IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.125195

A critical review of dominant features used in machine learning approaches in COVID-19 Severity risk prediction

Ranjan Kumar[1], Vaibhav Maheshwari[2], Aaditya Tripathi[3]

Department of Computer Science, Aryabhatta College, University of Delhi, Delhi, India[1,2,3]

Abstract: COVID-19, which is caused by SARSCoV2 (Severe Acute Respiratory Syndrome Coronavirus 2), has wreaked widespread havoc in recent years. Almost immediately after the epidemic, experts at nearly every public health center began investigating all possible sources of the virus. The spread surged dramatically in the early phases, which became a big concern in the medical community. Researchers employed numerous ML approaches in the computer age to study the causes and patterns of diffusion. As a result, several studies utilizing machine learning and artificial intelligence have been conducted in this field. This article examines the essential elements in predicting severity due to COVID-19 and summarizes new studies on predicting severity using machine learning. This study's reviewed research papers were published using various search techniques.

Keywords: COVID-19 Pandemic, Machine Learning, Feature Selection, Severity Risk, ML Techniques

I. INTRODUCTION

COVID-19, caused by the SARSCoV2 virus, has had a devastating global impact since it was first reported in December 2019 in Wuhan, China. The World Health Organization declared it a pandemic on March 11, 2020 [1]. The virus has led to economic setbacks and strained healthcare systems worldwide. Multiple virus mutations have contributed to millions of deaths globally. Measures such as lockdowns, travel restrictions, and border closures have been implemented to control the virus's spread.With limited knowledge initially, medical professionals have turned to ML and AI methods to detect and analyze different diseases [2]. ML techniques have been used to assess the severity of the COVID-19 virus and develop preventive solutions. ML and AI offer more accurate predictions and adapt to changing data, making them valuable tools in combating the pandemic. Numerous research papers have been published utilizing ML techniques to predict mortality and analyze clinical and laboratory data for COVID-19 patients. These models focus on creating systems that can be learned and improved from examples without explicitly programming.

Today, ML is widely used in health care to provide risk assessments, analyze significant data points, predict results, and other applications. Researchers lean towards ML to predict and understand the severity of COVID-19 and find solutions to prevent the spread of the virus. AI and ML techniques prevent human intervention [3] and provide much better results based on the available data. AI tools continually adapt to changing data, analyze and make predictions, and find efficient solutions to the problems. Many research papers were published using ML techniques in COVID-19 research to discuss the morality [4] [5].

Norah A. et al. [6] searched from January 2020 to 2021 yielded 645 results and reviewed 52 studies, including 76 models that diagnose COVID-19 and predict the risk and severity of mortality among patients infected by Coronavirus. This study includes 18 results for COVID-19 diagnosis and 54 results for predicting the risk and severity of death. The study focuses on predicting the potential of various ML techniques as well as dominant features in the specified domain. As per the study, LR is the most widely used model for prognostic and diagnostic models, followed by XGBoost and SVM. The most commonly used attributes for the diagnosis of COVID-19 are older age, LDH, CRP, fever, decreased calcium, male gender, WBC, hypokalemia, eosinophil, CRP + LDH + ferritin, platelets, lymphocytopenia, hemoglobin, basophil, AST, leukocytes, increased creatinine, CD3, neutrophil count, INR, and monocytes percentage. Whereas common features for predicting COVID-19 mortality and severity include male gender, older age, red blood cells, ferritin, oxygen saturation, comorbidities, decreased calcium, lymphocytopenia, albumin, respiratory rate, neutrophil count, creatinine, increased BUN, LDH, AST, CKD, CRP, procalcitonin, bilirubin, D-dimer, IL-6, cTnI, and cTnT. The study also pointed out the main issues in various

© IJARCCE This work is licensed under a Creative Commons Attribution 4.0 International License

DOI: 10.17148/IJARCCE.2023.125195

articles, with the most common being "unbalanced data".

In this study, we worked exhaustively on 23 similar studies and summarized their findings. The study focuses on two key goals. To begin, we concentrated on several ML approaches for predicting the two primary characteristics. The first is the fatality rate, and the second is the severity. These criteria are critical because critically sick patients require optimal decision-making and priority at the treatment institution [7]. Second, we examined the key characteristics that were most important in the early prediction of severity (or mortality).

II. MACHINE LEARNING ALGORITHMS

Machine Learning is a branch of computer algorithms that could routinely adapt without following explicit instructions, using ML algorithms and statistical models to understand and draw results from patterns in the data. ML algorithm uses historical data as input to predict new output values. It is capable of performing classification and regression (supervised learning). It can be used in dimensionality reduction and clustering high-dimensional data sets (unsupervised learning). It can also learn from unlabeled data by processing the dataset, extracting features, and identifying patterns. Nowadays, ML is mainly used in clinical data to design, conduct, and analyze clinical trials [8]. It has been recognized as the most potent and promising analytical tool in healthcare [9].

The study considers 23 research papers that include various ML techniques. Some common of them are:

2.1 Decision Tree (DT) :

A Decision Tree (DT)[10] is a tree-structured model that continuously divides the data according to specific parameters. It is a Supervised ML technique where each internal node and leaf node represent a test on a feature and a class label. Gini Impurity is the standard method used in selecting the best features for splitting a tree. It can be used for both classification and regression. The formula for calculating Gini Impurity -

$$\textbf{Gini} = 1-\sum_{i=1}^{n}(\mathrm{pi})2$$

2.2 Random Forest (RF) :

This is a widely used ML technique for regression and classification problems. This supervised ML algorithm builds decision trees on various samples and makes a majority vote on classifications. The mean, on the other hand, is used for regression. Many decision trees are built, each node is split, and a subset of randomly selected attributes is retrieved for each tree. Bagging and boosting [11] are two methods used in ensemble techniques [12] (combining multiple models). RF [13] can also handle continuous variable datasets for regression and categorical variables for classification.

2.3 Support Vector Machine (SVM) :

This is a classification algorithm where each item in the dataset can be represented in n-dimensional space (where "n" is the number of features), and the SVM [14] finds a hyperplane that uniquely classifies the data points. The classification is then performed by detecting a hyperplane that uniquely identifies the class. Given a record X; W denotes the weight vector, and b indicates bias, then the hyperplane can be represented as -

$$\mathbf{W.X + b = 0}$$

2.4 Linear Regression (LR) :

LR [15] is a supervised ML algorithm generally used in regression problems. This technique is used to find the linear relation between x (input) and y (output) and predicts the dependent variable value based on the independent variable given. LR’s hypothesis function is given by-

$$\mathbf{y = \theta_1 + x.\theta_2}$$

We’ve ‘x’ input training data; and ‘y’ labels to data during model training. To get the best regression fit line, we find the best value for intercept ‘$\boldsymbol{\theta_1}$’ and x coefficient ‘$\boldsymbol{\theta_2}$’.

© IJARCCE This work is licensed under a Creative Commons Attribution 4.0 International License

2.5 K-Nearest Neighbor (KNN) :

KNN [16] is a supervised ML algorithm that learns by comparing specific new, not labeled data based on specific labeled input data. The closer the two data points are, the more similar they are expected to be. The Euclidean distance equation is commonly used to calculate the distance between two data points. Later, we will classify a group of similar "K" data points into a single class. Euclidean distance is given by -

$$\mathbf{d,y}\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

2.6 Naive Bayes Classifier (NB) :

This is an algorithm that uses probabilistic classification based on Bayes Theorem. In this algorithm, all the dataset features are assumed to be independent of each other. Dataset is classified into feature matrices and a class variable for every row of the feature matrix. Let's say a record X and 'n' number of classes Y_1, Y_2, Y_3, …, Y_n. Naive Bayes [17] classifier maximizes $P(Y_i \mid X)$ using the formula -

$$\mathbf{P(Y_i|X)} = \frac{P(X|Yi)P(Yi)}{P(X)}$$

2.7 Artificial Neural Network (ANN) :

ANN [18] is a computational model inspired by the brain's functioning. It is an oriented graph that consists of nodes, with each arc consisting of weights. Each connection weight is kept to increase the model's accuracy during the learning phase. This technique works efficiently in the case of large linear and non-linear data.

2.8 eXtreme Gradient Boosting (XGBoost) :

When implemented along Gradient Boosting, Decision Trees are called XGBoost [19]. In this technique, weights on nodes play an important role and help create DT in sequential form. After each iteration, DT improves itself based on the results achieved in the previous step. It speeds up the execution of the model manifold. This algorithm is used in various domains such as regression, classification, and user-defined problems related to predictions and rankings.

III. METHODOLOGY

In this review paper, we searched articles using Advanced search on Google Search Engine and Google Scholar database. Our search query focused on the following keywords: COVID-19, Severity Prediction, and Machine Learning. We also provided our search timeframe from Feb 2021 to Jan 2022 so that we could review the recent research and studies. Our review mainly focused on prediction using clinical data and laboratory data. It includes a few papers that used various scan results, such as CT scans, X Rays, etc., for the COVID-19 severity prediction. Our search query retrieved 107 results, out of which we extracted 23 studies. These studies used around 26 different models for prediction.

DOI: 10.17148/IJARCCE.2023.125195

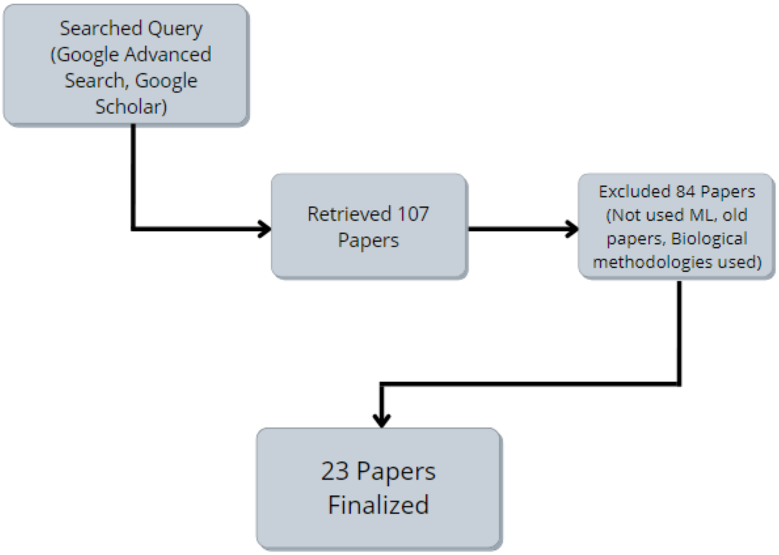
Searched Query (Google Advanced Search, Google Scholar)
Retrieved 107 Papers
Excluded 84 Papers (Not used ML, old papers, Biological methodologies used)
23 Papers Finalized

Fig.1 Study Selection Process

3.1 Critical Survey of Literature

Moulaei et al. [4] compare several ML models to predict mortality due to COVID-19. The study used 1500 patients' data(1386 survivors and 144 deaths) at the first time of admission. The study aims to predict COVID-19 mortality among patients early. The study employed Random Forest [13], XGBoost [19], KNN [16], MLP, LR [15], J48 DT [10], Naive Bayes [17]. As per the study, RF had better performance than other ML techniques with accuracy (95.03%), precision (94.23%), and ROC(99.02%). The study predicted six highly associated COVID-19 diagnosis biological variables: oxygen therapy, dyspnea, age, ICU admission, fever, and cough.

E. Jamshidi et al. [5] studied 797 COVID-19 patient data in Iran and the UK. The study aims to predict mortality among patients based on the lab data recorded on the day of the patient's admission to ICU. The study employed RF, LR, GB classifier, SVM, and ANN algorithms to compute the best result. According to the survey, Random Forest had the best performance with accuracy (79%) and precision (79%). The study also tested the model performance for some random validation sets, and the Random Forest classifier [13] predicted the outcome with specificity (75%) and sensitivity (70%). The study also identified the top 15 features out of 66 parameters: age, gender, BUN, creatinine, INR, albumin, etc.

Demichev et al. [20] studied a dataset of 168 patients with varying disease severity, out of which 50 most severely ill COVID-19 patients were used in model generation. The study aims to optimize the allocation and resources for patients admitted to the ICU. In the study, prominent features were RRT, ECMO, IL6, Comorbidity, Plasma Proteomes, Cell counts, enzyme tests, CRP, ferritin, thrombin(F2), and plasma kallikrein. The study employed the Charlson Comorbidity Index (CCI), APACHE-II, SOFA, and SVM as ML techniques to predict the survival of severely ill patients. As per the study, SVM [14] gives the best result compared with other techniques with AUROC (81%).

Hassan et al. [21] aimed to develop a model to predict the COVID-19 waiting times of patients on features like result interpretation, testing facility, and test date. The study employed Neural Network, SVM, KNN, Random Forest, Linear Regression, GB regression, and Decision Tree regression to predict the average waiting time with the highest accuracy. The study included 54730 patients recorded from 171 hospitals and 14 labs. The study used RMSE, MSE, and MAE to find the best ML technique. The smaller the RMSE value more accurately the model performed for the dataset. As per the study, the DT model [10] (RMSE value = 0.623) predicted the average waiting time with maximum accuracy.

Oyewola et al. [22] worked on the dataset of 5165 data samples with 13 features, out of which seven features were selected, namely sex, confirmed date, age, released date, country, symptom onset date, and the state. The study mainly focused on epidemiology and took the 'state' feature as the target column, which contained three classes (released, deceased, and isolated). The study employed classification techniques such as LR, DT, Bagging, Stochastic GB, BLSTM, KNN, and many

DOI: 10.17148/IJARCCE.2023.125195

more and achieved the best result in LR Technique with accuracy (82.77%). In the later phases, the study used AENN for improved outcomes and got the best result in Bagging Technique [11] and RF[13] model with balanced accuracy(100%).

Aurelle T. et al. [23] studied 5644 data of COVID-19 patients (who had gone through PCR tests). The study aims to identify whether a patient is COVID positive or not. The study divided 111 features into two sets: blood (features from the blood test) and viral (features from a virological test). The study employed five classification models - KNN, Bagging, Boosting, SVM, and Random Forest. As per the study, applying Bagging, AdaBoost classifier, and Random Forest leads to overfitting of the model, whereas SVM performed better compared to the other models with an accuracy (99.29%), sensitivity (92.79%), and specificity (100%). The study validated the SVM [14] model by applying it to another dataset, where accuracy (92.86%), sensitivity (93.55%), and specificity (90.91%) were obtained.

Zahra et al. [24] aim to develop a model to predict the risk of intubation in the COVID-19 patients. The study employed the Horse herd Optimization Algorithm (HOA) on a data set of 1225 patients with COVID-19 to extract the most important features (dry cough, high age, loss of smell, increased weight, fever, dyspnea, heart diseases, hypertension, CRP, ALT/ASP, SPO2, and leukocytosis) and ML techniques such as DT, SVM, MLP, and KNN were employed. Results were compared based on accuracy score. As per the study, the DT-based predictive model [10] achieved the best accuracy (93.8%) and precision (93.1%).

Singh V. et al. [25] studied the dataset containing records of 10,937 patients. The study used Deep Learning approaches for predictions and ML techniques such as SHAP values, LR, RF, and XGBoost. As per the study, the most important features were namely age, troponin-1, LDH, eosinophil, CRP, ferritin, lymphocyte, INR, creatinine, and d-DIMER and Random Forest [13] model achieved the best performance having AUC score (84%).

Altini et al. [26] studied 303 admitted patients diagnosed with COVID-19. The study aims to predict the risk of mortality among such patients. The study analyzed survival with Cox Regression and Kaplan–Meier curves to predict the most unfavorable features. The study employed ML techniques such as DT, Gaussian NB, SVM, KNN, RF, and AdaBoost. As per the study, the most preferable feature was C-Reactive Proteins, and the other preferable features were Ionized calcium max, AST min, CRP mean, CRP min, Total bilirubin min, Erythrocyte max, and the best ML model was Decision Tree [10] with an accuracy(88.52%) and precision(66.67%).

Patterson et al. [27] developed a severity prediction model for COVID-19 among patients. The dataset investigated for the study consists of 224 patients' records. The features mainly concentrated on blood test reports, and the dataset was balanced using SMOTE. The study categorized patients as acute, mild, or severe based on the frequent features IL-6, VEGF, IL-10, sCD40L, IFN-γ, and CCL4-MIP-1β. The ML technique which is used in the study was RF [13] which achieved accuracy (80%) and precision (62%).

Zhang et al. [28] aimed to classify patients as severe and non-severe. In this study, 422 COVID-19 patients were treated and classified based on several ML techniques, including RF, Gaussian NB, SVM, KNN, Logistic Regression, and ANN [18]. The data set was randomly divided into 80-20% for training and testing. RF model was used in feature importance analysis and identified T cell count, Helper T cell count, Cystatin C, Interleukin-6, and Suppressor T cell count as the essential features in classifying mild/severe. As per the study, the best classification model was Gaussian NB [17], with an AUC score of 0.90.

Ozan et al. [29] studied 166 patients' records and had 15 features, including symptoms present, demographic characteristics, disease histories, and blood test results. The study aimed to severity prediction of COVID-19 among patients. The study employed Hybrid ANN, SVM, and AdaBoost as ML techniques for classification. As per the study, the preferable features were age, CRP, respiratory rate, GCS, dyspnea, Spo2, DM, neutrophil, the first symptom of hospitalization, congestive heart disease, and fever, and the ANN [18] model performed better with accuracy (96%).

Alotaibi et al. [30] studied 80 patients' records with 52 features, out of which 32 features were selected during feature selection. The study aimed to predict the severity of patients using RBAs for feature selection. As ML classifiers, the study employed NN (Radial Basis and General Regression NN), SVM, NLP, and RF. The study also implemented the ML ensemble technique on Random Forest. The study selected dyspnea, age, loss of appetite, and oxygen saturation as the most important features. Random Forest with GentleBoost [13][11] performed better with accuracy (90.83%) and precision (90.83%).

Quiroz-Juárez et al. [31] aimed to identify high-risk(severe) patients. In the study, a dataset with 47,00,464 patients' who went under medical treatment in hospitals or clinics. The study includes 28 features, out of which 21 were selected for severity prediction as ICU, age, intubation, and hospitalization status were essential for severity prediction. As per the study, NN, LR, SVM, and KNN were some of the ML techniques used for classification, out of which NN [18] gave the best result with accuracy (93.5%).

Jia et al. [32] studied 3028 patients' records for risk prediction of COVID-19. In the study, the dataset included 82 features out of which 15 were selected as high-risk features -body-mass index (BMI), international normalized ratio, creatine kinase, LDH, prothrombin activity, D-dimer, prothrombin time (PT), hematocrit, heart rate, platelet count, magnesium, activated PTT, urine specific gravity, globulin and lymphocyte count (L%) As per the study, XGBoost and Logistic Regression were employed as the ML classifiers out of which XGBoost [19] resulted in higher accuracy(76.82%).

Emirena et al. [33] aimed to study the early prediction of death among patients. The study included 2782 patients, of which 2016 patients were from the first wave and 676 patients were included from the second wave of the COVID-19 outbreak in Italy. The dataset comprised 20 features, ten blood analytes (LDH, monocyte, %ferritin std, D-dimer, C-reactive protein, neutrophil/lymphocyte ratio, and lymphocyte %) Brescia chest X-ray score. As per the study, RF, GB, and LR were employed for the early prediction of in-hospital death. Random Forest [13] and SMOTE gave the best result with an AUC score of 0.97-0.98.

Mohsen et al. [34] used 520(270 discharged + 250 died) patients' data. In the study, 22 features were present, out of which the most prominent features were O_2 saturation, comorbidities, PCR, blood sugar, BUN, creatinine, LDH, SGOT, and WBC. The study aimed to predict mortality among patients using ML techniques such as NB, NN, AdaBoost, SVM, RF, DT, KNN, and ensembling techniques. The models that performed the best were NB and NN using an ensemble mechanism [17][18][12] with accuracy (80%) and an AUC score of 0.86.

Toshiki et al. [35] studied 1571(371 non-survivors) patients with COVID-19 from the Mount Sinai Health System. The study used two feature selection techniques, namely, LASSO and SHAP. The essential features extracted from each of the techniques were LASSO (diastolic blood race, coronary artery disease, pressure, blood urea nitrogen, heart rate, oxygen saturation, C-reactive protein, eGFR, systolic blood pressure, age, respiratory rate, hypertension, D-dimer, white blood cell count, hemoglobin, endotracheal intubation, and ICU admission) and SHAP (ICU admission, endotracheal intubation, age, oxygen saturation, hypertension, blood urea, and nitrogen). In the study, ML techniques used were LightGBM and LR, out of which LightGBM [11] had the best AUC (0.873).

H. Gull et al. [36] studied 992 records with nine symptoms, age, and gender as prominent features and ' condition ' feature as target attributes with categorical values(mild, moderate and severe). ML techniques employed in this study were LR, LDA, KNN, Gaussian NB, SVM, and RF. SVM [14] performed better than other techniques, with an accuracy of 60.27%.

Laatifi et al. [37] aimed to predict the severity-risk among patients. In the study, 337 COVID-19 patients (Cheikh Zaid Hospital). In the study, essential features were PLR, LDH, platelet count, D-DIMER, CRP, and comorbidities. Five feature selection techniques, namely ChiSquare, Mutual Information, ANOVA, PCA, and UMAP, were employed on ML techniques (LR, DT, GNB, SVM, and KNN). UMAP performed best and therefore was selected for classification on ML classifiers (XGBoost, AdaBoost, RF[13], ET, and KNN[16]) with 100% accuracy in each classifier except KNN(98%).

Nazir et al. [38] studied 287(243 survived and 44 deceased) COVID-19 patient records. The study's most prominent features were OX1, Temp2, Resp2, BP_Sys2, BP_Dsys2, Sum_sob, Pulse2, Age, fever, cough BP_Sys1, Temp1, Resp1, BP_Dsys1, OX2, MRN5, Gender, Sym_Others, Chr_dm. As per the study, three ML techniques were employed: LR, RF, and XGBoost. The study evaluated the three ML models with and without SMOTE analysis. Random Forest [13] with SMOTE achieved accuracy (95.2), sensitivity (0.949), and F1-score (0.955).

Xiong et al. [39] aimed to predict COVID-19 severity outcomes among patients during hospitalization. The study included 287(36.6%- severe 63.4%-non-severe) patients' records and 86 features in the initial dataset, later implemented LASSO CV, which reduced total features from 86 to 30. In the later phase, based on Spearman's Rank Correlation, literature evaluation, and expert opinions, the total number of features was reduced to 23. As per the study, the most important features were Chest CT scan, neutrophil to LDH, lymphocyte ratio, and D-dimer. Three ML models were employed: RF, SVM, and LR. RF [13]

DOI: 10.17148/IJARCCE.2023.125195

performed better than other models with AUC (0.970), sensitivity (96.7%), and accuracy (84.5%).

Mahdi et al. [40] studied 628 patients' records, and after exclusion of 136 patients' records, 492(65.8% severe and 34.2% non-severe) records were used for modeling. In the study, cough, dyspnea, and fever were the most observed and considered symptoms, and cardiovascular disease, diabetes mellitus, and hypertension were the most frequent comorbidities. The study employed NCA analysis on 37 biomarkers, out of which non-invasive features (SPO_2 and age) and laboratory biomarkers (BUN, PTT, and LDH) were the most prominent. The study made mortality predictions using three ML models: invasive, non-invasive biomarkers, and the joint (invasive + non-invasive) model. The joint model performed better than the other two, with an accuracy of 0.80 ± 0.03.

IV. DISCUSSIONS AND OBSERVATIONS

This paper demonstrates the extensive use of ML models in healthcare, specifically predicting the SARS-CoV-2 novel virus. During the global spread of the Corona Virus, it became essential for researchers to develop some mechanism to aid in the early prediction of the virus. The study aims to point out the most common and efficient ML models in classifying severity and mortality due to Coronavirus among patients. The study also aims to figure out the most prominent features of the same.

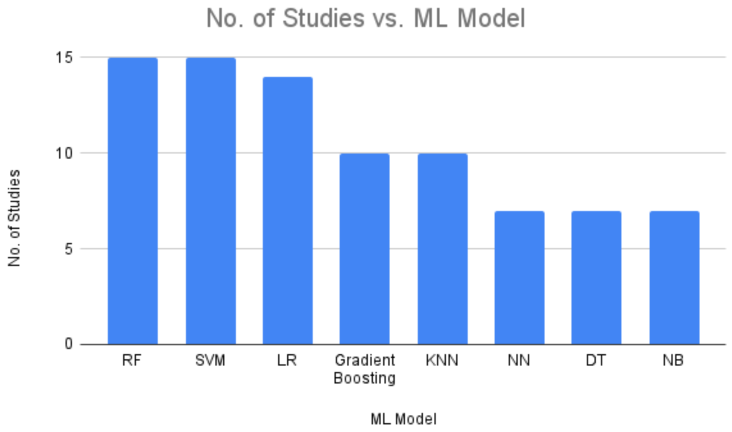

Fig.2 Frequency of ML models used in the reported studies

We observed that the most common Machine Learning models for classification and severity and mortality prediction due to COVID-19 are RF and SVM, followed by LR, then KNN, and Gradient Boosting. Fig. 2 shows the synopsis of the most frequent Machine Learning algorithms used in the studies used for reviewing. RF and SVM were commonly used ML models. RF exhibited promising predictive ability in terms of accuracy, reported as the best model by 7 out of 23 of the studies, followed by SVM. We mainly focused on laboratory and clinical data in the study, whereas it also included 1 study on chest X-ray data and another on CT scan data. We must note that, in both the studies, chest X-rays and CT scans, the best performing model was Random Forest.

Table1 summarizes studies being reviewed for severity and mortality risk prediction. In the table, rows represent the reviewed paper. In columns, we classify each paper based on severity definition, highest weighted features, ML techniques implemented, and their performance in the most efficient ML technique.

© IJARCCE This work is licensed under a Creative Commons Attribution 4.0 International License

DOI: 10.17148/IJARCCE.2023.125195

Table I Review table depicting the summary of various papers surveyed during the study

Author Name	Severity Definition	Highest weighted features	ML Approaches	Sample Size(No. Of Records)	Performance
Moulaei et al.[4]	Mortality	Dyspnea, age, ICU admission, fever, oxygen therapy, and cough	RF, XGBoost, KNN, MLP, LR, J48 DT, and Naive Bayes	1500 patient data(1386 survivors and 144 deaths)	RF Model: accuracy=95.03% precision=94.23% ROC=99.02%
E. Jamshidi. et al.[5]	ICU admission	gender, age, BUN, creatinine, respiratory disorders, mean corpuscular volume (MCV), red cell distribution width (RDW), lymphocyte count, and mean cell hemoglobin (MCH), along with a history of neurological, white blood cell count, cardiovascular segmented neutrophil count and albumin	RF, LR, GB classifier, SVM and ANN	797 data of patients	RF Model: accuracy=79% precision=79%
Demishev et al.[20]	Mortality	RRT, ECMO, IL6, Comorbidity, Plasma Proteomes, Cell counts, enzyme tests, CRP, ferritin, thrombin(F2), and plasma kallikrein.	Charlson Comorbidity Index(CCI), APACHE-II, SOFA, and SVM	168 patients	SVM Model: AUROC=81%
Hassan et al.[21]	Waiting time		NN, SVM, KNN, Random Forest, Linear Regression, GB regression, and DT regression	54730 patients	DT Model: RMSE value=0.623
Oyewola et al.[22]	N/A	country, state, sex, confirmed date, age, released date, and symptom onset date	LR, DT, Bagging, Stochastic GB, BLSTM, KNN, NB, RF, and SVM	5165 data samples	Bagging Technique and RF model after applying AENN: balanced accuracy=100
Aurelle T. et al.[23]	N/A	111 features: a.) blood tests results	KNN, Bagging, Boosting, SVM, and	5644 data of patients	SVM Model: accuracy=99.29%

DOI: 10.17148/IJARCCE.2023.125195

		b.) virological test results	RF		sensitivity=92.79% specificity=100% SVM Model(another data set): accuracy=92.86% sensitivity=93.55% specificity=90.91%
Zahra A. et al.[24]	N/A	C-reactive protein, oxygen saturation (SPO2), fever, dry cough, loss of smell, dyspnea, leukocytosis, cardiovascular diseases, high age, high weight, ALT/ASP, and hypertension	DT, SVM, MLP, and KNN	1225 COVID-19 patients	DT Model: accuracy=93.8% precision=93.1%
Singh V. et al.[25]	O2 Supplement, nasal cannula, BIPAP 3, SOFA b Score, transfer to hospice, mortality	age, troponin-1, LDH, eosinophil, CRP, ferritin, lymphocyte, INR, creatinine and d-DIMER	LR, RF, and XGBoost	10,937 patients	RF Model: AUC=84%
Altini et al.[26]	Admission to ICU	Erythrocyte max, Aspartate aminotransferase (AST) min, CRP, CRP min, CRP mean, Ionized calcium max, Total bilirubin min	DT, Gaussian NB, SVM, KNN, RF, and AdaBoost	303 admitted patients	DT Model: accuracy=88.52% precision=66.67%
Patterson et al.[27]	1. SpO2 < 94% 2. Heart Rate ≥ 125 bpm 3. Respiratory rate ≥ 30 bpm 4. (PaO2)/(FiO2) < 300mmHG 5. Lung infiltrate > 50%	IL-10, sCD40L, IL-6, VEGF, with IFN-γ and CCL4-MIP-1β	RF	224 patients	accuracy=80% precision=62%

DOI: 10.17148/IJARCCE.2023.125195

Zhang et. al.[28]	ICU and Mortality	Cystatin C, Helper T cell count, Suppressor T cell count, IL-6, T cell count	RF, Gaussian NB, SVM, KNN, LR and AN	422 patients	Gaussian NB Model: AUC=0.90
Ozan et al.[29]	Need of ventilation	Fever, congestive heart disease, first symptom, age, Spo2, respiratory rate, neutrophil, dyspnea, GCS, CRP, and DM	Hybrid ANN, SVM, and AdaBoost	166 patients	ANN Model: accuracy=96%
Alotaibi et.al.[30]	N/A	dyspnea, age, loss of appetite, and oxygen saturation	NN(Radial Basis and General Regression NN), SVM, NLP, and RF	80 patients	RF with GentleBoost Model: accuracy=90.83% precision=90.83%
Quiroz-Juárez et al.[31]	Mortality	ICU, age, intubation, and hospitalization status	NN, LR, SVM, and KNN	47,00,464 patients	NN Model: accuracy=93.5%
Jia et al.[32]	criteria defined National Health Commission of the PRC	body-mass index (BMI), international normalized ratio, creatine kinase, LDH, prothrombin activity, D-dimer, prothrombin time (PT), hematocrit, heart rate, platelet count, magnesium, activated PTT, urine specific gravity, globulin and lymphocyte count (L%)	XGB and LR	3028 patients	XGB Model: accuracy=76.82%
Emirana et al.[33]	Mortality	Brescia chest X-ray score and ten blood analytes(LDH, lymphocyte %, ferritin std, C-reactive protein, D-dimer, neutrophil/lymphocyte ratio, and monocyte %)	RF, GB, and LR	2782 patients comprised 2106 patients	RF with SMOTE Model: AUC=0.97-0.98
Mohsen et al.[34]	Mortality	O_2 saturation, comorbidities, PCR, blood sugar, BUN, creatinine, LDH, SGOT, and WBC	NB, NN, AdaBoost, SVM, RF, DT, KNN, and ensembling techniques	520(270 discharged + 250 died) patients	Ensembling(NB+NN) Model: accuracy=80% AUC=0.86
Toshiki et al.[35]	Mortality	LASSO(diastolic blood race, coronary artery disease, pressure, blood urea nitrogen, heart rate, oxygen saturation, C-reactive protein, eGFR, systolic blood pressure, age, respiratory rate, hypertension, D-dimer, white blood cell count, hemoglobin, endotracheal intubation, and ICU admission)	LightGBM and LR	1571 (371 non-survivors)	LightGBM Model: AUC=0.873

DOI: 10.17148/IJARCCE.2023.125195

		SHAP(ICU admission, endotracheal intubation, age, oxygen saturation, hypertension, blood urea, and nitrogen)			
H. Gull et al.[36]	N/A	age and gender Nine symptoms (runny-nose, fever, dry-cough, sore-throat, pains, diarrhea, difficulty-in-breathing, nasal-congestion, tiredness)	LR, LDA, KNN, Gaussian NB, SVM, and RF	992 records	SVM Model: accuracy=60.27%
Laatifi et al.[37]	N/A	PLR, LDH, platelet count, D-DIMER, C-reactive proteins, and comorbidities	XGBoost, AdaBoost, RF, ET, and KNN	337 COVID-19 positive patients	XGBoost, AdaBoost, RF, ET Model: accuracy=100%
Nazir et al.[38]	N/A	OX1, Temp2, Resp2, BP_Sys2, BP_Dsys2, Sum_sob, Pulse2, Age, fever, cough, BP_Sys1, Temp1, Resp1, BP_Dsys1, OX2, MRN5, Gender, Sym_Others, Chr_dm	LR, RF, and XGBoost	287(243 survived and 44 deceased)	RF with SMOTE Model: accuracy=95.2% sensitivity=94.9% F1-score=0.955
Xiong et al.[39]	N/A	neutrophil to lymphocyte ratio, Chest CT scan, D-dimer, and LDH	RF, SVM, and LR	287 patients	RF Model: accuracy=84.5% sensitivity=96.7% AUC=0.970
Mahdi et al.[40]	N/A	SPO$_2$, age, BUN, PTT and LDH	NCA Analysis	492(65.8% severe and 34.2% non-severe) records	accuracy=80±0.03%

Dimensionality Reduction is the typical strategy employed by most studies to decrease the complexity of their model, enhance the interpretability and increase the accuracy and precision of the model. It can be categorized into two, (i) Feature extraction generates a new and smaller set of features that captures all the valuable information, and (ii) Feature selection creates a subset of original features. We observed in the study that Laatifi et al. [37] used PCA and UMAP to reduce the number of components by combining features, Toshiki et al. [35], Singh V. et al. [25] used SHAP values to extract the essential features. However, most of the studies employed general variance importance coefficient factors for feature selection. Fig. 3 summarizes the reviewed studies' common feature selection and extraction techniques and their frequency count.

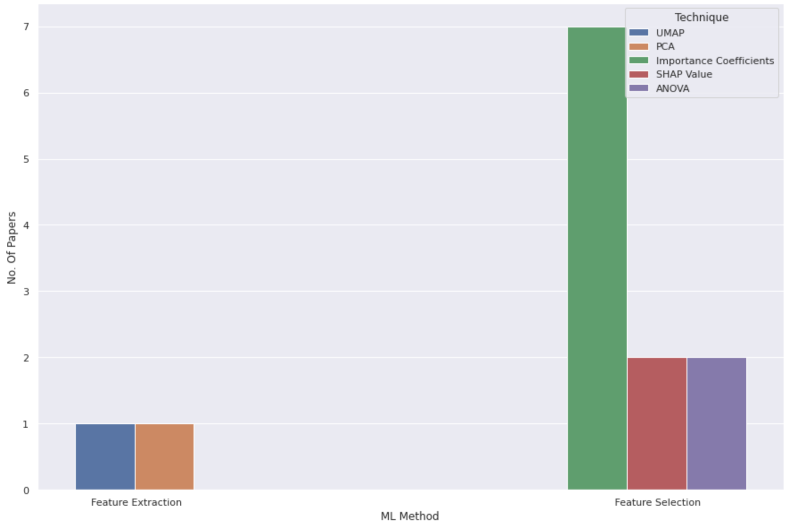

Fig.3 Major Feature Selection and Extraction methods used in the reported studies.

This study also aims to find the most relevant features for severity and mortality prediction among patients diagnosed with COVID-19. In the study, we try to observe the frequently occurring features among the papers, and Fig. 4 summarizes the frequent features. Age, CRP, LDH, O2 saturation, D-dimer, Lymphocyte, Dyspnea, BUN, Creatinine, Ferritin, and IL-6. ‘Age’ is the most prominent parameter in severity and mortality prediction. As per the study, patients of higher age are prone to get infected severely and therefore need to be prioritized in medical facilities. C-reactive protein is the second most important feature and can help in the early detection of severity. CRP level rises rapidly among patients suffering from COVID-19 and can be a prominent detector.

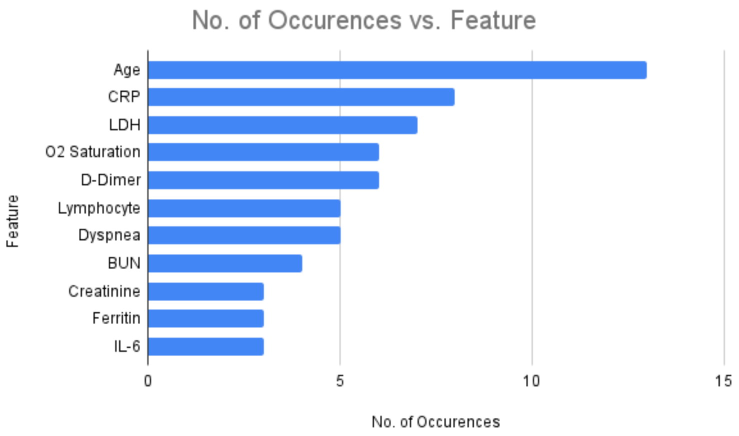

Fig.4 Frequency of prominent features selected by the reported studies

DOI: 10.17148/IJARCCE.2023.125195

V. CONCLUSION

This study surveyed existing articles related to mortality and severity prediction among patients diagnosed with COVID-19. We focused on studies that included clinical datasets and utilized Machine Learning techniques for classification and prediction. The study focuses on two broad categories, (i) to find out the best ML models for severity and mortality prediction and (ii) to extract the most frequent diagnostic and prognostic features from the articles. After studying 23 recent articles, we observed that the most frequently used ML models belong to the category of supervised machine learning algorithms. We also figured out various feature selection techniques that helped reduce the complexity of the dataset and increase the accuracy of the classification model. Feature selection techniques extract prominent columns from the dataset, thereby reducing the size of the data and reducing computation cost and irrelevant columns.

Although various researchers in this domain have done numerous studies, most of the work was still experimental. The developed models haven't been implemented and utilized in the real world. A few weaknesses of the studies performed in this field were, (i) dataset used in most of the studies were too small (or insufficient) for analysis; for an extensive study like COVID, such small datasets could lead to inappropriate results, and therefore, the model fails. (ii) demographically concentrated datasets were used in the studies, limiting the model to a specific region. Although the model performed well for a specific region, this does not guarantee the same accuracy for a different region (clinical features' importance might vary from region to region). Therefore, cannot be utilized globally. (iii) very few studies have classified patients with a distinct degree of severity due to coronavirus using some decision-making procedures. However, the prediction of mortality risk and severity among COVID-19 patients would be worth investigating further.

REFERENCES

[1] K.G. Andersen, A. Rambaut., W.I. Lipkin et al., “The proximal origin of SARS-CoV-2”, *Nat Med*, Vol. 26, pp. 450–452, 2020. https://doi.org/10.1038/s41591-020-0820-9

[2] R. Kumar, S. Maheshwari, A. Sharma et al., “Ensemble learning-based early detection of influenza disease.” *Multimed Tools Appl*., 2023. https://doi.org/10.1007/s11042-023-15848-2

[3] S. Dargan, M. Kumar, M.R. Ayyagari et al., “A Survey of Deep Learning and Its Applications: A New Paradigm to Machine Learning”, *Arch Computat Methods Eng*, Vol. 27, pp. 1071–109, 2020. https://doi.org/10.1007/s11831-019-09344-w

[4] K. Moulaei, M. Shanbehzadeh, Z. Mohammadi-Taghiabad et al., “Comparing machine learning algorithms for predicting COVID-19 mortality”, *BMC Med Inform Decis Mak*,Vol. 22(2), 2022. https://doi.org/10.1186/s12911-021-01742-0

[5] E. Jamshidi, A. Asgary, N. Tavakoli, A. Zali, S. Setareh, H Esmaily, S.H. Jamaldini, A. Daaee, A. Babajani, M.A. Sendani Kashi, M. Jamshidi, S. Jamal Rahi and N. Mansouri, “Using Machine Learning to Predict Mortality for COVID-19 Patients on Day 0 in the ICU”, Vol. 3:681608, 2022. https://doi.org/10.3389/fdgth.2021.681608

[6] Norah Alballa and Isra Al-Turaiki, “Machine learning approaches in COVID-19 diagnosis, mortality, and severity risk prediction: A review”, *Informatics in Medicine Unlocked*, Vol.24:100564,2021. https://doi.org/10.1016/j.imu.2021.100564

[7] X. Jiang, M. Coffee, A. Bari., J. Wang, X. Jiang.,J. Huang and Y. Huang, “Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity”, *Computers, Materials & Continua*, Vol. 63(1), pp. 537-551, 2021. https://doi.org/10.32604/cmc.2020.010691

[8] E.H. Weissler, T. Naumann, T. Andersson et al., “The role of ML in clinical research: transforming the future of evidence generation”, *Trials*, Vol. 22: 537, 2021. https://doi.org/10.1186/s13063-021-05489-x

[9] M.A. Jabbar, S. Samreen and R. Aluvalu, “The Future of Healthcare: Machine Learning”, *International Journal of Engineering and Technology*, Vol. 7, pp. 23-25, 2018. https://doi.org/10.14419/ijet.v7i4.6.20226

[10] J.R. Quinlan, “Induction of decision trees”, *Mach Learn*, Vol. 1, pp. 81–106, 1986. https://doi.org/10.1007/BF00116251

[11] P. Bühlmann, “Bagging, Boosting and Ensemble Methods”, *Handbook of Computational Statistics*, 2012. https://doi.org/10.1007/978-3-642-21551-3_33

[12] T.G. Dietterich, Ensemble Methods in Machine Learning, In: Multiple Classifier Systems, MCS 2000, Lecture Notes in Computer Science 1857, Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-45014-9_1

[13] L. Breiman, “Random Forests”, *Machine Learning*, Vol. 45 pp. 5–32, 2001. https://doi.org/10.1023/A:1010933404324

[14] N. Cristianini and E. Ricci, “Support Vector Machines”, *In: Kao MY. (eds) Encyclopedia of Algorithm*, Springer (2008), Boston,MA. https://doi.org/10.1007/978-0-387-30162-4_415

© IJARCCE This work is licensed under a Creative Commons Attribution 4.0 International License

DOI: 10.17148/IJARCCE.2023.125195

[15] R.R. Hocking, “Developments in Linear Regression Methodology: 1959-1982”, *Technometrics*, Vol. 25(3), pp. 219–230, 1983. https://doi.org/10.2307/1268603

[16] G. Guo, H. Wang, D.Bell, Y. Bi and K.Greer, “KNN Model-Based Approach in Classification”, *In: Meersman, R., Tari, Z., Schmidt, D.C. (eds) On The Move to Meaningful Internet Systems, Lecture Notes in Computer Science*, vol 2888. Springer, Berlin, Heidelberg.
https://doi.org/10.1007/978-3-540-39964-3_62

[17] Vikram kumar, B. Vijaykumar and Trilochan, “Bayes and Naïve Bayes Classifier”, 2014.
https://doi.org/10.48550/arXiv.1404.0933

[18] S.A. Kustrin and R. Beresford, “Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research”. *Journal of pharmaceutical and biomedical analysis*, 2000.
https://doi.org/10.1016/S0731-7085(99)00272-1

[19] Tianqi Chen and Carlos Guestrin, “XGBoost: A Scalable Tree Boosting System”, *In Proceedings of the 22nd ACM SIGKDD International conference on Knowledge Discovery and Data Mining(KDD2016)*,pp.785–794. https://doi.org/10.1145/2939672.2939785

[20] V. Demichev, P. Tober-Lau, T. Nazarenko, O. Lemke, S. Kaur Aulakh et al., “A proteomic survival predictor for COVID-19 patients in intensive care”, *PLOS Digital Health*, Vol. 1(1):e0000007,2022. https://doi.org/10.1371/journal.pdig.0000007

[21] H. Hijry, R. Olawoyin, W. Edwards, G. McDonald, D. Debnath and Y. Al-Hejri, “Predicting Average Wait-Time of COVID-19 Test Results and Efficacy Using Machine Learning Algorithms”, *International Journal of Industrial Engineering and Operations Management*, Vol. 3(2), pp. 75–88, 2021.
https://doi.org/10.46254/j.ieom.20210202

[22] D.O. Oyewola, E.G. Dada, S. Misra and R. Damaševičius, “Predicting COVID-19 Cases in South Korea with All K-Edited Nearest Neighbors Noise Filter and Machine Learning Techniques”, *Information*, Vol. 12:528, 2021. https://doi.org/10.3390/info12120528

[23] Aurelle Tchagna Kouanou, Thomas Mih Attia, Cyrille Feudjio, Anges Fleurio Djeumo, Adèle Ngo Mouelas, Mendel Patrice Nzogang, Christian Tchito Tchapga and Daniel Tchiotsop, “An Overview of Supervised Machine Learning Methods and Data Analysis for COVID-19 Detection”, *Journal of Healthcare Engineering*, Article ID 4733167, 2021. https://doi.org/10.1155/2021/4733167

[24] Zahra Asghari Varzaneh, Azam Orooji, Leila Erfannia and Mostafa Shanbehzadeh, A new COVID-19 intubation prediction strategy using an intelligent feature selection and K-NN method, Informatics in Medicine Unlocked 28:100825 (2022). https://doi.org/10.1016/j.imu.2021.100825

[25] V. Singh, R. Kamaleswaran, D. Chalfin, A. Buño-Soto, J. San Roman, E. Rojas-Kenney , R. Molinaro, S. von Sengbusch, P. Hodjat, D. Comaniciu and A. Kamen, “A deep learning approach for predicting severity of COVID-19 patients using a parsimonious set of laboratory markers”, *iScience*, Vol. 24(12):103523,2021. https://doi.org/10.1016/j.isci.2021.103523

[26] N. Altini, A. Brunetti, S. Mazzoleni, F. Moncelli et al., “Predictive Machine Learning Models and Survival Analysis for COVID-19 Prognosis Based on Hematochemical Parameters”, *Sensors (Basel)*, Vol. 21(24):8503, 2021. https://doi.org/10.3390/s21248503

[27] B.K. Patterson, J. Guevara-Coto, R. Yogendra, E.B. Francisco et al., “Immune-Based Prediction of COVID-19 Severity and Chronicity Decoded Using Machine Learning”, *Front. Immunol*, Vol. 12:700782, 2021.
https://doi.org/10.3389/fimmu.2021.700782

[28] R. Zhang, Q. Xiao, S. Zhu, H. Lin and M. Tang, “Using different machine learning models to classify patients into mild and severe cases of COVID-19 based on multivariate blood testing”, *J Med Virol*, Vol. 94, pp. 357- 365, 2021.
https://doi.org/10.1002/jmv.27352

[29] Ozan Kocadagli, Arzu Baygul, Neslihan Gokmen, Said Incir and Cagdas Aktan, “Clinical prognosis evaluation of COVID-19 patients: An interpretable hybrid machine learning approach”, *Current Research in Translational Medicine*, Vol. 70(1):103319,2021. https://doi.org/10.1016/j.retram.2021.103319

[30] A. Alotaibi, M. Shiblee and A. Alshahrani, “Prediction of Severity of COVID-19-Infected Patients Using Machine Learning Techniques”, *Computers*, Vol. 10(31), 2021.
https://doi.org/10.3390/computers10030031

[31] M.A. Quiroz-Juárez, A. Torres-Gómez, I. Hoyo-Ulloa, RdJ. León-Montiel and A.B. U’Ren, “Identification of high-risk COVID-19 patients using machine learning”, *PLOS ONE*, Vol. 16(9): e0257234, 2021.
https://doi.org/10.1371/journal.pone.0257234

[32] L. Jia, Z. Wei, H. Zhang et al., “An interpretable machine learning model based on a quick pre-screening system enables accurate deterioration risk prediction for COVID-19”, *Sci Rep*, Vol. 11: 23127, 2021.
https://doi.org/10.1038/s41598-021-02370-4

© IJARCCE This work is licensed under a Creative Commons Attribution 4.0 International License

DOI: 10.17148/IJARCCE.2023.125195

[33] Emirena Garrafa ,Marika Vezzoli, Marco Ravanelli, Davide Farina et al., “Early prediction of in-hospital death of COVID-19 patients: a machine-learning model based on age, blood analyses, and chest x-ray score”, *eLife*, Vol. 10:e70640, 2021. https://doi.org/10.7554/eLife.70640

[34] Mohsen Tabatabaie , Amir Hossein Sarrami , Mojtaba Didehdar , Baharak Tasorian et al., “Accuracy of Machine Learning Models to Predict Mortality in COVID-19 Infection Using the Clinical and Laboratory Data at the Time of Admission”, *Cureus*, Vol. 13(10): e18768, 2021. https://doi.org/10.7759/cureus.18768

[35] Toshiki Kuno, Matsuo So, Masao Iwagami, Mai Takahashi and Natalia N. Egorova, “The association of statins use with survival of patients with COVID-19”, *Journal of Cardiology*, Vol.79,pp.494-500,2022. https://doi.org/10.1016/j.jjcc.2021.12.012

[36] H. Gull, G. Krishna, M. I. Aldossary and S. Z. Iqbal, “Severity Prediction of COVID-19 Patients Using Machine Learning Classification Algorithms: A Case Study of Small City in Pakistan with Minimal Health Facility”, *in IEEE 6th International Conference on Computer and Communications (ICCC)*, pp. 1537-1541, 2020. https://doi.org/10.1109/ICCC51575.2020.9344984

[37] M. Laatifi, S. Douzi, A. Bouklouz et al., “Machine learning approaches in Covid-19 severity risk prediction in Morocco”, *J Big Data*, Vol. 9 (5), 2022. https://doi.org/10.1186/s40537-021-00557-0

[38] Shah Nazir, Sumayah S. Aljameel, Irfan Ullah Khan, Nida Aslam, Malak Aljabri and Eman S. Alsulmi, “Machine Learning-Based Model to Predict the Disease Severity and Outcome in COVID-19 Patients”, *Scientific Programming*, Article Id: 5587188, 2021. https://doi.org/10.1155/2021/5587188

[39] Y. Xiong, Y. Ma, L. Ruan et al., “Comparing different machine learning techniques for predicting COVID-19 severity”, *Infect Dis Poverty*, Vol. 11(19), 2022. https://doi.org/10.1186/s40249-022-00946-4

[40] M. Mahdavi, H. Choubdar, E. Zabeh, M. Rieder, S. Safavi-Naeini. et al., “A machine learning based exploration of COVID-19 mortality risk”, *PLOS ONE*, Vol. 16(7): e0252384, 2021. https://doi.org/10.1371/journal.pone.0252384

© IJARCCE This work is licensed under a Creative Commons Attribution 4.0 International License