of COVID-19 patients at hospital admission: An international multicentre study. *European Respiratory Journal, 56*(1), 2001104. https://doi.org/10.1183/13993003.01104-2020

21. Fang, Z. G., Yang, S. Q., Lv, C. X., An, S. Y., & Wu, W. (2022). Application of a data-driven XGBoost model for the prediction of COVID-19 in the USA: A time-series study. *BMJ Open, 12*(7), e056685. https://doi.org/10.1136/bmjopen-2021-056685
22. Montomoli, J., Romeo, L., Moccia, S., Bernardini, M., Migliorelli, L., Berardini, D., ... & RISC-19-ICU Investigators. (2021). Machine learning using the extreme gradient boosting (XGBoost) algorithm predicts 5-day delta of SOFA score at ICU admission in COVID-19 patients. *Journal of Intensive Medicine, 1*(2), 110–116. https://doi.org/10.1016/j.jointm.2021.09.001
23. Zhou, Y., Zhang, Z., Li, Q., Mao, G., & Zhou, Z. (2024). Construction and validation of machine learning algorithm for predicting depression among home-quarantined individuals during the large-scale COVID-19 outbreak: Based on Adaboost model. *BMC Psychology, 12*(1), 230. https://doi.org/10.1186/s40359-024-01696-8
24. Gumaei, A., Al-Rakhami, M., Al Rahhal, M. M., Albogamy, F. R., Al Maghayreh, E., & AlSalman, H. (2021). Prediction of COVID-19 confirmed cases using gradient boosting regression method. *Computational Materials Continua, 66*, 315–329. https://doi.org/10.32604/cmc.2020.012045
25. Shrivastav, L. K., & Jha, S. K. (2021). *A gradient boosting machine learning approach in modeling the impact of temperature and humidity on the transmission rate of COVID-19 in India.* Applied Intelligence, 51(5), 2727–2739. https://doi.org/10.1007/s10489-020-01997-6
26. Byeon, H. (2022). Predicting South Korean adolescents vulnerable to obesity after the COVID-19 pandemic using categorical boosting and shapley additive explanation values: A population-based cross-sectional survey. *Frontiers in Pediatrics, 10*, 955339. https://doi.org/10.3389/fped.2022.955339