Journal for Re Attach Therapy and Developmental Diversities
eISSN: 2589-7799
2023 August; 6 (10s): 1130-1141

Multivariate Analysis and Machine Learning: Mortality Predictions In COVID-19 Patients from Comorbidity, Demographic and Laboratory Findings

Husnul Khuluq[1], Prasandhya Astagiri Yusuf [2], Dyah Aryani Perwitasari [3], Abdul Fadhil [4]

[1] Dept. of Pharmacy, Faculty of Health Sciences, Universitas Muhammadiyah Gombong
[2] Dept. of Medical Physiology of Biophysis / Medical Technology Cluster IMERI, Faculty of Medicine, Universitas Indonesia
[3] Faculty of Pharmacy, Universitas Ahmad Dahlan
[4] Dept. of Electrical Engineering, Technical Faculty, Universitas Ahmad Dahlan

Received: 24- June -2023
Revised: 27- July -2023
Accepted: 21- August -2023

Abstract

Objective: COVID-19 Patients were constantly at a risk of death. It has been demonstrated that the utilization of machine learning (ML) algorithms could be a possible strategy for prediction mortality. Aim: This study aimed to analysis six Machine Learning (ML) algorithms in an multivariate analysis to identify key clinical, demographic and laboratory finding to predict mortality in COVID-19 pandemic Materials and methods: This retrospective study consisted of persons-under-investigation for COVID-19. Dataset taken from data science community (kaggle.com), predictive models of mortality were constructed and compared using six supervised machine learning algorithms: KNN, naivebayes, SVM, decision tree, random forest and logistics regression using 10-fold cross-validation and multivariate analysis. The performance of algorithms was assessed using precision, recall, F-measure accuracy and area under the receiver operating characteristic curve (ROC). The Waikato Environment for Knowledge Analysis (WEKA) version 3.8.6 for analysis. Multivariate analysis using Logistic regression were used to predict mortality. Results: A total of 4711 patients were included in the analysis. The top 4 mortality predictors were Mean Artery Pressure (MAP) ($p<0.001$; OR 17.071(12.233-23.820), stroke ($p<0.001$;OR 3.499(1.883-6.503), Age ($p<0.001$;OR 3.23(2.716-3.830), IL6 ($p<0.001$; OR 2.03(1.512-2.725. Logistic regression was the best ML algorithms predicted mortality with 81% ROC. Conclusion: This study identifies important independent clinical variables that predict COVID-19 infection-related mortality. The prediction method is helpful, easily improved, and easily retrained with new data. This method can be applied right away and may help front-line doctors make clinical decisions in situations where there are limited resources and time.

Keywords: big data study, data mining research, machine learning algorithm, prediction models

1. INTRODUCTION

Clinical Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV-2), the responsible agent of novel coronavirus (COVID-19 or 2019-nCoV), appeared in late 2019 and likely to come from Hubei Province, China called Wuh*an*[1] [2] *I*t is suspected that COVID-19, which is quickly spreading in humans, was initially originated from bats and likely spread to humans through intermediate hosts, the raccoon dog (*Nyctereutes procyonoides*) and palm civet (*Paguma larvata*) [3][4] The earliest symptom of SARS-COV-2 were fever, coughing, and shortness of breath, which frequently matched the flu. [2] Since then, COVID-19 advanced to a critical stage and spread globally, infecting numerous people. Human-to-human transmission of COVID-19 from infected patients with moderate symptoms has also been documented.[5]. Nevertheless, no drug or vaccine has been clinically shown to cure COVID-19 pandemic, so other non-clinical or non-medical therapeutic techniques, such as data mining techniques, machine learning, and expert systems, among other artificial intelligence techniques, are needed to contain and prevent further outbreak.

Data Mining (DM) is a sophisticated AI methods for finding new, practical, and reliable hidden patterns or knowledge from datasets. [6] The method identifies connections, information, or patterns between the datasets in multiple or a specific dataset. [7][2] It is also frequently utilized for disease diagnosis and prognosis, such as Severe Acute Respiratory Syndrome Coronavirus (SARS-COV) and Middle East Respiratory Syndrome

Coronavirus (MERS-CoV) that were so far discovered in 2003 and 2012, respectively [2] A valuable resource to be mined and evaluated for new, relevant, and innovative knowledge or patterns extraction for better decision-making to contain the COVID-19 pandemic is the huge dataset generated daily around the world in relation to the 2019-nCoV pandemic. Data mining has been successfully used in the healthcare system for a range of functions, including patient outcome prediction, health outcome modeling, hospital ranking, and evaluation of treatment efficacy and infection control, stability, and recovery. [8][9]

We created a data mining model in this study to predict 2019-nCoV mortality. The models predict when patients with COVID-19 infection would survive, as well as those who might not survive due to the COVID-19 pandemic. The models aid medical professionals in identifying COVID-19 pandemic survivors who are infected.

2. METHODOLOGY

A. Dataset Collection and Description

The dataset was created by reviewing the main medical records and data from a healthcare monitoring software program (Clinical Looking Glass [CLG]; Streamline Health, Atlanta, Georgia). Age, comorbidities, and laboratory tests were collected at the presentation were posted on the Kaggle website. (https://www.kaggle.com/datasets/harshwalia/mortality-risk-clinincal-data-of-covid19-patients) [3]. There were 4711 instances and 45 attributes make up the dataset. and dataset cases of the 2019-nCoV pandemic-infected records that survived and did not survive were taken into consideration.

B. Data Mining Techniques/Algorithms

1. Logistic Regression (LR)

Logistic regression (LR) is applied in order to establish the relationship between categorical dependent variables and independent variables. [10] LR is used when the dependent variable has two values such as 0 and 1, yes and no or true and false and thus it is called binary logistic regression [11]. Furthermore, multinomial logistic regression is performed when the dependent variable has more than two values. Prediction of a modification of the dependent variables is made using a mathematical model of a set of explanatory variables for LR..The mathematical formula for the LR transformation is:

$$i = \text{Logistic regression } (p) = \text{Ln } \frac{(P)}{(1-P)}$$

Let, presume the dependent values are numerical of 1 and 0 where 0 reflect negative value and 1 positive value as a binary variable. Therefore, the mean of the binary variable will the proportion of positive values. If p is the proportion of observations with an outcome of 1, then 1 − p is the probability of an outcome of 0. The ratio p/(1 − p) is called the odds and the LG is the logarithm of the odds or just log odds.

2. Support Vector Machine (SVM)

One of the supervised learning techniques used for classification and regression is the support vector machine (SVM). [12] SVM's classification task requires training and test data that include some examples of the data. [13] The main objective of SVM is to create a model that will predict the target value or values because every instance in the training dataset has one or more target values. [12] SVM is used for regression by providing a different loss function that may be linear or nonlinear. [13]

3. Naive Bayes (NB)

One type of data mining classification method, known Naive Bayes, is used to identify dataset instances based on predetermined attributes [13]. NB is a probabilistic classifier that performs classification problems using Bayes theorem [5]. Below is the Bayes theorem:

$$P(A|B) = \frac{(P(B|A)P(A))}{P(B)}$$

4. Random Forest (RF) (2)

An ensemble learning method for classification and regression problems in data mining is called the random forest (RF) algorithm. When training, the algorithm creates a large number of decision trees. [9]. RF data mining algorithm is the best to be used for any decision tree with overfitting to its training dataset [14]

5. K-Nearest Neighbor (K-NN)

The non-parametric and supervised data mining classifier K-nearest neighbor (K-NN) is used for regression and classification problems. [15] In both tasks, the input variables consist of the K closes training dataset in the feature space. K-NN requires labeled input data to learn a function and provide the desired results from unlabeled input data. [16] When using K-NN classification, the outcome is a class membership in which each data instance is assigned to the class that has received the most support from its K-nearest neighbors. In contrast, the output of a K-NN regression is the property value of a data instance, which is the average of the value of the K-nearest neighbors.[17]

6. Decision Tree (DT)

Decision tree (DT) is an effective technique for classification problems in data mining as it can handle both categorical and continuous data, is simple to understand, and is constructed into phases that include growth and pruning phases, respectively.[14][7] [18].

Because it evaluates and matches the input data and categorizes them into a tree-like structure, a decision tree is a classification method that is more frequently used in medical diagnostic protocols because it is simple to learn and interpret. Analyzing decision trees in this study is attractive because they can produce sufficient visual information to determine whether or not cases. It generates criteria as an algorithm that divides data in stages based on the values of predictors, constructing a tree with roots and leaves. [19][20]

C. Data Mining Evaluation Technique

1. Conventional statistics

The original data was first processed using microsoft excel 13; new variables were derived were appropriate. All variables utilized in this study were categorical, and thus, percentages were used for summarization. We used IBM SPSS 25 for conventional statistics analysis. A chi-square tests to assess association and attributes with *p* value more than 0,25 removed [21]. Logistic regression test to ranked between attributes which selected from chi-square test.

2. Machine learning techniques

We used the WEKA Platform (version 3.8.5). WEKA's EXPLORER module to determine the optimal parameters for each algorithm used. A ten-fold cross-validation process system was used in all algorithms. Then, run all algorithms 10 times, using repeated ten-fold cross-validation, to facilitate comparison of the predictive performance based on the different evaluation criteria that are available in WEKA[22]

3. Algorithm evaluation

The critical feature of machine learning is data mining evaluation method since it provides as a basis for evaluating the accuracy and effectiveness of any data mining model or algorithm. [23]. It has been used to evaluate the effectiveness of data mining methods or models. [24] As a result, in this study, the performance matrices below are utilized to assess and choose the best data mining methods for analysis of the COVID-19 dataset.

To analyze the algorithms' effectiveness, we used the calculation of accuracy, specificity, precision, recall, F-measure, and the area under curve ROC (AUC).

A patient who not survived can be classified correctly (true positive-TP) or incorrectly (false negative-FN) and a patient who survived can be classified correctly (true negative-TN) or incorrectly (false positive-FP). [25]

The evaluation parameters are what we defined as:

a. True positive (TP) indicates how many patients the algorithm correctly identified as not survived.

b. True negative (TN) is the proportion of patients who were correctly classified by the algorithm as survived.

c. False positive (FP) means the number of patients who the algorithm incorrectly classified as not survived while being survived.

d. False negative (FN) depicts the number of patients who are not survived and incorrectly identified by the algorithm as survived.

e. Specificity represents the percentage of patients who are survived and correctly identified by the algorithm and is determined as follows:

TN/ (TN + FP) [18]

f. Sensitivity represent the percentage of patients who are not survived and correctly identified by the algorithm and is determined as follows:

TP/(TP + FN)[18]

g. Accuracy represent the percentage of patients who are correctly identified by the algorithm and is determined as follows:

(TP + TN)/(TP + TN + FP + FN).[18]

h. Precision is the proportion of patients that are correctly predicted as not survived among those labelled as not survived. Precision = TP/(TP + FP)[25]

i. F-measure. A measure that combine both Precision and Recall. F-measure = (2 x Precision x Recall)/ (Precision + Recall).[25]

3. RESULT AND DISCUSSION

A. Demographic Characteristics

The study included 4711 patients from datasheet. The mortality rate was higher among patient older than 65 years old . p-value <0,001 (table 1)

B. Co-morbidities

Non survival patients had a significantly higher prevalence of PVD than survived patients (15.07% versus 2.93% respectively; p-value <0.001). Besides, deceased patients were associated with higher prevalence of CHF (8.3 % vs. 3.18%, p-value <0.05) , COPD (9% versus 3.4%, p-value <0.05), Renal disease (12.69% Vs. 4.00%, p-value <0.05) , Stroke (0.64% vs. 0.59%, p value <0.001). There were no significant difference between survived and no-survived patients in terms of MI, CPD, DEMENT, DM Complicated, DM Simple and seizure (table 1).

C. Demographics

Non survival patients had a significally higher risk for Patients older than 65 years vs survived patients (31.2% versus 17.43% , p-value <0.001). Besides, deceased patients were associated with white and asian (p<0.05). There were no significant difference between survived and no-survived patients in terms of black and latino (table 1)

D. Vital sign

Deceased patients had significantly at lower MAP (p-value< 0.001) lower saturation rate (p-value <0.001) and high temperature (p <0.002). There was no significant difference between survived and no-survived patients in terms of systolic and diastolic blood pressure on admission (p-value 0.762 and p-value 0.577, respectively), (table 1).

E. Laboratory Test

Deceased patients were associated with higher value D-Dimer, AST, WBC, Lymphocytes, Procalcitonin, IL6 , INR, BUN, CrtnScore, Sodium, Ferritin, C-Reactive Prot, Troponin (p-value <0.001), PltsScore (p-value 0.009), CRP (p-value <0.001), Serum ferritin (p-value 0.001), and D-dimer (p-value <0.001) at time of admission. In

addition, deceased patients had lower platelet count (p-value 0.017) (Table 1). There were no significant difference between survived and no-survived patients in terms of Glucose and ALT (table 1)

Table 1. Bivariat analysis of Demographic, comorbidity, laboratory findings by survived and non-survived group

Parameter	All (4711)		survived		Non survived		
Comorbidity	*N*	*%*	*N*	*%*	*N*	*%*	*P*
MI	201	(4.27)	148	(3.14)	53	(1.13)	0.274
PVD	848	(18)	710	(15.07)	138	(2.93)	0.000*
CHF	541	(11.48)	391	(8.3)	150	(3.18)	0.031*
CVD	506	(10.74)	373	(7.92)	133	(2.82)	0.157
Dementia	372	(7.9)	269	(5.71)	103	(2.19)	0.069
COPD	265	(5.63)	185	(3.93)	80	(1.7)	0.015*
'DM Complicated'	495	(10.51)	380	(8.07)	115	(2.44)	0.287
'DM Simple'	686	(14.56)	518	(11)	168	(3.57)	0.485
Renal Disease	833	(17.68)	598	(12.69)	235	(4.99)	0.003*
Stroke	58	(1.23)	30	(0.64)	28	(0.59)	0.000*
Seizure	38	(0.81)	28	(0.59)	10	(0.21)	0.451
Demographics							
Age >65	2291	(48.63)	1470	(31.2)	821	(17.43)	0.000*
black	1743	(37)	1335	(28.34)	408	(8.66)	0.127
white	466	(9.89)	332	(7.05)	134	(2.84)	0.013*
asian	121	(2.57)	83	(1.76)	38	(0.81)	0.045*
Latino	1753	(37.21)	1348	(28.61)	405	(8.6)	0.064
Laboratory test and vital sign							
O_2 Sat < 94	1862	(39.52)	1261	(26.77)	601	(12.76)	0.000*
'Temp > 38'	854	(18.13)	613	(13.01)	241	(5.12)	0.002*
'D-Dimer > 3'	1151	(24.43)	730	(15.5)	421	(8.94)	0.000*
'Glucose <60 or > 500'	114	(2.42)	83	(1.76)	31	(0.66)	0.270
'AST > 40'	2121	(45.02)	1458	(30.95)	663	(14.07)	0.000*
'ALT > 40'	1292	(27.43)	966	(20.51)	326	(6.92)	0.208
WBC <11 or > 4.8'	3891	(82.59)	2895	(61.45)	996	(21.14)	0.000*
Lymphocytes < 1	2124	(45.09)	1498	(31.8)	626	(13.29)	0.000*
'IL6 > 150'	291	(6.18)	144	(3.06)	147	(3.12)	0.000*
Procalciton > 0.1'	1724	(36.6)	1098	(23.31)	626	(13.29)	0.000*
MAP < 70	339	(7.2)	288	(6.11)	51	(1.08)	0.000*

PltsScore>150or<450	1178	(25.01)	860	(18.26)	318	(6.75)	0.009*
INR > 1.2	1151	(24.43)	730	(15.5)	421	(8.94)	0.000*
BUN > 30	1285	(27.28)	787	(16.71)	498	(10.57)	0.000*
Creatinine >1.4	1703	(36.15)	1050	(22.29)	653	(13.86)	0.000*
Sodium < 139 or > 154	592	(12.57)	400	(8.49)	192	(4.08)	0.000*
Ferritin > 270	2561	(54.36)	1865	(39.59)	696	(14.77)	0.000*
C-Reactive Prot > 2	1853	(39.33)	1208	(25.64)	645	(13.69)	0.000*
Troponin > 0.4	450	(9.55)	248	(5.26)	202	(4.29)	0.000*

Data are presented as number (percentage). * *p* value based on chi-square test (<0.05)

MI: Myocardial Infarction; PVD : Peripheral Vascular Disease, CHF: Congestive Heart Failure, CVD: Cardiovascular Disease, COPD: Chronic Obstructive Pulmonary Disease, DM: Diabetes Mellitus, AST : Aspartat Aminotransferase ,ALT: Alanine Aminotransferase , BUN: Blood Urea Nitrogen , WBC: White Blood Cell , Plts: Platelets, INR: International Normalized Ratio

Table 2. Multivariate analysis of Demographic, comorbidity and laboratory findings

Parameter	OR (95%CI)		*p* value
Comorbidity			
PVD	0,503	(0,393-0,644)	0,000*
CHF	1,063	(0,802-1,409)	0,670
CPD	0,952	(0,722-1,254)	0,724
DEMENT	1,135	(0,844-1,526)	0,402
COPD	1,266	(0,911-1,758)	0,159
Renal Disease	1,291	(1,020-1,634)	0,034*
Stroke	3,499	(1,883-6,503)	0,000*
Demographics			
Age >65	3,225	(2,716-3,830)	0,000*
black	0,742	(0,592-0,930)	0,010*
white	0,930	(0,694-1,247)	0,629
asian	1,552	(0,950-2,535)	0,079
Latino	0,796	(0,640-0,990)	0,041*
Laboratory test and vital sign			
'O2 Sat < 94'	1,561	(1,323-1,841)	0,000*
Temp > 38'	1,133	(0,923-1,391)	0,233
D-Dimer > 3	1,141	(0,945-1,377)	0,171
AST > 40'	1,555	(1,286-1,880)	0,000*
'ALT > 40'	0,835	(0,679-1,028)	0,089
'WBC <11 or > 4.8'	1,105	(0,873-1,399)	0,405
'Lymphocytes < 1'	1,238	(1,052-1,457)	0,010*

IL6 > 150	2,030	(1,512-2,725)	0,000*
Procalciton > 0.1	0,649	(0,540-0,779)	0,000*
MAP < 70	17,071	(12,233 -23,820)	0,000*
PltsScore>150or<450	1,309	(1,087-1,577)	0,005*
BUN > 30	1,095	(0,889-1,350)	0,393
CrtnScore<139 or >154	1,810	(1,475-2,221)	0,000*
Sodium < 139 or > 154	1,305	(1,038-1,641)	0,023
Ferritin > 270	0,649	(0,540-0,779)	0,000*
C-Reactive Prot > 2	1,526	(1,268-1,837)	0,000*
Troponin > 0.4	1,417	(1,102-1,822)	0,007*

CI : confidental interval; OR: odds Ratio * *p* value based on logistic regression test ($p<0.05$)

MI: Myocardial Infarction; PVD : Peripheral Vascular Disease, CHF: Congestive Heart Failure, CVD: Cardiovascular Disease, COPD: Chronic Obstructive Pulmonary Disease, DM: Diabetes Mellitus, AST : Aspartat Aminotransferase ,ALT: Alanine Aminotransferase , BUN: Blood Urea Nitrogen , WBC: White Blood Cell , Plts: Platelets, INR: International Normalized Ratio

F. Multivariate Analysis

On multivariate analysis with logistic regression analysis, we found that age >65 years (OR = 2.84, 95% CI (2.41–3.36, p<0.001), black (OR=0.742, 95%CI(0.592-0.930,*p*=0.010), Latino (OR=0.796,95%CI(0.640-0.990,*p*=0.041), PVD (OR=0.503, 95% CI(0.393-0.644, *p*<0.001), CHF(OR=0.820, 95%CI (0.671-1.003,*p*=0.031), COPD (OR=0.731, 95% CI(0.557-0.959, *p*=0.015), Renal Disease (OR=0.784,95% CI (0.662-0.927,*p*=0.003), Stroke(OR=0.340,95% CI(0.202-0.571,*p*<0.001), significantly increased the risk of mortality among hospitalized patients. Concerning vital signs, we found that saturation (OR=0.499 ,95% CI (0.436-0.570, *p*<0.001), temperature (OR=0.782,95%CI (0.662-0.924,*p*=0.002), MAP (OR=17.071, 95% CI (12.233-23.820,p<0.001) significantly increased the risk of mortality among hospitalized patients (table 2)

G. Performance of Machine Learning

ML was employed to examine the performance of six algorithms (decision tree, random forest, logistic regression, KNN, SVM and Naïve Bayes) using eleven variables of which two were demographics, eleven were comorbidity, nineteen were laboratory test and vital sign. Confusion matrix (table 3) was applied for first step of evaluation, then the parameters evaluation was constructed using six algorithm using 10-foldcrossvalidation. The results shows that logistic regression and random forest more or less similar results, which were slightly better than those from the others classifiers in terms of accuracy, precision, sensitivity and F-Measure.(figure 2)

H. Discussion

The results of hospitalized patients with verified RT-PCR are reported in this study. The patients who had positive RT-PCR for SARS-CoV-2 died as a result of this investigation.

The algorithms' performance is evaluated using the Weka machine learning software. Table 3 and Figure 2 illustrate the results of the performance evaluation. With an ROC value of 0.817, the logistic regression algorithm proved to be the best algorithm on the dataset. As a result of the performance evaluation, it is shown that the Logistic regression algorithm can predict both death and alive patients in relation to COVID-19.

The multivariate approach demonstrated that demographics, clinical characteristics, comorbidities,

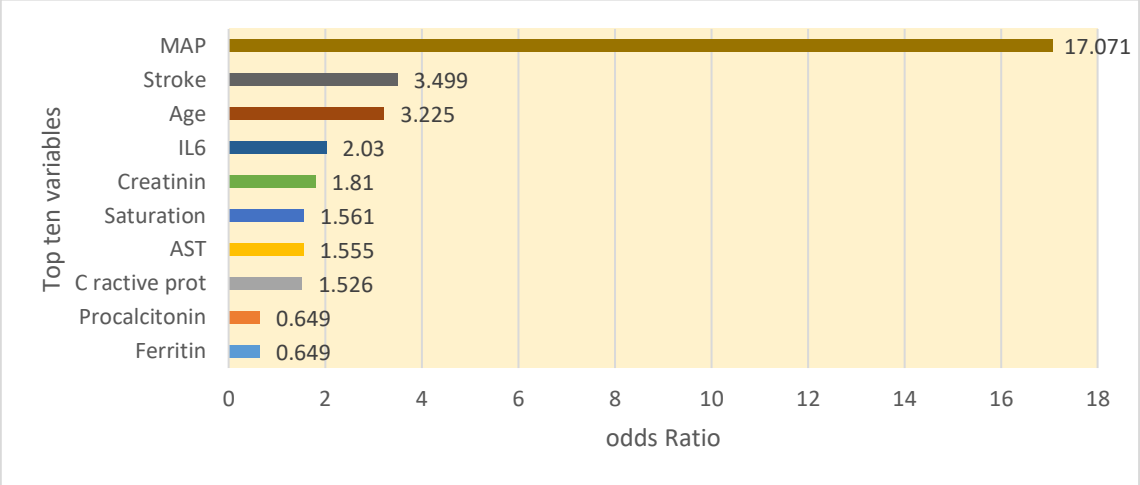

Figure 1. Illustrates the order of the top ten variables in terms of information gain in predicting mortality. MAP (Mean Artery Pressure) contributed the most to the prediction of mortality with a contribution value of 17.701 followed by a lower contribution by the stroke, age, IL6 , creatinine, saturation, AST, C-reactive protein, procalcitonin and ferritin (3.499,3.225, 2.03, 1.81, 1.561, 1.55, 1.526, 0.469 and 0.469 respectively).

Table 3.Confusion Matrix

Model	True Positive (TP)	False Positive (FP)	False Negative (FN)	True Negative (TN)
KNN	3217	346	327	821
Logistic Regression	3379	184	690	458
Random Forest	3384	179	734	414
SVM	3373	190	723	425
Naïve Bayes	3043	520	548	600
Decision Tree	3260	303	702	446

KNN :k-Nearest Neighbor ; SVM: Support vector machine

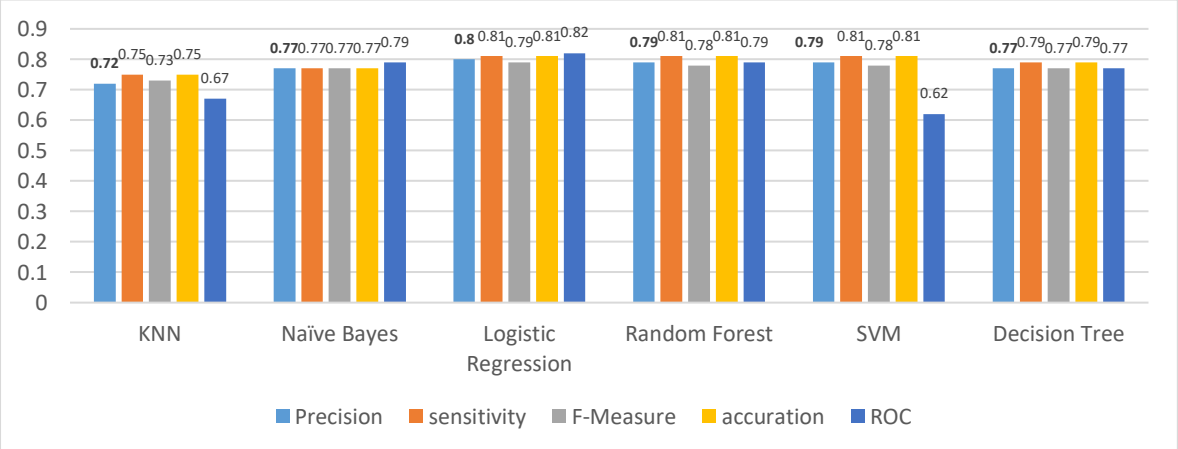

Figure 2. Performance of the machine learning algorithms, represents the performance of the classifiers used and shows that logistic regression and random forest more or less similar results, which were slightly better than those

from the others classifiers in terms of accuracy, precision, sensitivity and F-Measure and biochemical markers of patients can be applied to predict hospital mortality outcomes.

According to this study , for patients older than 65 years, the odds ratio of COVID-19 mortality is expected to be higher by 322.5%. Similar to our study older age was associated with increasing mortalities in studies including different populations [26][27]. Corcoles et al (2021) reported that getting older was related with an increased risk of death [28]. The mortality rate varied widely among people with different ages. For instance, the overall COVID-19 case fatality rate in China was estimated as 0.32% in those aged <60 years and significantly increased to 6.4% in those aged > 60 years [29] Among those aged 80 years and older, this rate was as high as 13.4%[29] Similarly, in Italy, the mortality rate increased from 0.3% among patients aged 30–39 years to 20.2% among those aged >80 years [30]

The sensitivity of older persons to severe COVID-19 disease and death is mostly related to immune system remodeling or immunosenescence, as well as the possibility of immunopathology in aged patients with decreased B and T lymphocyte capabilities.[31] Impaired type-1 interferon (IFN) response is associated with age-related changes in innate and adaptive immunity. Furthermore, many SARS-CoV-2 non-structural proteins suppress the type-1 IFN activity, resulting in a lower CD8+ T-cell response to viral infection.[32]

Inflammaging, chronic low-grade inflammatory phenotype (CLIP), serious viral infection, e.g., CMV, and other possible factors, such as smoking, decreased sex steroid secretion, and accumulated adipose tissue, all make a significant contribution to an unstable pro-inflammatory milieu in elderly adults, which increases further inflammatory reactions upon SARS-CoV-2 infection, and an exacerbated cytokine storm. It also has an impact on ACE-2 expression and viral infection. [31][33]

Our findings showed comorbidities including PVD, renal disease and stroke associated will increased mortality. Pre-existing chronic medical conditions were commonly related with higher disease severity and mortality in COVID-19 patients. Higher mortality was associated with older age, male gender, cardiac disease, lung disease other than bronchial asthma, chronic renal insufficiencies, chronic hepatic disease, malignancy, and dementia, according to a UK study.[34].

A systematic review and meta-analysis with eight studies with a total sample size of 19.399 COVID-19 patients reported that patients with COVID-19 who developed stroke had significantly higher mortality than those without stroke [35]. Li Zhang et al (2021) showed a meta-analysis with total 344,431 participants from 34 studies, chronic kidney disease (CKD) was related with an enhanced risk of progression and mortality in COVID-19 patients [36]. Then, another systematic review and meta- analysis that included 1576 hospitalized patients in China on risk for predicting mortality of COVID 19 patients demonstrated reported that hypertension, chronic respiratory conditions, and cardiovascular disease are associated with severe COVID-19[37]

Our study did not show Diabetes Mellitus as a factor associated with mortality. This is in contrast to other study. Barron et all (2020) reported a nationwide analysis in England show that type 1 and type 2 diabetes were both independently associated with a significant increased odds of in-hospital death with COVID-19 [38]. In human monocytes, elevated glucose levels directly increase SARS- CoV-2 replication, and glycolysis sustains SARS-CoV-2 replication via the production of mitochondrial reactive oxygen species and activation of hypoxia-inducible factor 1α [39]

In lab features, the deceased patients had persistent and more severe lymphopenia compared with recovered patients, and the lymphocyte count was selected and incorporated into the predictive model. Defects in function of lymphocytes are age-dependent and are associated with inflammation levels.[40] Pro-inflammatory markers like Ferritin and C Reactive Protein associated with higher mortality. This is consistent with various international studies [41][42].

This study suggest that elevated AST and C-Reactive Protein, lower lymphocytes was associated with higher mortality. Leon et al. (2021) applied the ML approach to cluster the patients with COVID into 3 groups including higher, moderate, and low rate of mortality, and showed that the higher AST, C-Reactive Protein and number of neutrophils were associated with a higher rate of mortality, respectively[43]. Significantly elevated C-Reactive Protein levels in the early stages of COVID-19 disease are associated with disease severity and the extent of

internal tissue abnormalities. [44] A significant increase in neutrophils with a decreased in lymphocytes, monocytes, and eosinophils may suggest clinical deterioration and an increasing risk of poor prognosis in COVID-19 patients [45]

In this study, higher ferritin and lower platelets count were associated with higher mortality. A systemic review and meta analysis including fifty-eight studies (44,305 patients) found similar results [46]. The inflammatory effects of a high ferritin, and low platelet counts could both precipitate or be the result of thrombotic and coagulopathic effects[47].

IL6 as pro inflammatory cytokines was associated with mortality in this study. Twelve studies involving in a systematic review and meta- analysis showed IL-6 is an appropriate predictor of severe infection in patients infected with COVID-19[48]. IL-6 levels have been identified a valid indication of disease severity and prediction of ventilatory support since the early stages of the COVID-19 pandemic [49] A major biomarker of inflammation is IL-6, a chemokine produced by macrophages and T cells to stimulate an immunological response. It is also consists of several cell types that respond to a variety of pathological circumstances such as inflammation, infection, and cancer[50]

Limitation

First, the data utilized in this investigation did not have information about radiology findings, which could be relevant as a predictive factors[51] [52]. Second, patients' treatments can have a significant impact on prognosis; therefore, we assumed that all of these patients were on standard therapy.

Strength

This study take a big data sample of N=4711 case to estimate the real distribution, however a very good estimation as N increases.

4. CONCULSIONS

This study shows that Age above 65 years and comorbidities including the presence of stroke and renal disease; Vital sign and laboratory test including lower Mean Arterial pressure and O_2 saturation, higher AST, IL6, troponin, ferritin; abnormally platelets score, sodium and creatinine were independent predictors of mortality in patients with COVID-19. We found that this factor could be combined in logistic regression machine learning model to create effective predictor of mortality with an ROC of 81.7%

The model will give insight on the population groups most impacted by the epidemic. Moreover, the study may be useful not just for COVID-19 prediction but also for other pandemics that the country has experienced.

References

1. R. Wölfel *et al.*, "Virological assessment of hospitalized patients with COVID-2019," *Nature*, vol. 581, no. 7809, pp. 465–469, 2020, doi: 10.1038/s41586-020-2196-x.
2. M. L. Jibril and U. S. Sharif, "Power of Artificial Intelligence to Diagnose and Prevent Further COVID-19 Outbreak: A Short Communication," pp. 0–3, 2020, [Online]. Available: http://arxiv.org/abs/2004.12463
3. 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, 2020, [Online]. Available: https://www.medrxiv.org/content/10.1101/2020.02.27.20028027v3%0Ahttps://www.medrxiv.org/content/10.1101/2020.02.27.20028027v3.abstract
4. R. Dolin and S. Perlman, "Novel Coronavirus From Wuhan , China 2019-2020," *Mand. Douglas, Bennett's Princ. Pract. Infect. Dis.*, p. January 31, 2020.
5. C. Rothe *et al.*, "Transmission of 2019-nCoV Infection from an Asymptomatic Contact in Germany," *N. Engl. J. Med.*, vol. 382, no. 10, pp. 970–971, 2020, doi: 10.1056/nejmc2001468.
6. L. J. Muhammad, M. M. Islam, S. S. Usman, and S. I. Ayon, "Predictive Data Mining Models for Novel Coronavirus (COVID-19) Infected Patients' Recovery," *SN Comput. Sci.*, vol. 1, no. 4, pp. 1–7, 2020, doi: 10.1007/s42979-020-00216-w.
7. L. J. Muhammad *et al.*, "Using Decision Tree Data Mining Algorithm to Predict Causes of Road Traffic Accidents, its Prone Locations and Time along Kano –Wudil Highway," *Int. J. Database Theory Appl.*, vol. 10, no. 1, pp. 197–206, 2017, doi: 10.14257/ijdta.2017.10.1.18.

8. A. Rahaman, M. Islam, R. Islam, M. S. Sadi, and S. Nooruddin, “Revue d ’ Intelligence Artificielle Developing IoT Based Smart Health Monitoring Systems : A Review,” vol. 33, no. 6, pp. 435–440, 2020.
9. M. M. Islam, A. Rahaman, and M. R. Islam, “Development of Smart Healthcare Monitoring System in IoT Environment,” *SN Comput. Sci.*, vol. 1, no. 3, pp. 1–11, 2020, doi: 10.1007/s42979-020-00195-y.
10. A. Field, “Logistic regression Logistic regression Logistic regression,” *Discov. Stat. Using SPSS*, pp. 731–735, 2012.
11. S. I. Ayon, M. M. Islam, and M. R. Hossain, “Coronary Artery Heart Disease Prediction: A Comparative Study of Computational Intelligence Techniques,” *IETE J. Res.*, vol. 0, no. 0, pp. 1–20, 2020, doi: 10.1080/03772063.2020.1713916.
12. M. M. Islam, H. Iqbal, M. R. Haque, and M. K. Hasan, “Prediction of breast cancer using support vector machine and K-Nearest neighbors,” *5th IEEE Reg. 10 Humanit. Technol. Conf. 2017, R10-HTC 2017*, vol. 2018-Janua, pp. 226–229, 2018, doi: 10.1109/R10-HTC.2017.8288944.
13. M. E. Mavroforakis and S. Theodoridis, “A geometric approach to support vector machine (SVM) classification,” *IEEE Trans. Neural Networks*, vol. 17, no. 3, pp. 671–682, 2006, doi: 10.1109/TNN.2006.873281.
14. L. J. Muhammad, A. Abba Haruna, I. A. Mohammed, M. Abubakar, B. G. Badamasi, and J. Musa Amshi, “Performance evaluation of classification data mining algorithms on coronary artery disease dataset,” *2019 9th Int. Conf. Comput. Knowl. Eng. ICCKE 2019*, no. 978, pp. 1–5, 2019, doi: 10.1109/ICCKE48569.2019.8964703.
15. N. S. Altman, “An introduction to kernel and nearest-neighbor nonparametric regression,” *Am. Stat.*, vol. 46, no. 3, pp. 175–185, 1992, doi: 10.1080/00031305.1992.10475879.
16. Onel. H, “Machine learning basics with the K-nearest neighbors algorithm, towards data science,” . *https //towar dsdat ascie nce. com/machi ne-learn ing-basic s-with-the-k-neare st-neigh bors-algor ithm-6a6e7 1d017 61*.
17. et al. Everitt BS, *Miscellaneous clustering methods in cluster analysis. 5th ed. Chichester:*, vol. 14, no. 1. 2011. doi: 10.1007/BF00154794.
18. B. Z. Yahaya, L. J. Muhammad, N. Abdulganiyyu, F. S. Ishaq, and Y. Atomsa, “An Improved C4.5 Algorithm using L’ Hospital Rule for Large Dataset,” *Indian J. Sci. Technol.*, vol. 11, no. 47, pp. 1–5, 2017, doi: 10.17485/ijst/2018/v11i47/132538.
19. M. Tayefi *et al.*, “The application of a decision tree to establish the parameters associated with hypertension,” *Comput. Methods Programs Biomed.*, vol. 139, pp. 83–91, 2017, doi: 10.1016/j.cmpb.2016.10.020.
20. J. Brownlee, “Statistical Methods for Machine Learning: Discover How to Transform Data into Knowledge with Python,” *Mach. Learn. Mastery*, 2019.
21. S. Dahlan, *Statistik Untuk kedokteran dan Kesehatan*. Jakarta: Epidemologi Indonesia, 2020.
22. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA data mining software,” *ACM SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, 2009, doi: 10.1145/1656274.1656278.
23. M. K. M. Bharti Suri, “‘Performance Evaluation of Data Mining Techniques’. In: Mishra D., Nayak M., Joshi A. (eds) Information and Communication Technology for Sustainable Development. Lecture Notes in Networks and Systems,” *Springer, Singapore*, vol. vol 9., 2019.
24. A. A. Haruna, L. T. Jung, V. Arputharaj, and L. J. Muhammad, “Incentive-Scheduling Algorithms to Provide Green Computational Data Center,” *SN Comput. Sci.*, vol. 2, no. 4, 2021, doi: 10.1007/s42979-021-00633-5.
25. L. Serviá *et al.*, “Machine learning techniques for mortality prediction in critical traumatic patients: anatomic and physiologic variables from the RETRAUCI study,” *BMC Med. Res. Methodol.*, vol. 20, no. 1, pp. 1–12, 2020, doi: 10.1186/s12874-020-01151-3.
26. M. Zunyou Wu, “Characteristics of and Important Lessons From theCoronavirus Disease 2019 (COVID-19) Outbreak in ChinaSummary of a Report of 72 314 Cases From the ChineseCenter for Disease Control and Prevention,” 2019, doi: 10.1001/jama.2020.2648.
27. S. Richardson *et al.*, “Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients Hospitalized With COVID-19 in the New York City Area,” *JAMA*, vol. 323, pp. 2052–2059, 2020, doi: 10.1001/jama.2020.6775.
28. A. Vila-corcoles *et al.*, “COVID19-related and all-cause mortality risk among middle-aged and older adults across the first epidemic wave of SARS- COV-2 infection : a population-based cohort study in Southern Catalonia , Spain , March – June 2020,” pp. 1–15, 2021.
29. R. Verity *et al.*, “Estimates of the severity of coronavirus disease 2019: a model-based analysis,” *Lancet Infect. Dis.*, vol. 20, no. 6, pp. 669–677, 2020, doi: 10.1016/S1473-3099(20)30243-7.
30. Graziano Onder,Giovanni Rezza, “Case-Fatality Rate and Characteristics of Patients Dying in Relation to COVID-19 in Italy,” 2020, doi: 10.1001/jama.2020.4683.
31. Y. Chen *et al.*, “Aging in COVID-19: Vulnerability, immunity and intervention,” *Ageing Res. Rev.*, vol. 65, no. October 2020, p. 101205, 2021, doi: 10.1016/j.arr.2020.101205.
32. S.-Y. Fung, K.-S. Yuen, Z.-W. Ye, C.-P. Chan, and D.-Y. Jin, “A tug-of-war between severe acute respiratory syndrome coronavirus 2 and host antiviral defence: lessons from other pathogenic viruses,” 2020, doi: 10.1080/22221751.2020.1736644.

33. M. J. Nasiri *et al.*, "COVID-19 Clinical Characteristics, and Sex-Specific Risk of Mortality: Systematic Review and Meta-Analysis," *Front. Med.*, vol. 7, no. July, pp. 1–10, 2020, doi: 10.3389/fmed.2020.00459.
34. C. M. Petrilli *et al.*, "Factors associated with hospital admission and critical illness among 5279 people with coronavirus disease 2019 in New York City: prospective cohort study(No Title)", doi: 10.1136/bmj.m1966.
35. A. Yassin *et al.*, "Mortality rate and biomarker expression within COVID-19 patients who develop acute ischemic stroke: a systematic review and meta-analysis," *Futur. Sci. OA*, vol. 7, no. 7, 2021, doi: 10.2144/fsoa-2021-0036.
36. L. Zhang, J. Hou, F. Z. Ma, J. Li, S. Xue, and Z. G. Xu, "The common risk factors for progression and mortality in COVID-19 patients: a meta-analysis," *Arch. Virol.*, vol. 166, no. 8, pp. 2071–2087, 2021, doi: 10.1007/s00705-021-05012-2.
37. J. Yang *et al.*, "Prevalence of comorbidities and its effects in coronavirus disease 2019 patients: A systematic review and meta-analysis," *Int. J. Infect. Dis.*, vol. 94, pp. 91–95, 2020, doi: 10.1016/j.ijid.2020.03.017.
38. E. Barron *et al.*, "Associations of type 1 and type 2 diabetes with COVID-19-related mortality in England: a whole-population study," *Lancet Diabetes Endocrinol.*, vol. 8, no. 10, pp. 813–822, 2020, doi: 10.1016/S2213-8587(20)30272-2.
39. S. Lim, J. Hyun Bae, H.-S. Kwon, and M. A. Nauck, "COVID-19 and diabetes mellitus: from pathophysiology to clinical management", doi: 10.1038/s41574-020-00435-4.
40. G. Chen *et al.*, "Clinical and immunological features of severe and moderate coronavirus disease 2019," *J. Clin. Invest.*, vol. 130, no. 5, pp. 2620–2629, 2020, doi: 10.1172/JCI137244.
41. J. bo Xu *et al.*, "Associations of procalcitonin, C-reaction protein and neutrophil-to-lymphocyte ratio with mortality in hospitalized COVID-19 patients in China," *Sci. Rep.*, vol. 10, no. 1, pp. 1–10, 2020, doi: 10.1038/s41598-020-72164-7.
42. D. Ji *et al.*, "Clinical characteristics predicting progression of COVID-19," *Lancet*, 2020.
43. J. Benito-León, M. D. Del Castillo, A. Estirado, R. Ghosh, S. Dubey, and J. I. Serrano, "Using unsupervised machine learning to identify age- And sex-independent severity subgroups among patients with COVID-19: Observational longitudinal study," *J. Med. Internet Res.*, vol. 23, no. 5, 2021, doi: 10.2196/25988.
44. C. Tan *et al.*, "C-reactive protein correlates with computed tomographic findings and predicts severe COVID-19 early," *J. Med. Virol.*, vol. 92, no. 7, pp. 856–862, 2020, doi: 10.1002/jmv.25871.
45. B. M. Henry, M. H. S. De Oliveira, S. Benoit, M. Plebani, and G. Lippi, "Hematologic, biochemical and immune biomarker abnormalities associated with severe illness and mortality in coronavirus disease 2019 (COVID-19): A meta-analysis," *Clin. Chem. Lab. Med.*, vol. 58, no. 7, pp. 1021–1028, 2020, doi: 10.1515/cclm-2020-0369.
46. E. H. Taylor *et al.*, "Factors associated with mortality in patients with COVID-19 admitted to intensive care: a systematic review and meta-analysis," *Anaesthesia*, vol. 76, no. 9, pp. 1224–1232, 2021, doi: 10.1111/anae.15532.
47. G. Xiang *et al.*, "The effect of coagulation factors in 2019 novel coronavirus patients: A systematic review and meta-analysis," *Medicine (Baltimore).*, vol. 100, no. 7, p. e24537, 2021, doi: 10.1097/MD.0000000000024537.
48. X. Liu, H. Wang, S. Shi, and J. Xiao, "Association between IL-6 and severe disease and mortality in COVID-19 disease: A systematic review and meta-analysis," *Postgrad. Med. J.*, pp. 1–9, 2021, doi: 10.1136/postgradmedj-2021-139939.
49. T. Herold *et al.*, "Elevated levels of IL-6 and CRP predict the need for mechanical ventilation in COVID-19," *J. Allergy Clin. Immunol.*, vol. 146, no. 1, pp. 128-136.e4, 2020, doi: 10.1016/j.jaci.2020.05.008.
50. C. Garbers, S. Heink, T. Korn, and S. Rose-John, "Interleukin-6: Designing specific therapeutics for a complex cytokine," *Nat. Rev. Drug Discov.*, vol. 17, no. 6, pp. 395–412, 2018, doi: 10.1038/nrd.2018.45.
51. M. Yuan, W. Yin, Z. Tao, W. Tan, and Y. Hu, "Association of radiologic findings with mortality of patients infected with 2019 novel coronavirus in Wuhan, China," *PLoS One*, vol. 15, no. 3, pp. 1–10, 2020, doi: 10.1371/journal.pone.0230548.
52. Z. Feng *et al.*, "Early prediction of disease progression in COVID-19 pneumonia patients with chest CT and clinical characteristics," *Nat. Commun.*, vol. 11, no. 1, pp. 1–9, 2020, doi: 10.1038/s41467-020-18786-x.