

A machine learning model for predicting recovery rates of COVID-19 patients

Amani Abdo^{1,2}, Kholoud Mohamed Elzalama², Ahmed Elsayed Yakoub²

¹Faculty of Computing, Arab Open University, Cairo, Egypt

²Faculty of Computers and Artificial Intelligence, Helwan University, Cairo, Egypt

Article Info

Article history:

Received Jan 29, 2023

Revised May 17, 2023

Accepted May 19, 2023

Keywords:

Classification

COVID-19

Ensemble method

Prediction

Recovery rate

SARS-COV-2

ABSTRACT

During disease epidemics, any trial to improve healthcare systems entails preserving lives. Therefore, predicting which patients are at high risk becomes critical and challenging when confronted with a novel virus. The recent COVID-19 changed many people's perspectives on how to approach diseases. According to the lack of medical resources, it is important to identify the patients who need instant medical care. This research proposes a machine learning model to identify high-risk patients that require specific medical attention. Specifically, extreme gradient boosting (XGboost), random forest (RF), and logistic regression (LR) are used in the ensemble method to classify COVID-19 patients at high risk. The dataset consists of 361 medical records for severe COVID-19 patients which have included 195 survivors. The most correlated features (neutrophils (%), hypersensitive c-reactive protein, lactate dehydrogenase, age, procalcitonin, and neutrophils count) are selected to be used in classification. Different machine learning classifiers are applied to the mentioned dataset to find out the optimum classifiers to be used in the ensemble method. 98% is the most optimal accuracy achieved with the proposed model.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Kholoud Mohamed Elzalama

Faculty of Computers and Artificial Intelligence, Helwan University

Cairo, Egypt

Email: KholoudElzalama@gmail.com

1. INTRODUCTION

The epidemic of COVID-19 began in Wuhan, China [1], then spread to the rest of the world. COVID-19, an infectious disease, was designated a Public Health Emergency of International Concern on January 30, 2020, by the World Health Organization (WHO) [2]–[5]. WHO classified the viral outbreak a pandemic on March 11, 2020 [3], after it expanded to 200 nations [6]. On 22 November 2022, WHO received reports of 635,229,101 confirmed cases of COVID-19, including 6,602,552 fatalities which affected the global economy. Hospitals throughout the world struggle to care for a growing number of patients who need to be hospitalized. This calls for methods to prioritize patients and assess illness severity [7]. The COVID-19 pandemic has raised awareness about the dangers of developing zoonotic diseases and sparked broad concern and interest in taking action to avoid future pandemics [8]. At the beginning of any new disease or pandemic, there is usually no prognostic biomarker available to differentiate people who require prompt medical intervention and their related fatality rate. Given the shortage of medical resources, it's essential to provide priority to patients who need immediate medical attention.

Machine learning for this pandemic is not well defined because the COVID-19 is a relatively new field of study. Machine learning is used in a variety of disciplines such as computer vision, speech recognition, bio-surveillance, and so on [9]. As a result, we should utilize machine learning with COVID-19 medical

information. Therefore It will help us to extract insights that can be referenced to predict any future pandemic. For this reason, a number of studies on COVID-19 have recently been published. Hamzah *et al.* [4], Hu *et al.* [10], some of these studies attempted to predict the COVID-19 outbreak and the end of this pandemic. Hu *et al.* [10] proposed an artificial intelligence forecasting model for the COVID-19 outbreak in China. The size, lengths, and ending time of Coronavirus are estimated by a modified stacked autoencoder and clustering. The used data is from surging news network and WHO. The estimated average error was between 0.73% and 2.27%. Hamzah *et al.* [4] used the susceptible-exposed-infected removed (SEIR) model to predict the COVID-19 outbreak and evaluate the political and economic impact of it by analyzing the COVID-19 latest news. They discovered that negative articles outnumber positive ones. Cooperation and individual strength in the face of this pandemic are the content of the top five positive articles, whereas the top five negative articles are uncertainty and bad virus outcomes such as mortality. They applied their model to data from John Hopkins University, WHO, and Ding Xiang Yuan. Other studies analyze patient information to determine which features could predict recovery or fatality rates. In order to save human lives. Al-Rousan and Al-Najjar [11] used statistical analysis and the χ^2 test to uncover the correlation between the dependent and independent variables of the COVID-19 data. Data was collected from Korean centers for disease control. The findings revealed that sex, area, birth year, and infection causes are the most correlated features to the target variables. DeCapprio *et al.* [12] developed three models that reflect a balance between the convenience of implementation, and accuracy. The goal of these models is to detect COVID-19 patients at high risk. Logistic regression, gradient boosted trees (all features), and gradient boosted trees (diagnosis+age) are used to build the three models. The AUC of the logistic regression, xgboost diagnosis history+age, and XGBoost full feature set are .731, .810, and .810 respectively. Moulai *et al.* [13] proposed a machine learning model to forecast the mortality rate. The random forest algorithm performed better than other classifiers used in this research with 95.03% accuracy. The model was applied to data collected from the registry of Ayatollah Taleghani Hospital. Feng *et al.* [14] proposed diagnosis aid system based on the Lasso regression model using medical information gathered from fever clinics. The Lasso feature selection is used to select the most relevant features. 84.1% is the accuracy of the model performance. To predict the survival of COVID-19 patients, a machine learning model built on the XGboost algorithm was presented by Yan *et al.* [15], [16]. Blood samples from COVID-19 patients were obtained from Wuhan, China's Tongji Hospital, and the model was used for those samples. Around 97% of the model's performance is accurate.

The healthcare business is enormous and necessitates the collecting and processing of medical data in real-time [17]. Additionally, the difficulty of data processing is at the core of this industry and requires real-time prediction and information dissemination to practitioners in order to give fast medical action [17]. However, there is an increasing necessity to continue analyzing COVID-19 data because COVID-19 medical information is not available, and if it is, it is confidential. Therefore, all of the previous studies are based on sparse data.

Due to the preciousness of our human life, we aim to enhance the recovery rate prediction of COVID-19 patients even though medical information is insufficient. Therefore, we propose a machine learning model to predict the recovery rate of patients and determine which patients need to be hospitalized the most. We aim to achieve the best results possible to classify COVID-19 patients' recovery. Classification in machine learning refers to a predictive modeling task in which a class label is determined for a given sample of data [18]. The training dataset is used by the classification model to assign samples to specified classes. Feature selection is an important approach to figure out the most correlated features to the class label [19], [20]. Consequently, classification accuracy will be improved and execution time will be decreased [20], [21]. In this research paper, we stacked the feature selection approaches with Machine learning classification techniques to predict the recovery rate of a COVID-19 patient.

This research paper is divided into 3 main sections. In section 2 contains the proposed model phases and how they are developed. In section 3 contains a discussion of the model results and some attempts were performed to achieve the best results. In section 4 contains a brief of the research paper.

2. METHOD

In this research, a machine learning model is employed to estimate the probability of COVID-19 patient recovery. In this section, the proposed model is discussed. As shown in Figure 1 the model is divided into three stages. The first stage is data preprocessing. In this stage, the missing values are handled, next, the most relevant features to the target value are chosen, and finally, all features are scaled. The second stage is data splitting to validate our model in 2 ways. The third stage is classification, which involves training the ML classifier with the training dataset and finally, determine the recovery rate for a new patient. These three phases are explained in the following subsections.

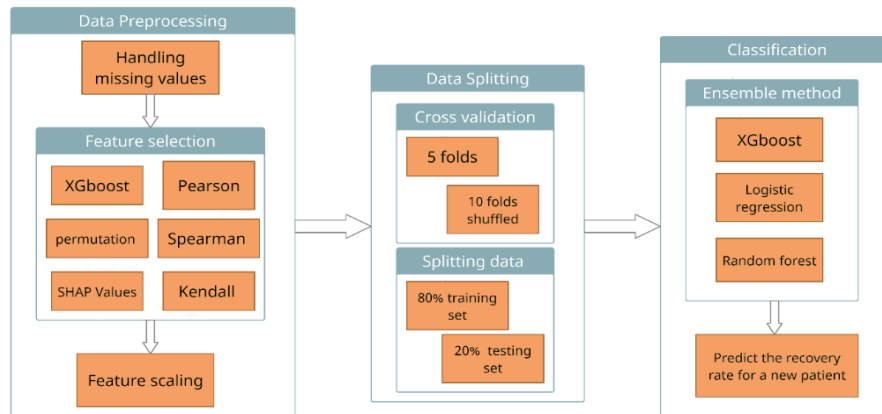


Figure 1. The proposed model architecture

2.1. Data processing

The medical dataset used by Yan *et al.* [15], [16] has been employed in this study. The dataset contains data about 375 patients from Tongji Hospital in Wuhan. 14 patients are excluded because of a lack of data as shown in Figure 2. For each day the patient was hospitalized, laboratory results are included in the dataset, such that the dataset has several records for each patient.

The data is preprocessed at this phase to prepare it for the following stage and to verify that it is clean. This stage consists of three major steps. Step 1: fill the missing values of the patient with the mean of his own values since each patient gets a row for each day he spent in the hospital. The median of all patients is then applied to the remaining missing values. We first combine all rows for each patient in order to apply feature selection techniques, then we unmerge rows in order to utilize each row of the patient's days in the hospital as an instance representing a distinct state for the patient in order to train the model as we augmented the dataset to enhance the model accuracy. The feature selection is the second step. In this step, Person, Kendall, Spearman, and XGboost feature selection techniques are used to select the most correlated features to the target as shown in Figure 3. The most correlated features are obtained by combining the first three features for each approach.

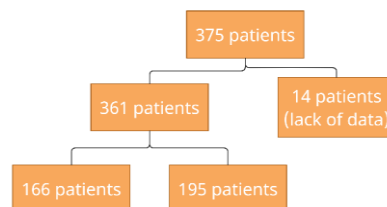


Figure 2. Chart of patient enrollment

The most three correlated features to the target using XGboost feature importance are lactate dehydrogenase, hypersensitive c-reactive protein, and age with scores of 0.46673, 0.06271822, and 0.05778807 respectively as shown in Figure 3(a). The most three correlated features to the target using XGboost permutation-based feature importance are lactate dehydrogenase, procalcitonin, and hypersensitive c-reactive protein with scores of 0.0232687, 0.00554017, 0.00443213 respectively as shown in Figure 3(b). Lactate dehydrogenase, hypersensitive c-reactive protein, and procalcitonin are the most correlated features to the target according to XGboost feature importance computed with SHAP values as shown in Figure 3(c).

The highest three correlation coefficients according to the pearson technique are 0.7460, and 0.71446 for neutrophils (%) and, hypersensitive c-reactive protein respectively. 0.68 is the score of neutrophils count, and lactate dehydrogenase as shown in Figure 3(d). The most three correlated features to the target using Kendall are lactate dehydrogenase, neutrophils (%), and hypersensitive c-reactive protein with scores of 0.6645, 0.637, and 0.5649 respectively as shown in Figure 3(e). The most three correlated features to the target using Spearman are lactate dehydrogenase, neutrophils (%), and hypersensitive c-reactive protein with scores of 0.8124, 0.7789, and 0.7649 respectively as shown in Figure 3(f).

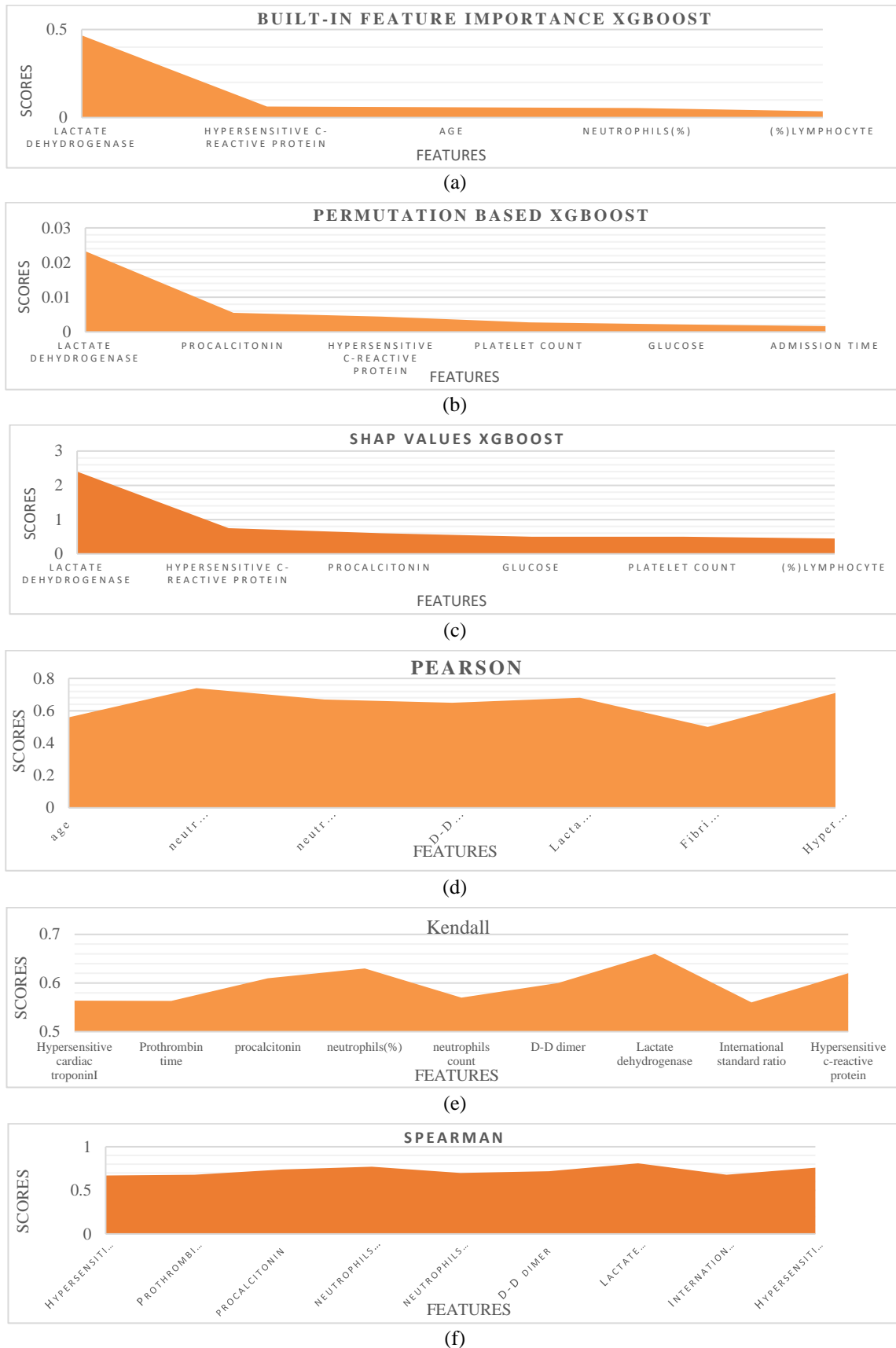


Figure 3. Feature importance scores are calculated by; (a) XGboost builtin feature importance, (b) XGboost permutation importance, (c) SHAP values, (d) Pearson correlation coefficient, (e) Kendall correlation coefficient, and (f) Spearman correlation coefficient

The first three features in each technique were picked to be utilized in the prediction model. The selected features are: neutrophils (%), hypersensitive c-reactive protein, lactate dehydrogenase, age, procalcitonin, and neutrophils count. The last step involves feature scaling. The main objective of this step is to normalize the range of independent features.

2.2. Splitting the dataset

The dataset has been partitioned in order to verify the proposed model with two distinct techniques. The first technique is to divide the whole dataset into training and testing sets with an 8:2 ratio, similar to the machine learning algorithm XGBoost [15], [16]. Therefore, the first technique is used to compare the proposed model to the model proposed by Li Yan [15], [16]. The second technique is cross validation, which utilizes each data record as a validation record once and a training record the other times [22]. In order to generate essentially unbiased error estimates [22].

2.3. Classification

The main goal of the proposed model is to classify the COVID-19 patient into 2 classes which are death and recovery classes. The classification is accomplished using the ensemble method. The ensemble method used more than a machine learning classifier to enhance the accuracy [23]. To choose which machine learning classifiers will be used in the ensemble method, the most popular classifiers in healthcare systems are applied to the dataset as shown in Figure 4. The random forest (RF) Algorithm is a cutting-edge machine learning approach that is capable of both regression and classification [23]. The logistic regression (LR) Algorithm is a prominent mathematical modeling approach used in machine learning for epidemiology datasets [23]. Decision trees (DTs) are applicable when the issue is straightforward, and the dataset is modest. K-nearest neighbor (KNN) is a well-known supervised classification technique with good accuracy that is widely utilized in a variety of industries [23]. Numerous classification or regression applications have previously implemented XGBoost, demonstrating that it is an excellent and efficient classifier construction approach [24]. So that all of these algorithms are applied to the dataset Then the most three accurate classifiers (RF, XGBoost, and LR) are used in the ensemble method. The random forest algorithm achieved the best results with 97% accuracy and 97% K-fold cross validation. There are different approaches for the ensemble method boosting, stacking, and bagging [25]. The soft voting technique is used to combine the results [26] of the RF, LR, and XGboost classifiers. The soft voting method accuracy was 98%, which is more accurate than the best three algorithms as shown in Figure 4. The recovery rate is estimated by the probability of the recovery class. After training the model with the dataset now we can predict the recovery rate of a new patient.

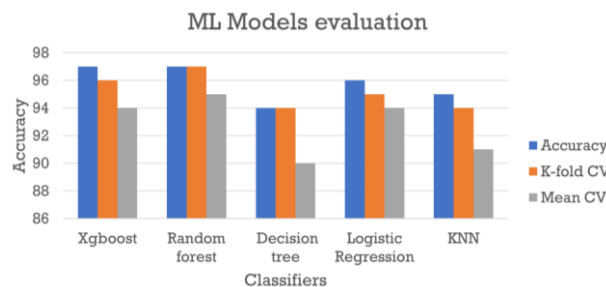


Figure 4. Machine learning models evaluation

3. RESULTS AND DISCUSSION

This section contains a discussion of the proposed model result and an explanation of various attempts are performed to make the machine learning model with the best results for this study. The proposed model is evaluated using the confusion matrix, accuracy, and cross validation. The accuracy was used to assess the ratio between the predicted label \hat{y}_s and the actual label y_s according to $\text{accuracy} = \frac{1}{S} \sum_{s=1}^S [\hat{y}_s = y_s]$. As shown in the Figure 5, the confusion matrix contains the true positives, false positives, true negatives, and false negatives:

- 119 survived states (true positives).
- 5 survived states (false positives).
- 1 dead state (false negative).
- 116 dead states (true negatives).

The last evaluation method was cross validation with 10 folds. The K-fold cross validation average score was 97%. To build the proposed model with the highest accuracy, 4 attempts are performed.

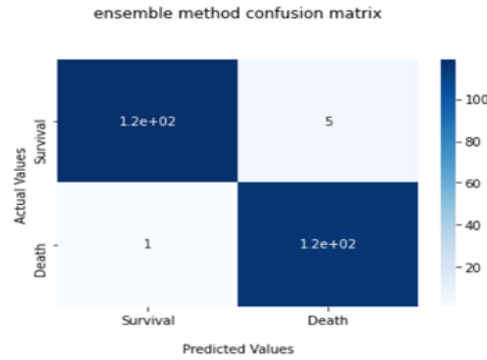


Figure 5. The confusion matrix of the ensemble method

3.1. Attempt 1

First, data preprocessing is carried out by, merging rows with the same patient id by getting the average of these values using the mean technique. Then replace the missing values with the median to avoid the outliers. Second, the feature selection techniques are applied. The first three features in each technique are then combined to be the most correlated features to the target. The most important features are neutrophils (%), hypersensitive c-reactive protein, lactate dehydrogenase, age, procalcitonin, neutrophils count. Finally, XGboost, random forest, decision tree, logistic regression, and KNN classifiers are applied to the previous features. As shown in Figure 6. The best model was random forest with 96% accuracy.

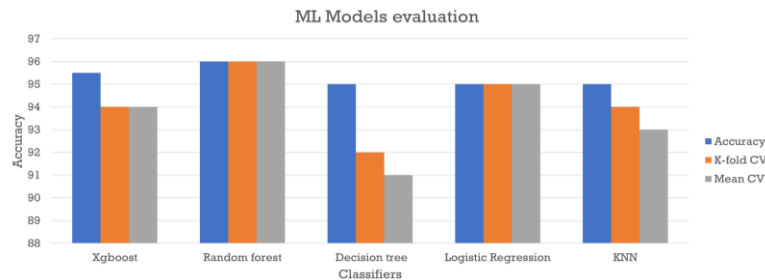


Figure 6. ML models evaluation for attempt 1

3.2. Attempt 2

Each patient has more than one row (each row is a day in the hospital). We use each row as a different state for the patient. The missing values are filled for the patient with the mean of his own values. Then used the median of all patients to fill in the rest missing values. Followed by applying the feature selection techniques. The first two features in each technique are then combined to be the most correlated features to the target. The most important features are neutrophils (%), hypersensitive c-reactive protein, lactate dehydrogenase, basophil (%), procalcitonin, and platelet count. Finally, XGboost, random forest, decision tree, logistic regression, and KNN classifiers are applied to the previous features. As shown in Figure 7. The best model was the random forest with 95% accuracy and 96% k-fold cross validation mean.

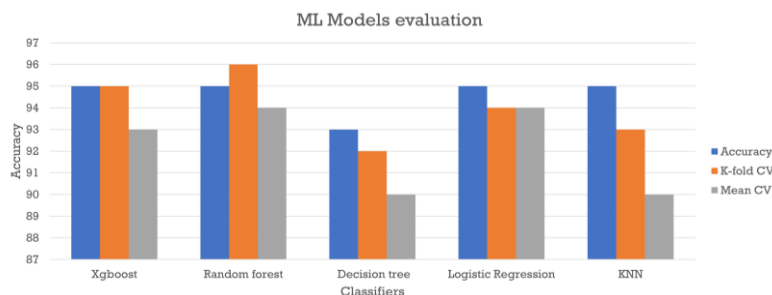


Figure 7. ML models evaluation for attempt 2

3.3. Attempt 3

Each patient has more than one row (each row is a day in the hospital). We use each row as a different state for the patient. The missing values are filled for the patient with the mean of his own values. Then used the median of all patients to fill in the rest missing values (same as attempt 2). After that, the same selected features in attempt 1 are used. Finally, all classifiers are applied. As shown in Figure 4. The best model was random forest and XGBoost with 97% accuracy and 97% k-fold cross validation mean for random forest.

3.4. Attempt 4

Each patient has more than one row (each row is a day in the hospital). The missing values of the patient are filled with the mean of his own values. Then fill in the rest missing values with a constant number (1). After that, we included only the last day of the patient in the hospital. Followed by applying the feature selection techniques and include the first two features in each technique. The selected features are hypersensitive c-reactive protein, lactate dehydrogenase, neutrophils (%), platelet count, (%) lymphocyte, D-D dimer. Finally, all classifiers are applied. As shown in Figure 8. The best accuracy was 94% for random forest and the XGboost. The best k-fold cross validation mean was 97% for random forest.

The results of all attempts are summarized in Table 1. The best result was for random forest in attempt 3 with 97% accuracy and 97% k-fold cross validation mean as shown in Table 1. The ensemble method was then used for all attempts, and the top three outcomes for the algorithms were chosen in each attempt to be used in the ensemble method. The best ensemble method result was for attempt 3 which was used in our model in this study.

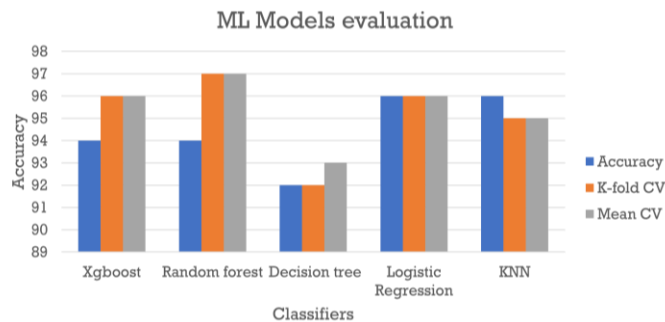


Figure 8. ML models evaluation for attempt 4

Table 1. Comparisons of the results of all algorithms in all attempts

Attempt	Evaluation matric	XGboost	Random forest	Decision tree	Logistic regression	KNN
1	Accuracy	95.5%	96%	95%	95%	93%
	K-fold mean	94%	96%	92%	95%	94%
	Cross validation	94%	96%	91%	95%	93%
	Mean					
2	Accuracy	95%	95%	93%	95%	95%
	K-fold mean	95%	96%	92%	94%	93%
	Cross validation	93%	94%	90%	94%	90%
	Mean					
3	Accuracy	97%	97%	94%	96%	95%
	K-fold mean	96%	97%	94%	95%	94%
	Cross validation	94%	95%	90%	94%	91%
	Mean					
4	Accuracy	94%	94%	92%	96%	96%
	K-fold mean	96%	97%	92%	96%	95%
	Cross validation	96%	97%	93%	96%	95%
	Mean					

Yan [15], [16] applied a machine learning model to predict the survival of severe COVID-19 patients using the XGboost algorithm. When applying their model to the dataset using the same three medical features (lactate dehydrogenase, hypersensitive c-reactive protein, (%) lymphocyte) obtained 94% accuracy and 96% k-fold cross validation. The model confusion matrix is shown in Figure 9. In addition to the results of the proposed model and the XGBoost machine learning algorithm [15], [16] are shown in Table 2.

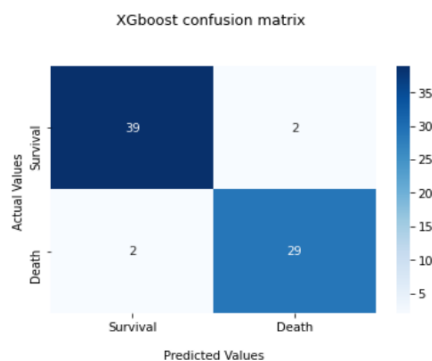


Figure 9. XGboost confusion matrix

Table 2. Comparison of the proposed model scores and the XGBoost machine learning algorithm [15], [16]

Model	Averaging strategy	Precision	Recall	F1-score	Accuracy
The model proposed by Li Yan [15], [16]	Macro avg	0.943	0.943	0.943	
	Weighted avg	0.944	0.944	0.944	94%
Our proposed model	Macro avg	0.975	0.976	0.975	
	Weighted avg	0.976	0.975	0.975	98%

4. CONCLUSION

In this study, a machine learning model is applied to predict the recovery rate of COVID-19 patients. The model aims to identify patients at high risk. At first, the model selects the most correlated medical features to the target value. The feature selection step is applied to provide the greatest accuracy outcomes with the least amount of processing in the shortest amount of time. The most relevant features are neutrophils (%), hypersensitive c-reactive protein, lactate dehydrogenase, age, procalcitonin, and neutrophils count. The recovery rate of COVID-19 patients is predicted using the ensemble method based on XGboost, random forest, and logistic regression. The model is applied to medical information for COVID-19 patients with an accuracy score of 98%. In future work, the model will be applied to a bigger dataset to enhance the model performance.




REFERENCES

- [1] S. Zhang *et al.*, "Development and validation of a risk factor-based system to predict short-term survival in adult hospitalized patients with COVID-19: a multicenter, retrospective, cohort study," *Critical care*, vol. 24, p. 438, Jul. 2020, doi: 10.1186/s13054-020-03123-x.
- [2] C. Anastassopoulou, L. Russo, A. Tsakris, and C. Siettos, "Data-based analysis, modelling and forecasting of the COVID-19 outbreak," *PLOS ONE*, vol. 15, no. 3, p. e0230405, Mar. 2020, doi: 10.1371/journal.pone.0230405.
- [3] D. Liu *et al.*, "A machine learning methodology for real-time forecasting of the 2019-2020 COVID-19 outbreak using Internet searches, news alerts, and estimates from mechanistic models." *arXiv*, Apr. 08, 2020. doi: 10.48550/arXiv.2004.04019.
- [4] F. A. B. Hamzah *et al.*, "CoronaTracker: world-wide COVID-19 outbreak data analysis and prediction," Mar. 2020, doi: 10.2471/BLT.20.255695.
- [5] Z. Du *et al.*, "Risk for transportation of coronavirus disease from Wuhan to other cities in China," *Emerging infectious diseases.*, vol. 26, no. 5, pp. 1049–1052, May 2020, doi: 10.3201/eid2605.200146.
- [6] I. A. Hassan, A. A. M. Al-Azzawi, and M. L. Talal, "The use of information technology applications in the provision of health care to a COVID-19 patients," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 28, no. 1, Art. no. 1, Oct. 2022, doi: 10.11591/ijeecs.v28.i1.pp277-283.
- [7] T. J. Levy *et al.*, "Development and validation of a survival calculator for hospitalized patients with COVID-19," *medRxiv*, p. 2020.04.22.20075416, Jun. 2020, doi: 10.1101/2020.04.22.20075416.
- [8] S. O. Petrovan *et al.*, "Post COVID-19: a solution scan of options for preventing future zoonotic epidemics," *Biological reviews of the Cambridge Philosophical Society.*, vol. 96, no. 6, pp. 2694–2715, 2021, doi: 10.1111/brv.12774.
- [9] R. Behera and K. Das, "A survey on machine learning: concept, algorithms and applications," *Journal of Intelligent Learning Systems and Applications*, vol. 2, Feb. 2017, doi: 10.15680/IJIRCCCE.2017.0502001.
- [10] Z. Hu, Q. Ge, S. Li, L. Jin, and M. Xiong, "Artificial intelligence forecasting of COVID-19 in China." *arXiv*, Mar. 01, 2020. doi: 10.48550/arXiv.2002.07112.
- [11] N. Al-Rousan and H. Al-Najjar, "Data analysis of coronavirus COVID-19 epidemic in South Korea based on recovered and death cases," *Journal of medical virology*, vol. 92, no. 9, pp. 1603–1608, Sep. 2020, doi: 10.1002/jmv.25850.
- [12] D. DeCapprio, J. Gartner, C. J. McCall, T. Burgess, S. Kothari, and S. Sayed, "Building a COVID-19 vulnerability index." *medRxiv*, p. 2020.03.16.20036723, Mar. 30, 2020. doi: 10.1101/2020.03.16.20036723.
- [13] K. Moulaei, M. Shanbehzadeh, Z. Mohammadi-Taghiabad, and H. Kazemi-Arpanahi, "Comparing machine learning algorithms for predicting COVID-19 mortality," *BMC medical informatics and decision making*, vol. 22, no. 1, p. 2, 2022, doi: 10.1186/s12911-021-01742-0.
- [14] C. Feng *et al.*, "A novel artificial intelligence-assisted triage tool to aid in the diagnosis of suspected COVID-19 pneumonia cases in fever clinics," *Annals of translational medicine*, vol. 9, no. 3, p. 201, Feb. 2021, doi: 10.21037/atm-20-3073.




- [15] L. Yan *et al.*, “Prediction of criticality in patients with severe COVID-19 infection using three clinical features: a machine learning-based prognostic model with clinical data in Wuhan.” *medRxiv*, p. 2020.02.27.20028027, Mar. 03, 2020. doi: 10.1101/2020.02.27.20028027.
- [16] L. Yan *et al.*, “A machine learning-based model for survival prediction in patients with severe COVID-19 infection.” *medRxiv*, p. 2020.02.27.20028027, Mar. 17, 2020. doi: 10.1101/2020.02.27.20028027.
- [17] C. Iwendi *et al.*, “COVID-19 patient health prediction using boosted random forest algorithm,” *Front. Public Health*, vol. 8, 2020, Accessed: Jan. 17, 2023. [Online]. Available: <https://www.frontiersin.org/articles/10.3389/fpubh.2020.00357>, doi:10.3389/fpubh.2020.00357.
- [18] K. Selvakuberan, D. Kayathiri, B. Harini, and M. I. Devi, “An efficient feature selection method for classification in health care systems using machine learning techniques,” in *2011 3rd International Conference on Electronics Computer Technology*, Apr. 2011, pp. 223–226. doi: 10.1109/ICECTECH.2011.5941891.
- [19] G. V. Gopal and G. R. M. Babu, “An ensemble feature selection approach using hybrid kernel based SVM for network intrusion detection system,” *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 23, no. 1, Art. no. 1, Jul. 2021, doi: 10.11591/ijeecs.v23.i1.pp558-565.
- [20] J. Miao and L. Niu, “A survey on feature selection,” *Procedia Computer Science*, vol. 91, pp. 919–926, Jan. 2016, doi: 10.1016/j.procs.2016.07.111.
- [21] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, “Heart disease identification method using machine learning classification in e-healthcare,” *IEEE Access*, vol. 8, pp. 107562–107582, 2020, doi: 10.1109/ACCESS.2020.3001149.
- [22] M. Rafałó, “Cross validation methods: Analysis based on diagnostics of thyroid cancer metastasis,” *ICT Express*, vol. 8, no. 2, pp. 183–188, Jun. 2022, doi: 10.1016/j.icte.2021.05.001.
- [23] S. M. D. A. C. Jayatilake and G. U. Ganegoda, “Involvement of machine learning tools in healthcare decision making,” *Journal of healthcare engineering*, vol. 2021, p. e6679512, Jan. 2021, doi: 10.1155/2021/6679512.
- [24] D. Zhang, L. Qian, B. Mao, C. Huang, B. Huang, and Y. Si, “A data-driven design for fault detection of wind turbines using random forests and XGboost,” *IEEE Access*, vol. 6, pp. 21020–21031, 2018, doi: 10.1109/ACCESS.2018.2818678.
- [25] K. Raza, “Chapter 8 - Improving the prediction accuracy of heart disease with ensemble learning and majority voting rule,” in *U-Healthcare Monitoring Systems*, N. Dey, A. S. Ashour, S. J. Fong, and S. Borra, Eds., in *Advances in Ubiquitous Sensing Applications for Healthcare*. Academic Press, 2019, pp. 179–196. doi: 10.1016/B978-0-12-815370-3.00008-6.
- [26] R. Islam and Md. A. Shahjalal, “Soft voting-based ensemble approach to predict early stage DRC violations,” in *2019 IEEE 62nd International Midwest Symposium on Circuits and Systems (MWSCAS)*, Aug. 2019, pp. 1081–1084. doi: 10.1109/MWSCAS.2019.8884896.

BIOGRAPHIES OF AUTHORS






Prof. Amani Abdo    professor of information systems at Arab open University, Faculty of Computing and Faculty of computers and artificial intelligence, Helwan University. She was born in 1980, she obtained a bachelor’s degree in computers and information in 2000, and she obtained a Master’s degree in information systems in 2004; Then obtained her Ph.D. in bioinformatics in 2010 from the department of information systems, Faculty of Computers and Information, Helwan University. She has supervised many graduation projects in the field of data mining, big data, and artificial intelligence. She also supervised many master’s and doctoral dissertations in the fields of information systems, medical informatics, and bioinformatics. She can be contacted at email: amanyabdo_80@yahoo.com.



Kholoud Mohamed Elzalama    received the B.Sc. degree in software engineering Helwan University, Egypt in 2017, Since 2018 up till the present day, she has been working as a teaching assistant at the Faculty of computers and artificial intelligence at Helwan University. She can be contacted at email: kholoudelzalama@gmail.com.



Dr. Ahmed Elsayed Yakoub    obtained his B.Sc. in information systems from Helwan University (excellent with honor) in 2008. He obtained his MSc from Helwan University in 2014 and his Ph.D. in Information Systems in 2019. He worked as a teaching assistant at the faculty of Computers and Information, Helwan University from 2010 to 2014. Then from 2014 to 2016, he was a lecturer assistant at the same faculty. He is now a lecturer at the Faculty of computers and artificial intelligence, Helwan University. Machine learning, Big data analytics, and data management are some of his research interests. He can be contacted at email: eng_ahmedyakoup@yahoo.com.