Global Medical and Health Communication

Online submission: https://ejournal.unisba.ac.id/index.php/gmhc
DOI: https://doi.org/10.29313/gmhc.v11i3.12119

GMHC. 2023;11(3):200–208
pISSN 2301-9123 | eISSN 2460-5441

RESEARCH ARTICLE

Effectiveness of Machine Learning for COVID-19 Patient Mortality Prediction using WEKA

Husnul Khuluq,[1] Prasandhya Astagiri Yusuf,[2] Dyah Aryani Perwitasari[3]

[1]Department of Pharmacy, Faculty of Health Sciences, Universitas Muhammadiyah Gombong, Kebumen, Indonesia, [2]Department of Medical Physiology and Biophysics/Medical Technology Cluster IMERI, Faculty of Medicine, Universitas Indonesia, Central Jakarta, Indonesia, [3]Faculty of Pharmacy, Universitas Ahmad Dahlan, Yogyakarta, Indonesia

Abstract

Timely detection of patients with a high mortality risk in coronavirus disease 2019 (COVID-19) can substantially improve triage, bed allocation, time reduction, and potential outcomes. A potential solution is using machine learning (ML) algorithms to predict mortality in COVID-19 hospitalized patients. The study's objective was to create and verify individual risk assessments for mortality using anonymous demographic, clinical, and laboratory findings at admission, as well as to assess the possibility of death using machine learning. We used a standardized format and electronic medical records. Data from 2,313 patients were collected from two Muhammadiyah hospitals from January 2020 to July 2022. Utilizing each patient's clinical manifestation state at admission and laboratory parameters, 24 demographic, clinical, and laboratory results were studied. The algorithms analyzed were AdaBoost, logistic regression, random forest, support vector machine, naïve Bayes, and decision tree, which were applied through WEKA version 3.8.6. Random forest performed better than the other machine learning techniques, with precision, sensitivity, receiver operating characteristic (ROC), and accuracy of 78.6%, 78.7%, 85%, and 78.65%, respectively. The three top predictors were septic shock (OR=21.518, 95% CI=4.933–93.853), respiratory failure (OR=15.503, 95% CI=8.507–28.254), and D-dimer (OR=3.288, 95% CI=2.510–4.306). Machine learning–based predictive models, especially the random forest algorithm, may make it easier to identify patients at high risk of death and guide physicians' appropriate interventions.

Keywords: Data mining, inpatient mortality, machine learning algorithm, prediction model

Introduction

In 2019, Wuhan province in China identified the first case of a novel coronavirus, which is considered to have been transferred from animals to humans.[1] The virus is severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Originally known as the 2019 novel coronavirus (2019-nCoV), the disease is now known as COVID-19.[2] In March 2020, the World Health Organization declared COVID-19 a pandemic.[3]

When COVID-19 patients are first admitted, doctors frequently can only adequately determine their prognosis once the disease has worsened. Additionally, the course of COVID-19 can change abruptly, causing a patient with a stable status to develop a critical condition quickly.[4] Elderly age, male, and the presence of various comorbidities, such as diabetes, high blood pressure, high cholesterol levels, cardiovascular disease, and chronic kidney disease, have been linked to increased mortality rates and severe outcomes in individuals affected by COVID-19.[5,6]

Artificial intelligence (AI) is a discipline in computer science that aims to comprehend and construct intelligent entities, typically manifested as software programs.[7] AI research has utilized machine learning techniques, which can consider intricate relationships to detect patterns within the given data. Standard machine learning algorithms can be broadly categorized into two sorts based on the tasks they aim to solve: supervised and unsupervised.[7]

The study about mortality prediction in COVID-19 patients using supervised machine learning conducted in Korea shows that LASSO and linear SVM demonstrated ROC values of 94,6% and 97,7% in predicting mortality.[8] Another international study in India using eXtreme Gradient Boosting shows a ROC value of 85.8%.[9] Compared to the machine learning models, the numerous studies applying conventional statistical models have significant methodological weaknesses and provide a substantial risk of bias within multiple fields of study.[10]

PKU Muhammadiyah Yogyakarta and PKU

Received: 19 June 2023; Revised: 6 October 2023; Accepted: 8 December 2023; Published: 25 December 2023

Correspondence: Husnul Khuluq. Department of Pharmacy, Faculty of Health Sciences, Universitas Muhammadiyah Gombong. Jln. Yos Sudarso No. 461, Gombong, Kebumen 54412, Central Java, Indonesia. E-mail: husnulkhuluq@unimugo.ac.id

Muhammadiyah Gamping General Hospital, Yogyakarta, are reference hospitals that applied structural electronic medical records; the electronic medical records can be easily and rapidly accessed compared to manual medical records.

The first aim of this study was to construct a model for predicting death from the first day of patient admission. The secondary purpose was to investigate predictors of COVID-19 mortality. Six machine learning models were employed, utilizing the Waikato Environment for Knowledge Analysis (WEKA) version 3.8.6. WEKA is open-source software, and compared to other software such as Rapid Miner and Orange, WEKA offers a broader range of machine learning methods and facilitates SMOTE, a technique used to address imbalanced datasets in machine learning. It also enables data mining activities by providing an extensive array of tools for data preprocessing, classification, attribute selection, and visualization. Several standard familiar file formats can be utilized with WEKA, such as xls and csv.[11]

Methods

This retrospective observational study included a total population of 2,882 patients, all consecutive COVID-19 patients admitted to the PKU Muhammadiyah Yogyakarta and PKU Muhammadiyah Gamping General Hospital, Yogyakarta, Indonesia, from January 2020 to July 2022. Of these, 68 patients were pregnant, 74 were children under 18, and 427 were missing or incomplete data. Then, the patients who met the inclusion criteria were 2,313 and applied for analysis.

The inclusion criteria were (1) SARS-CoV-2 infection confirmed by RT-PCR assays on material collected by a nasopharyngeal and oropharyngeal swab, (2) hospitalized patients, and (3) age above 18 years. Excluded from the analysis were patients who died during admission, patients who did not have primary data, pregnancies, and patients who were relocated to other designated hospitals while hospitalized.

The required patient data acquired from their medical records were age, gender, cardiovascular risk factors (high blood pressure, type 2 diabetes mellitus, and lipid disorders); primary comorbidities, including chronic renal disease; history of coronary artery disease, chronic obstructive pulmonary disease, and peripheral vascular disease; and laboratory results. Hospitalized COVID-19 patients who were deceased and those who lived were analyzed differently. Figure 1 shows the study design visualization.

The main objective of the study was to construct similar models with enhanced accuracy parameters to provide a mortality risk predictor.

Primary patient data, such as age and gender, were included within the clinical variables. A record of associated chronic diseases was

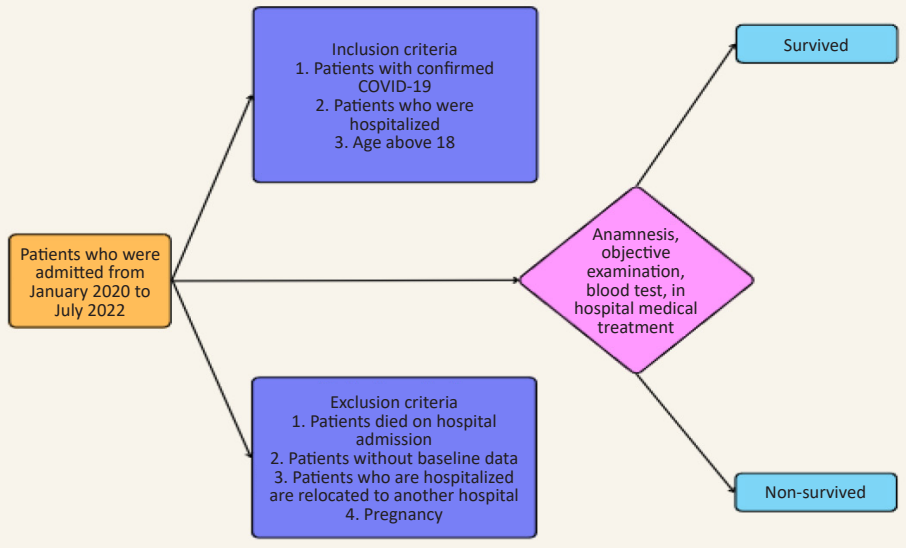

Figure 1 Study Design Visualization

collected, including diabetes, high blood pressure, cardiac and kidney disorders, and cancer. Laboratory parameters, such as lymphocytes, leukocytes, thrombocytes, neutrophils, D-dimer, glucose, creatinine, and hemoglobin, were evaluated. Out of 24 predictors, we selected these based on their odds ratio, p value, and relevance in clinical practice.

New variables were initially derived using Microsoft Excel 2013 to analyze the original data. Descriptive statistics for categorical variables used in this study are presented as absolute numbers (percentages). Since the study only used categorical variables, percentages were used to describe the data using chi-square tests to examine for correlations. The chi-square test was performed in bivariate analysis. The Pearson chi-square test or Fisher exact test was used, depending on which was suitable for analysis, and the odds ratio was computed. If $p<0.05$, it was considered significant. SPSS version 25.0 was used to analyze the data.

Data preprocessing is an essential part of the data mining process. It includes cleaning, transforming, and integrating raw data for analysis. The steps of data preparation are cleaning the data, integrating the data, reducing the amount of data, and changing the data. Data cleaning is getting clear of wrong, missing, or inaccurate information from a dataset. Integrating data involves combining information from different sources into a single file.

Data reduction is the process of making a dataset smaller by getting rid of duplicate or unimportant data. Changes are made to the format or layout of data to be used in the mining process.

We evaluate the predictors of hospitalized mortality using a set of machine learning algorithms that were adjusted. Six supervised machine learning algorithms were used in WEKA version 3.8.6 to construct mortality prediction models utilizing the preprocessed data: AdaBoost, logistic regression (LR), random forest (RF), support vector machine (SVM), naïve Bayes, and decision tree (DT).

For this research's categorization problem, supervised machine learning was selected among different methods. The nonlinearity in the data was clarified by supervised machine learning, which also constructs a performance that maps the input (predictor variables) to the output (mortality). As input data are processed, the outcomes of supervised machine learning become more accurate, and the predictions are more likely to fall throughout the allowed range.[12]

Researchers used the synthetic minority oversampling technique (SMOTE) to make more examples for the surviving class, which is in the minority group. SMOTE oversamples the minority class by making more artificial samples, and this increases the size of the class with fewer samples. After that, the researchers implemented the spread subsample to decrease the number of subjects in the majority class or surviving class, thus reducing it to balance with the minority class. When we oversample the minority class and undersample the majority class or cut off several samples in the class with more samples, the classifier will work better.[13]

The optimum hyperparameters for each model were obtained using WEKA's explorer module. The selected hyperparameters were those with the highest performance values. The effectiveness and general error of the comprehensive classification models were assessed using a tenfold cross-validation process system. All models were tested ten times using WEKA's experimenter module, and repeating ten-fold cross-validation was utilized to make comparisons of the performance-based prediction.

To produce the performance metrics (sensitivity, specificity, accuracy, precision, and ROC) generated from testing alone, the validation findings from ten experimental models were combined.[14]

Building an accurate machine learning model requires a fundamental component called model performance evaluation. Utilizing performance metrics for the confusion matrix, the predictive models were assessed (Table 1).

We used assessment indicators comprising accuracy, specificity, precision, sensitivity, and ROC chart criteria to assess the performance of the predictive models. To determine the best model for predicting COVID-19 mortality, All of these evaluation measures were contrasted based on their performance (Table 2).

The Research Ethics Committee of PKU Muhammadiyah Gamping Hospital approved the study protocol with exemption number 144/KEP-PKU/VII/2022). Additionally, because the study was retrospective, informed consent was not required.

Results

Between January 2020 and July 2022, a total

Table 1 Confusion Matrix

Output	Predicted Values	
	Non-survival (+)	Survival (–)
Actual value		
Non-survival (+)	TP	TN
Survival (–)	FP	FN

Note: TP: true positive is the number of cases the algorithm correctly classifies as positive; FP: false positive is the number of cases the algorithm incorrectly classifies as positive; FN: false negative is the number of cases the algorithm incorrectly classifies as negative; TN: true negative is the number of cases the algorithm correctly classifies as negative

Table 2 Performance Evaluation Measures

Performance Criteria	Item
Accuracy	(TP+TN)/(TP+TN+FP+FN)
Precision	TP/(TP+FP)
Sensitivity/recall	TP/(TP+FN)
Specificity	TN/(TN+FP)

Note: TP: true positive, TN: true negative, FP: false positive, FN: false negative

of 2,313 consecutive PCR-confirmed COVID-19 patients were retrospectively analyzed. Six hundred and thirty-eight COVID-19 patients (27.6% of the total) died at the time of their hospitalization, while the remaining 1,675 patients (73% of the total) survived. Table 3 provides more information on the symptoms, comorbidities, vital signs, and laboratory findings. Significantly, the three top odds ratios were septic shock (21,518), respiratory failure (15,503), and D-dimer (3,288). Type 1 diabetes mellitus was the top predictor among comorbidities (1,453).

In Table 3, 3 datasets were created. The first is the original dataset, which has 1,675 instances of the class that survived and 638 instances that did not survive.

Following the processing of the SMOTE algorithm, the number of cases belonging to the minority class was increased by generating synthetic samples, and the resulting dataset was then saved in the second dataset. This dataset contains 1,675 cases that related to the survived class and 1,276 cases that related to the updated non-survived class.

The researchers then executed the spread subsample technique, which undersamples the majority class to create a balanced dataset, which was then saved in the third dataset. By executing the SMOTE and Spread Subsample steps, the researchers created a balanced dataset, which was then used to train and test the COVID-19 predictor (Figure 2).

Figure 3 overviews the machine learning models' final testing results. The random forest algorithm was the best machine learning, with a precision of 78.6%, sensitivity of 78.7%, ROC of 85%, and accuracy of 78.65%. The 85% ROC value indicates good accuracy. The ROC needs to be higher than 0.5 for a diagnostic test to be considered meaningful. ROC≥0.8 is typically regarded as acceptable.[15]

Discussion

We use a ten-fold cross-validation technique to increase data utilization for training and

Table 3 SMOTE and Spread Subsample Methods' Results

Dataset Number	Technique Used	No. of Cases from the Surviving Class	No. of Cases from the Non-survived Class
1	–	1,675	638
2	SMOTE	1,675	1,276
3	Spread subsample	1,675	1,276

Table 4 Baseline Characteristics of Study Participants According to Mortality (Survival Vs Non-survival)

Parameters	Survival		Non-survival		OR (95% CI)	p
	n	%	n	%		
Total sample	2,126	72.4	638	27.6		
Comorbidity						
Pneumonia	47	2.0	10	0.4	0.552 (0.277–1.098)	0.086
Hypertension	422	18.2	163	7.0	1.019 (0.826–1.259)	0.861
Septic shock	2	0.1	16	0.7	21.518 (4.933–93.853)	0.000*
Chronic kidney disease	36	1.6	18	0.8	1.322 (0.745–2.345)	0.339
Acute kidney disease	3	0.1	3	0.1	2.633 (0.530–13.080)	0.356
Type 2 DM	235	10.2	115	5.0	1.347 (1.055–1.721)	0.017**
Type 1 DM	176	7.6	93	4.0	1.453 (1.110–1.903)	0.006**
Asthma	39	1.7	6	0.3	0.398 (0.168–0.945)	0.031**
Anemia	40	1.7	13	0.6	0.850 (0.452–1.600)	0.615
Respiratory failure	13	0.6	69	3.0	15.503 (8.507–28.254)	0.000*
COPD	7	0.3	2	0.1	0.749 (0.155–3.617)	1.000
Cerebral infarction	20	0.9	4	0.2	0.522 (0.178–1.533)	0.229
CHF	14	0.6	11	0.5	2.081 (0.940–4.609)	0.065
Myocardial infarction	8	0.3	2	0.1	0.655 (0.139–3.094)	0.736
Vital signs/laboratory result						
High blood pressure	1,092	47.2	477	20.6	1.582 (1.289–1.942)	0.000*
SPO_2<90%	1,329	57.5	563	24.3	1.954 (1.494–2.556)	0.000*
Lymphocytes	885	38.3	485	21.0	2.830 (2.304–3.475)	0,000*
Leukocytes	447	19.3	261	11.3	1.902 (1.571–2.303)	0.000*
Thrombocytes	339	14.7	149	6.4	1.201 (0.965–1.494)	0.101
Neutrophils	1,069	46.2	534	23.1	2.911 (2.307–3.673)	0.000*
D-dimer	1,192	51.5	568	24.6	3.288 (2.510–4.306)	0.000*
Glucose	667	28.8	357	15.4	1.915 (1.593–2.303)	0.000*
Creatinine	787	34.0	403	17.4	1.935 (1.137–2.334)	0.000*
Hemoglobin	772	33.4	303	13.1	1.058 (0.881–1.270)	0.546
Demographics						
Gender (male)	877	37.9	383	16.6	1.367 (1.136–1.645)	0.001**
Age>65 years	72	3.1	54	2.3	2.059 (1.429–2.966)	0.001**

Note: *p<0.001 and **p<0.05 considered significant, DM: diabetes mellitus, SPO_2: peripheral oxygen saturation, COPD: chronic obstructive pulmonary disease, CHF: congestive heart failure, OR: odds ratio, CI: confidence interval

validation instead of overfitting or data overlaps between the test and validation sets. Furthermore, this technique helped to decrease the deviation in prediction error and frequently used and recommended validation methodology in machine learning and data mining.[16]

Random forest was an effective machine learning algorithm to categorize the mortality risk in a study population of patients admitted to the PKU Muhammadiyah Gamping and PKU Muhammadiyah hospitals in Yogyakarta. This study was similar to other international studies, such as a study in Italy suggesting that random forest was the best machine learning with an ROC score of 88%,[17] and another study in Iran showed an ROC score of 83.6%.[18] The random forest algorithm is a popular ensemble learning method that exploits the combination of numerous decision trees in order to create accurate predictions. Employing an ensemble technique helps reduce overfitting problems and enhance the overall generalized performance. Random forests have been observed to exhibit computational inefficiency and prolonged training times when used to huge datasets.[19] Several international studies have shown different results on the best machine learning algorithm. A study conducted on the Korean population suggests that the decision tree algorithm is more effective at predicting the probability of death among COVID-19-infected patients compared to different algorithms, such as the support

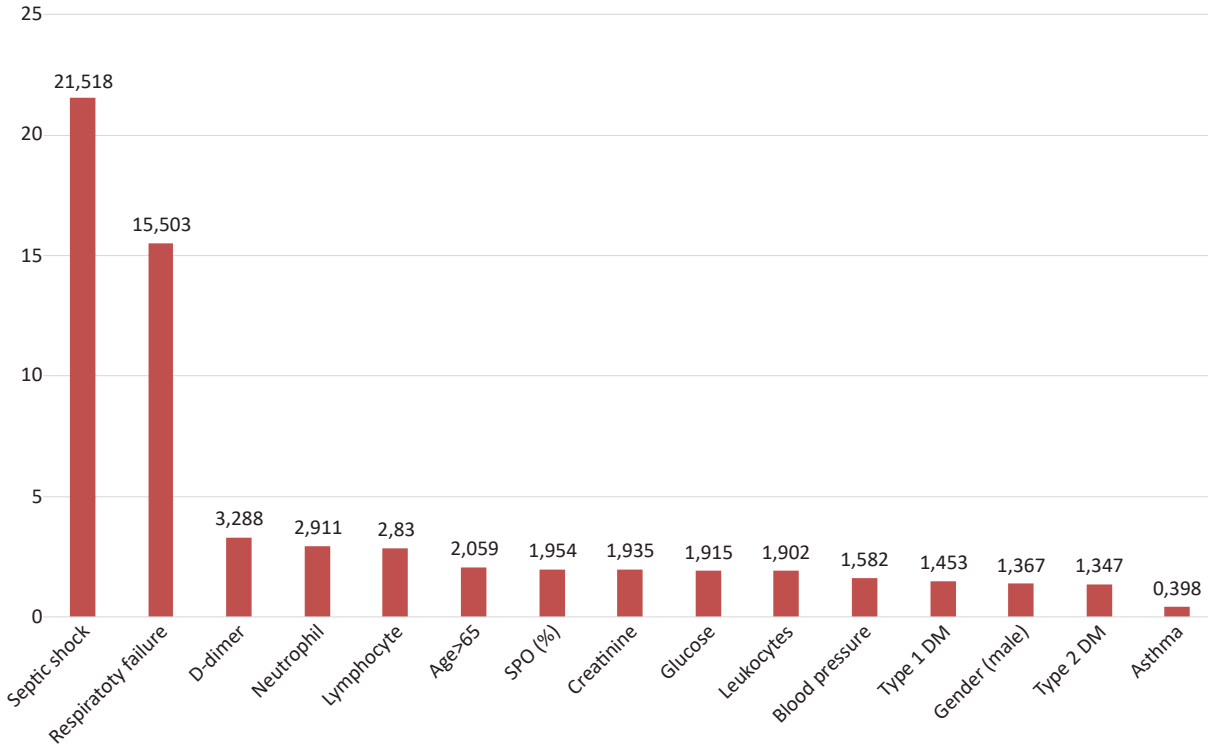

Figure 2 Odds Ratios of Top 15 Predictors based on Bivariate Analysis (p<0.05)

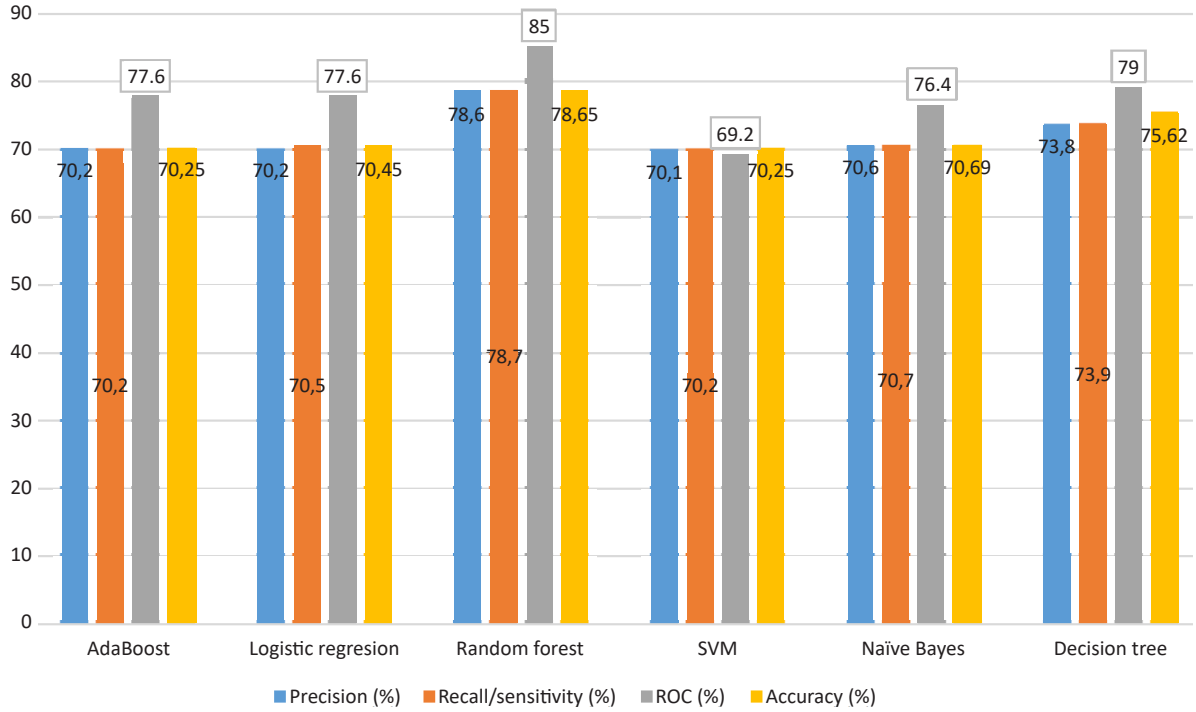

Figure 3 Visual Comparison of Machine Learning Algorithm Capabilities for COVID-19 Mortality Prediction

vector machine, naïve Bayes, logistic regression, random forest, and K-nearest neighbor.[20] The results could differ due to the differences in the datasets used and the features selected.[21]

In this study, the patient's age over 65 years was the main predictor and high mortality risk, significantly indicated here with a high odds ratio (OR=2.059, 95% CI=1.429–2.966). Other studies also suggest that older age predicts an increased mortality risk.[22] Comorbidities that are common among the elderly, including type 2 diabetes mellitus, high blood pressure, coronary artery disease, and pulmonary disease, are indicators of a poor prognosis, more significant mortality, and morbidity rates. The absence of symptoms in the older population and their ability to carry a significant viral load renders them a viable vector for viral transmission.[23] Additionally, odds ratios (OR=1.367, 95% CI=1.136–1.645) for men indicate a greater risk of mortality, which is consistent with a study conducted by Doerre and Doblhammer[24] that death rates are twice as high for men than for women in every age group. The expression of angiotensin-converting enzyme 2 (ACE2) receptors, which have a role in facilitating the entry of the SARS-CoV-2 virus and its transmission between humans, exhibits variations across individuals of various sexes. Estradiol has the potential to exert an effect on the expression of ACE2, a gene that is situated on the X chromosome. This chromosomal location may confer susceptibility to evading X-inactivation in females.

Septic shock was a high predictor in this study (OR=1.518, 95% CI=4.933–93.853). It is consistent with several international studies. An increased risk of death has been associated with respiratory symptoms, including respiratory failure and low oxygen levels (SPO_2<90%, see Table 3). It has been extensively researched in United States populations.[26]

In this study, comorbidities also had a minor impact on mortality risk prediction. These results are similar to another study[27] but only partially consistent with international studies, which have demonstrated that comorbidities play a significant role in risk prediction.[28]

Diabetes mellitus was the only comorbidity with an implication on mortality in this study. The odds ratios of types 2 and 1 diabetes mellitus were 1.347 (95% CI=1.055–1.721) and 1.453 (95% CI=1.110–1.903), similar to the meta-analysis carried out by Kumar et al.[29] Even though all types of diabetes mellitus have been linked to an elevated risk of in-hospital COVID-19-related mortality, our findings revealed that the risk was higher in type 1 diabetics than in type 2 diabetics. Various factors could explain this finding. Types 1 and 2 diabetes mellitus differ in terms of COVID-19-related mortality for a variety of reasons, including their distinct causes and pathophysiologies, patterns of complications or iatrogenic harms (such as hypoglycemia), treatments, intensity and duration of glycemia, and the effects of comorbidities either not taken into account in these analyses or were not appropriately considered.[30]

Among the laboratory features tested during admission, lymphopenia (<1 mg/dl) had an OR of 2.830 (95% CI=2.304–3.475), which matches the findings of a systemic review and meta-analysis.[31] Other significant factors including increased leucocytes (≥11 mg/dl), neutrophils (≥6 mg/dl), creatinine (≥1.2 mg/dl), and D-dimer (≥0.5 mg/dl) were shown to be mortality risk factors. These findings were similar to another international study.[32]

Conclusions

Independent predictors of mortality in patients with COVID-19 were age above 65 years, male, and diabetes mellitus. They have vital signs and laboratory tests: septic shock, respiratory failure, O_2 saturation, higher leucocytes, neutrophils, creatinine, and D-dimer. These parameters could be combined in a random forest machine learning model to provide a moderate-accuracy predictor of mortality with an ROC of 85%.

Conflict of Interest

None declared.

Acknowledgments

This research is funded by the Doctoral Research Grant from the Ministry of Education Culture Research and Technology in the *Penelitian Disertasi Doktor*, Indonesia, in 2022.

References

1. Centers for Disease Control and Prevention. COVID data tracker [Internet]. Atlanta: Centers for Disease Control and Prevention;

2023 [cited 2023 April 10]. Available from: https://covid.cdc.gov/covid-data-tracker.
2. Singhal T. A review of coronavirus disease-2019 (COVID-19). Indian J Pediatr. 2020;87(4):281–6.
3. Cucinotta D, Vanelli M. WHO declares COVID-19 a pandemic. Acta Biomed. 2020;91(1):157–60.
4. Cascella M, Rajnik M, Aleem A, Dulebohn SC DNR. Features, evaluation, and treatment of coronavirus. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2023 [cited 2023 April 20]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK554776.
5. Cummings MJ, Baldwin MR, Abrams D, Jacobson SD, Meyer BJ, Balough EM, et al. Epidemiology, clinical course, and outcomes of critically ill adults with COVID-19 in New York city: a prospective cohort study. Lancet. 2020;395(10239):1763–70.
6. Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. Lancet. 2020;395(10229):1054–62.
7. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng. 2018;2(10):719–31.
8. An C, Lim H, Kim DW, Chang JH, Choi YJ, Kim SW. Machine learning prediction for mortality of patients diagnosed with COVID-19: a nationwide Korean cohort study. Sci Rep [Internet]. 2020;10(1):18716.
9. Kar S, Chawla R, Haranath SP, Ramasubban S, Ramakrishnan N, Vaishya R, et al. Multivariable mortality risk prediction using machine learning for COVID-19 patients at admission (AICOVID). Sci Rep. 2021;11(1):12801.
10. Kwok SWH, Wang G, Sohel F, Kashani KB, Zhu Y, Wang Z, et al. An artificial intelligence approach for predicting death or organ failure after hospitalization for COVID-19: development of a novel risk prediction tool and comparisons with ISARIC-4C, CURB-65, qSOFA, and MEWS scoring systems. Respir Res. 2023;24(1):79.
11. Attwal KPS, Dhiman AS. Exploring data mining tool-WEKA and using WEKA to build and evaluate predictive models. Adv Appl Math Sci. 2020;19(6):451–69.
12. Brownlee J. Statistical methods for machine learning: discover how to transform data into knowledge with python [e-book]. San Juan: Machine Learning Mastery; 2019 [cited 2023 May 10]. Available from: https://machinelearningmastery.com/statistics_for_machine_learning.
13. Brownlee J. SMOTE for imbalanced classification with python [Internet]. San Juan: Machine Learning Mastery; 2021 [cited 2023 May 15]. Available from: https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification.
14. Indarto, Utami E, Raharjo S. Mortality prediction using data mining classification techniques in patients with hemorrhagic stroke. In: Proceeding Virtual Conference of the 2020 8th International Conference on Cyber and IT Service Management (CISTM); 2020 October 23–24; Pangkalpinang, Indonesia. Piscataway: Institute of Electrical and Electronics Engineers; 2020 [cited 2023 May 20]. p. 1–5. Available from: https://ieeexplore.ieee.org/document/9268802.
15. Nahm FS. Receiver operating characteristic curve: overview and practical use for clinicians. Korean J Anesthesiol. 2022;75(1):25–36.
16. Wu CC, Yeh WC, Hsu WD, Islam MM, Nguyen PAA, Poly TN, et al. Prediction of fatty liver disease using machine learning algorithms. Comput Methods Programs Biomed. 2019;170:23–9.
17. Laino ME, Generali E, Tommasini T, Angelotti G, Aghemo A, Desai A, et al. An individualized algorithm to predict mortality in COVID-19 pneumonia: a machine learning based study. Arch Med Sci. 2022;18(3):587–95.
18. Akbarzadeh M, Alipour N, Moheimani H, Zahedi AS, Hosseini-Esfahani F, Lanjanian H, et al. Evaluating machine learning-powered classification algorithms which utilize variants in the GCKR gene to predict metabolic syndrome: Tehran Cardio-metabolic Genetics Study. J Transl Med. 2022;20(1):164.
19. Zhou S, Mentch L. Trees, forests, chickens, and eggs: when and why to prune trees in a random forest. Stat Anal Data Min. 2023;16(1):45–64.
20. Muhammad LJ, Islam MM, Usman SS, Ayon SI. Predictive data mining models for novel coronavirus (COVID-19) infected patients' recovery. SN Comput Sci. 2020;1(4):206.
21. Salari N, Kazeminia M, Sagha H,

Daneshkhah A, Ahmadi A, Mohammadi M. The performance of various machine learning methods for Parkinson's disease recognition: a systematic review. Curr Psychol. 2022;42:16637–60.
22. Chen R, Liang W, Jiang M, Guan W, Zhan C, Wang T, et al. Risk factors of fatal outcome in hospitalized subjects with coronavirus disease 2019 from a nationwide analysis in China. Chest. 2020;158(1):97–105.
23. Dadras O, SeyedAlinaghi SA, Karimi A, Shamsabadi A, Qaderi K, Ramezani M, et al. COVID-19 mortality and its predictors in the elderly: a systematic review. Health Sci Rep. 2022;5(3):e657.
24. Doerre A, Doblhammer G. The influence of gender on COVID-19 infections and mortality in Germany: insights from age- and gender-specific modeling of contact rates, infections, and deaths in the early phase of the pandemic. PLoS One. 2022;17(5):e0268119.
25. Peckham H, de Gruijter NM, Raine C, Radziszewska A, Ciurtin C, Wedderburn LR, et al. Male sex identified by global COVID-19 meta-analysis as a risk factor for death and ITU admission. Nat Commun. 2020 Dec;11(1):6317.
26. Bahl A, Van Baalen MN, Ortiz L, Chen NW, Todd C, Milad M, et al. Early predictors of in-hospital mortality in patients with COVID-19 in a large American cohort. Intern Emerg Med. 2020;15(8):1485–99.
27. Tezza F, Lorenzoni G, Azzolina D, Barbar S, Leone LAC, Gregori D. Predicting in-hospital mortality of patients with COVID-19 using machine learning techniques. J Pers Med. 2021;11(5):343.
28. Kim HR, Jin HS, Eom YB. Association between manba gene variants and chronic kidney disease in a Korean population. J Clin Med. 2021;10(11):2255.
29. Kumar A, Arora A, Sharma P, Anikhindi SA, Bansal N, Singla V, et al. Is diabetes mellitus associated with mortality and severity of COVID-19? A meta-analysis. Diabetes Metab Syndr. 2020;14(4):535–45.
30. Barron E, Bakhai C, Kar P, Weaver A, Bradley D, Ismail H, et al. Associations of type 1 and type 2 diabetes with COVID-19-related mortality in England: a whole-population study. Lancet Diabetes Endocrinol. 2020;8(10):813–22.
31. Lu Q, Wang Z, Yin Y, Zhao Y, Tao P, Zhong P. Association of peripheral lymphocyte and the subset levels with the progression and mortality of COVID-19: a systematic review and meta-analysis. Front Med (Lausanne). 2020;7:558545.
32. Li X, Xu S, Yu M, Wang K, Tao Y, Zhou Y, et al. Risk factors for severity and mortality in adult COVID-19 inpatients in Wuhan. J Allergy Clin Immunol. 2020;146(1):110–8.