Engineered Science
DOI: https://dx.doi.org/10.30919/es1169

# Predicting Pandemic Fatality Based on Supervised Machine Learning Methods

Jamil Al Shaqsi,[1,*] Mohamed Borghan,[2] Osama Drogham,[3] Gholam Hossein Roshani[4,*] and Salim Al Whahaibi[5]

## Abstract

This study aimed to conduct intensive experiments on predictive machine learning models to predict the parameters behind death due to the coronavirus (COVID-19). Datasets were obtained from Mexican Government and Various machine learning algorithms were used to detect the hidden patterns from the obtained datasets. Feature selection technique was used to optimize the precision of the supervised machine learning algorithms. Given a large employed sample (N = 67300 patients), the 10-fold Cross-Validation technique was used to validate the experimental results. Several metrics were used to measure the algorithms' accuracy and then compare the outcomes of the conducted experiments. Based on the analysis of multiple algorithms, it has been found that J48 algorithm outperformed other algorithms' classification performance. More importantly, it has been confirmed that some parameters significantly contribute to the death of infected patients. Oxygen saturation, pneumonia, and age are the leading predictors for predicting mortality. The obtained results will help in minimizing the risk of death by re-structuring the treatment protocol.

*Keywords:* Machine learning; Covid 19; Feature selection; Knowledge discovery.
Received: 06 March 2024; Revised: 16 May 2024; Accepted: 06 June 2024.
Article type: Research article.

## 1. Introduction

The respiratory syndrome coronavirus 2 (aka SARS-CoV-2) was first reported in December 2019 in China. The global spread of the virus alerted the global community. As a result, in March 2020, the World Health Organization (WHO) declared an outbreak of the infectious virus as a global pandemic.[1] By Feb 2023, it has been reported that the cases of the positive infection exceeded 755 million, among them more than 6.8 million deaths.[2] COVID-19 is a viral infection responsible for inducing severe acute respiratory syndrome,[1] fever, dry cough, fatigue, dyspnea, dizziness, abdominal pain, pharyngalgia, and other symptoms.[3] Rahimi *et al.*[4] conducted a comprehensive review on COVID-19 forecasting models aiming to spot the light on the machine learning approaches capable of predicting the global outbreak. It has been found that the various is transmitted by direct contact, airborne, handshaking, droplet, and sneezing. [5-7] According to Dhouib W. *et al.*[8] the incubation period of COVID-19 is estimated to be 10.3 to 16 days. The unnoticed virus incubation among asymptomatic cases makes this infectious disease more uncertain.[8] Cenggoro and Pardamean[9] and Badiola-Zabala *et al.*[10] conducted a systematic literature reviews to address the best possible method to diagnose COVID-19 more quickly and accurately to minimize the spread of the diseases. The study's findings revealed that chest X-rays were the preferred data source for categorizing COVID-19 cases, indicating their widespread use in medical imaging for diagnosing the disease.[9] Additionally, the study employed transfer learning techniques to leverage knowledge from one domain or task to improve performance in another domain or task.[9] This approach highlights the innovative use of machine learning methodologies to enhance the accuracy and efficiency of COVID-19 categorization using chest X-ray

[1] *Department of Information Systems, Sultan Qaboos University, P.O. Box 20, PC 123, Muscat, 999046, Oman.*
[2] *Department of Pre-medicine, National University of Science and Technology, Muscat, 999046, Oman.*
[3] *Department of Information and Communication Technology Systems, Al-Balqa Applied University, Al-Salt,999045, Jordan.*
[4] *Electrical Engineering Department, Kermanshah University of Technology, Kermanshah 6715685420, Iran.*
[5] *Falha Medical Solutions, Muscat, 999046,Oman.*
**Email:* alshaqsi@squ.edu.om (J. Al Shaqsi), hosseinroshani@kut.ac.ir (G. H. Roshani)

© Engineered Science Publisher LLC 2024

data. Prince *et al*.[11] utilized machine learning algorithms to diagnose medical images. In aim of the study was to implement texture descriptors on X-rays of COVID-19 patients' lungs and utilize the extracted features in frameworks designed to accurately assess COVID-19 patients. The experimental results demonstrated that combining different types of texture descriptors led to increased accuracy, resulting in improved metrics for detecting and diagnosing COVID-19.[12]

In the initial phase of the COVID-19 pandemic, there was a shortage of vaccines, antiviral medications, and ventilators. Consequently, countries implemented quarantine measures and other non-pharmaceutical strategies to control the spread of the virus. The various strains of COVID-19 reduce the efficacy of human antibodies in combatting the virus, allowing the virus to evade the immune system both before and after vaccination.[13] Although some individuals have recuperated from COVID-19 without medical intervention, there are cases, as noted by the World Health Organization,[7] where individuals have experienced severe impacts. These severe cases worsen due to several reasons. Examples of these reasons are the weak the immune system and the patient's medical history with conditions including chronic diseases such as hypertension, diabetes, and cardiovascular and respiratory diseases.[7] While some countries like Oman have ceased tracking the COVID-19 case counts, Mexico continues to witness a steady rise in active cases. As of September 12, 2023, the total has surpassed 7.5 million cases, according to WorldMeter COVID-19 Data,[14] see Fig. 1.

ML is an application of Artificial Intelligence (AI) to uncover new knowledge by exploring patterns that emerge in a given dataset. Finding relationships between parameters (variables) is at the heart of ML approach. Once relationships are identified, the prediction of possible outcomes is made possible. Data from the medical field are reach of volumes, variety, and variability.[15,16] This fact allowed ML possible for scholars interested in uncovering hidden patterns in medical data.[17-19] Several works have been done concerning health outcome prediction.[20-25] For example, classification and clustering analysis with ML is employed in protein structure, drug grouping, gene interaction, and cancer prediction.[17,26-30] The objective of this research is to carry out in-depth experiments utilizing predictive machine learning models for the purpose of developing and comparing prognostic tools aimed at identifying the factors contributing to COVID-19 related fatalities.[31] Due to the complexity of COVID-19, the available studies have some limitations[21,22,32] Mainly, the limitations are summarized in the following points: (1) the size of the dataset and the time frame of the data collection,[25,33] (2) the dataset is imbalance, (3) one single algorithm is used in the experiments,[23] (4) no comparison with other ML algorithms to confirm the optimal classification accuracy,[23] (5) the used dataset is limited to a clinical unstructured data,[32] (6) datasets were limited to the images and there was no consideration for the blood parameters in their experiments[11,24,34] (7) studies some were restricted to the spread of COVID-19 rather than the diagnosis,[35] Although, recently, few vaccines have been released and around 13,228,728,467 patients have been vaccinated worldwide,[2] identifying the infected blood parameters is essential as COVID-19 variants might trigger new variants in the future. The findings of this study will help the clinical diagnosis in fine tuning the abnormal parameters, which are outside the range of the appropriate threshold, in order to revitalize the immune system.

This paper is organized by offering a section on materials and the methods. The second section highlights the material and methods. The machine learning analysis will be presented in the third section. In this section, evaluation methods, covering accuracy, precision, recall, F-Measure, and ROC Curve meters are presented. Then the light will be spotted

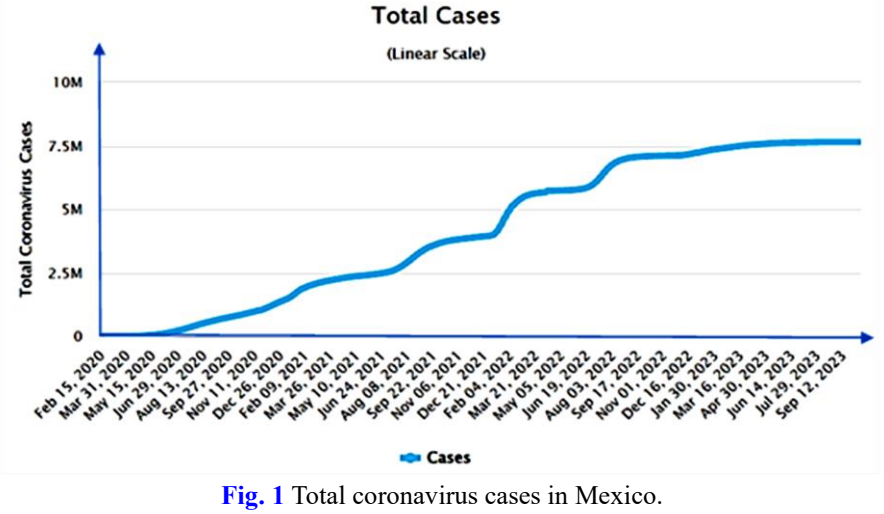


**Fig. 1** Total coronavirus cases in Mexico.


© Engineered Science Publisher LLC 2024

on the results of the machine learning. After that, the study will present the discussion followed with the implications for both theory and practice, Finally, yet importantly, the future work will be presented and the study's limitations will be highlighted.

## 2. Material and methods

To carrying out the machine learning the experiments, this research adopts the Knowledge Discovery (KDD) methodology.[10] cc helps researchers to uncover hidden patterns with reliable and valid insights, given the input data is of high quality. KDD can perform different analyses such as classification and clustering problems.[36] Fig. 2 illustrate the main stages of the KDD methodology are: defining the problem, gathering the needed resources, performing appropriate cleansing, conducting a pre-processing, selecting the appropriate algorithms (data mining), evaluating of the generated results, interpreting of the obtained results, and exploiting of the results.

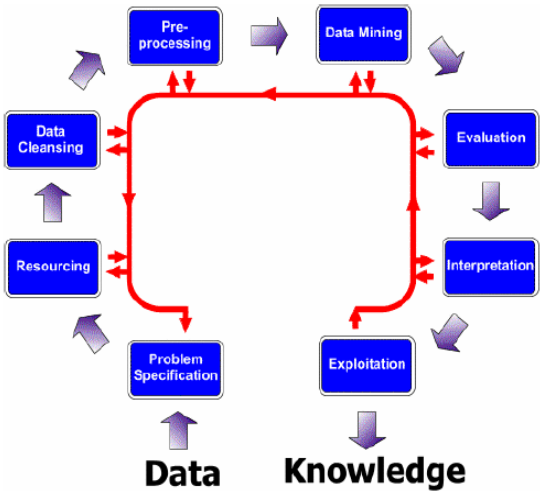


**Fig. 2** Stages of the KDD Methodology.

### 2.1 Problem specification and resourcing

The initial phase of problem specification sets the boundaries of the KDD problem clearly.[36] Within this crucial stage, key processes include preliminary assessment and adjustment of datasets, identification of necessary tasks, and evaluation of essential resources such as hardware and software. This phase focuses on defining the problem, determining both high-level tasks (HLT) and low-level tasks (LLT). Given the aim of predicting patient status (recovered or unrecovered), the HLT in this study revolves around building a predictive model. [37,38] Regarding the LLT, the selection of the following supervised machine learning algorithms was based on their respective strengths, as outlined by Sen P.C. *et al.*[38]: J48, Naïve Base, Random Tree, IBK, and REPTree Random Forest. Here is a brief discerption about each algorithm:

- **J48 (C4.5 Decision Trees):** J48 is an implementation of the C4.5 algorithm, a popular decision tree approach in machine learning.[39,40] It creates decision trees from a given dataset using information entropy. J48 can handle both categorical and numerical data, making it suitable for classification tasks.
- **Naïve Bayes:** Naïve Bayes is a straightforward probabilistic classifier that applies Bayes' theorem with strong independence assumptions between features.[41] It is easy to implement and performs well with large datasets.
- **Random Tree:** Random Tree is an ensemble learning technique that builds a group of decision trees during training and predicts the class based on the mode of the classes predicted by individual trees.[42] While similar to random forests, it constructs only one tree.
- **IBK (Instance-Based Learning with k-Nearest Neighbors):** IBK is a simple, instance-based learning algorithm that approximates functions locally and defers computation until evaluation.[43] It uses a distance metric to identify the k-nearest neighbors of a query instance and predicts based on the majority class among these neighbors.
- **REPTree (Regression Tree):** REPTree is a decision tree learner designed for regression tasks.[43] It constructs a decision tree where each leaf node represents a constant value, which is the average of the target attribute values of the instances reaching that node. REPTree can handle both numerical and categorical data.
- **Random Forest:** Random Forest is an ensemble learning method that builds a collection of decision trees during training and predicts the class based on the mode of the classes predicted by individual trees.[43] It improves performance over single decision trees by reducing overfitting and increasing accuracy, making it widely used for classification and regression tasks in machine learning.

### 2.2 Data Cleansing

This stage is dedicated to cleaning and refining the dataset by addressing outliers, missing values, and database balancing. The output of this stage is high-quality data. Real-world datasets often contain cases with missing values, noise, or inconsistencies due to various reasons such as incomplete information or data unavailability. The obtained dataset was collected by the Mexican government from multihospital database.[44] It consists of 5666020 anonymized samples and 21 parameters and most the parameters are of type numeric. Originally, the dataset has many missing values in most of the parameters. Newman D.A.[45] proposed three different mechanisms for missing the data: (1) missing completely at

© Engineered Science Publisher LLC 2024

random (MCAR), (2) missing at random (MAR), and (3) not missing at random (NMAR). Handling the issue of missing values is important to generates robust and reliable dataset. However, it is a challenge task as making the right decision require scientific justification. The most common methods which are used by the researchers to sort out these issues are:

- ignoring the data sample by removing it from the dataset
- ignoring the parameters that have many missing values
- replacing the missing value with either a global constant or the parameter means
- employing the data mining algorithm to estimate or predict the most probable value

Due to the fact that this is a medical data and it is not clear what was the reason behind missing a vast amount of data, in this study, all samples with missing values were removed apart from the values in the pregnancy feature at it represents the male patients. The missing values (97, 98, and 99) were replaced with NA. Other samples with missing values were removed. The value 99-99-9999 which represents the patients that have recovered were replaced with No. The irrelevant parameters such as Contact Other Covid, Entry Date, Date Dymptoms were removed as well. As the purpose of the study is to address the infected patients with the Covid19 pandemic to extract the hidden knowledge behind the patient recovery and unrecovered, all the uninfected patients were removed from the given dataset, resulting in 67300 patients. Among these patients, 20732 have diabetes, 2422 have COPD, 1534 have history of asthma, 23137 have hypertension disease, 2818 have cardiovascular,15865 have obesity disease and 3237 have history of renal chronic. The overall study population included 63% male and 37% female. Almost all ages were represented in our patients: 15.8% were less than 40, 18.9% were within the range of 40 to 50, 24.5% were within the range of 50 to 60, 21.6% were with the range of 60 to 70 and 19.2 were older than 70. The average age of all death patients is 61 years old and the average age of recovered patients is 52 years old. The number of patients who were admitted to ICU was 5667 (8%) with the average age of 55 years old, among whom 2896 (51%) had died with COVID-19.

### 2.3 Pre-processing

Generally, this stage is typically the most critical in any machine learning and data mining project. Here, the focus is on identifying and retaining the most relevant parameters while eliminating unproductive ones. Feature selection is an important step in the data preprocessing phase of machine learning. It involves detecting and selecting the most relevant and informative features (or parameters) from the given dataset. The primary objective of this process is to reduce the dimensionality of the dataset by eliminating irrelevant or redundant parameters; consequently, enhancing the performance and efficiency of the machine learning model. This will result in a more interpretable model, as it focuses on the most relevant features that contribute to the model's predictive power. By considering the most productive parameters, the develop model will have better ability to generalize to potential unseen data. In this study five feature selections algorithms were use in order to identify the most productive parameters from the given dataset: InfoGain and Gain Ration, Gini, $X^2$, ReliefF and FCBF. These algorithms are commonly used in literature by many researchers.[46,47] Here is a brief description about each algorithm:

- **InfoGain and Gain Ratio** are utilized in decision tree algorithms to select features. Information Gain gauges the decrease in entropy or uncertainty following a dataset split, while Gain Ratio normalizes this by the intrinsic information of the split.
- **Gini** is another metric in decision trees, measuring dataset impurity or disorder, with lower values indicating purer subsets.
- **$X^2$ (Chi-squared)** is a statistical test for determining variable independence. In feature selection, it assesses the relationship between a feature and the target variable.
- **ReliefF** is a feature selection method that evaluates feature relevance by considering instance weights. It samples instances iteratively, updating feature weights based on their ability to distinguish between similar instances.
- **Fast Correlation-Based Filter (FCBF)** is a feature selection algorithm that assesses feature merit based on class correlation and redundancy with other features. It aims to identify feature subsets highly correlated with the class and minimally redundant with each other.

The significant of the parameters are presented in Table 1. As shown that the algorithms highly ranked the first eight parameters. The only difference is that the $X^2$ algorithm prioritized the ICU parameter over diabetes. This study will be considered the first eight parameters: Age, Intubed, Pneumonia, hypertension (HBP), Diabetes, ICU, renal_chronic, and COPD. This selection of parameters is essential as they are deemed most significant in predicting certain outcomes or conditions in COVID-19 patients. For instance, age is a known risk factor, while parameters like pneumonia, hypertension, diabetes, and chronic renal and pulmonary diseases can complicate COVID-19 cases. Understanding the significance of these parameters can aid in better managing and treating patients, particularly in critical care settings like ICUs.

 © Engineered Science Publisher LLC 2024

**Table 1**. Significance of the parameters.

| | | # | Info. gain | Gain ratio | Gini | $\chi^2$ | ReliefF | FCBF |
|---|---|---|---|---|---|---|---|---|
| 1 | N age | | 0.051 | 0.026 | 0.032 | 3985.912 | 0.007 | 0.036 |
| 2 | C intubed | 2 | 0.032 | 0.071 | 0.021 | 2833.735 | -0.008 | 0.048 |
| 3 | C pneumonia | 2 | 0.027 | 0.030 | 0.017 | 814.324 | 0.002 | 0.000 |
| 4 | C hypertension (HBP) | 2 | 0.011 | 0.012 | 0.007 | 708.436 | 0.000 | 0.000 |
| 5 | C diabetes | 2 | 0.008 | 0.009 | 0.005 | 496.710 | 0.002 | 0.000 |
| 6 | C ICU | 2 | 0.007 | 0.016 | 0.005 | 607.391 | -0.002 | 0.000 |
| 7 | C renal_chronic | 2 | 0.003 | 0.012 | 0.002 | 312.006 | -0.004 | 0.006 |
| 8 | C copd | 2 | 0.002 | 0.008 | 0.001 | 177.434 | 0.000 | 0.000 |
| 9 | C Gender | 2 | 0.002 | 0.002 | 0.001 | 53.487 | 0.000 | 0.002 |
| 10 | C cardiovascular | 2 | 0.001 | 0.005 | 0.001 | 120.052 | 0.000 | 0.000 |
| 11 | C other_disease | 2 | 0.001 | 0.003 | 0.000 | 63.695 | 0.004 | 0.000 |
| 12 | C obesity | 2 | 0.000 | 0.000 | 0.000 | 21.290 | 0.004 | 0.000 |
| 13 | C tobacco | 2 | 0.000 | 0.000 | 0.000 | 17.636 | -0.004 | 0.000 |
| 14 | C inmsupr | 2 | 0.000 | 0.001 | 0.000 | 18.737 | 0.000 | 0.000 |
| 15 | C asthma | 2 | 0.000 | 0.001 | 0.000 | 14.644 | -0.008 | 0.000 |
| 16 | C patient_type | 1 | 0.000 | 0.000 | 0.000 | NA | 0.000 | 0.000 |

### 2.4 Validation Method

To validate the model, the 10-fold cross validation method was used. It is a method used to measure the performance of the selected algorithms. It is widely employed in the machine learning to run a comparison between the used algorithms to select the most predictive one.[48] Compared to the train/test method, 10-fold cross-validation is characterized as less biased/optimistic, producing dependable classification results, as highlighted by Brownlee.[48] The procedures of this model are summarized as follow:

Rearrange the given dataset randomly.

1. Split the dataset into a number of subsets k
2. For each generated subset:

a. Consider the first subset, k = 1, as the test dataset

b. Consider the remaining sample in the other subsets as a training dataset

c. Use the training set to train the model and then evaluate it on the test set

d. Maintain the obtained score of the estimation and reject the model

3. Present a summary of the overall score of the model based of the scored obtained above.

### 2.5 Model evaluation

One of the most important issues in predictive analysis is measure and evaluate the classification quality, usually in terms of accuracy. Several metrics are employed to assess the model's effectiveness, which encompass Accuracy, Recall, Precision, F-Measure, and ROC Area.

**a) Accuracy**

This metric is commonly applied in the field of classification. In practical terms, there are two commonly utilized measurements to evaluate classification accuracy or error. Accuracy, denoted as 'r' can be assessed through the following formula: $r = \frac{1}{n}\sum_{i=1}^{k} a_i$[49-52] where $a_i$ represents the samples with the majority label in class $i$, and $n$ signifies the total samples count in the dataset. Consequently, the classification error can be derived as $e = 1 - r$. The smaller value of $e$ indicates more favorable results. The same idea is implemented in Ref. [53] but it is presented in a different format:

$$E_c = \frac{\sum_{i=1}^{k}(s_i - M_i)}{\sum_{i=1}^{k} s_i} = \frac{\sum_{i=1}^{k}(s_i - M_i)}{n}$$

© Engineered Science Publisher LLC 2024

where $S_i$ represents the size of class $i$ and $M_i$ counts the majority samples with the matching label within class $i$.

**b) Recall**

Recall is an evaluation method employed to measure the algorithm's performance by quantifying the ratio of correctly identified true positives to the total actual positive instances. It can be mathematically computed as follows:

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

**c) Precision**

Precision determines the overall accurate classifications produced by the algorithm for relevant outcomes, including correct assessments of deceased and recovered cases. Its definition is as follows:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

**d) F-Measure**

It is the average of Precision and Recall. It is calculated as:

$$F - Measure = \frac{2 * Precision * Recall}{Precision + Recall}$$

**e) ROC Curve**

It represents Receiver Operating Characteristics, which is employed to visually depict how a classifier distinguishes between genuine and incorrect classes, with the aim of pinpointing the most effective threshold for their separation. This graph is formed by plotting the True Positive Rate against the False Positive Rate at various threshold levels.

- **Matthews Correlation Coefficient (MCC)**

MCC is a metric used to evaluate the quality of binary classifications. It considers true and false positives and negatives and provides a balanced measure even for imbalanced datasets. The MCC value ranges from -1 to +1, where +1 indicates a perfect prediction, 0 suggests no better than random prediction, and -1 indicates total disagreement between prediction and observation.

- **AUC-PR**

AUC-PR is a metric used to evaluate binary classification models, particularly in datasets with imbalanced class distributions. It measures the area under the precision-recall curve, which shows the trade-off between precision (positive predictive value) and recall (sensitivity) at various threshold settings. A higher AUC-PR indicates better performance, with a maximum value of 1 representing a perfect balance between precision and recall.

## 3. Results and discussion

Figure 3 presents a comparative analysis of the accuracy achieved by the selected algorithms. Based on the results, the J48 algorithm secured the top position with an accuracy of 69.6%. Following closely, REPTree attained the second-highest accuracy at 69.3%. The Radom Tree and IBK shared the third position with an accuracy of 68.6%. RandomForest exhibited the lowest accuracy among the algorithms, registering at 68.5%.

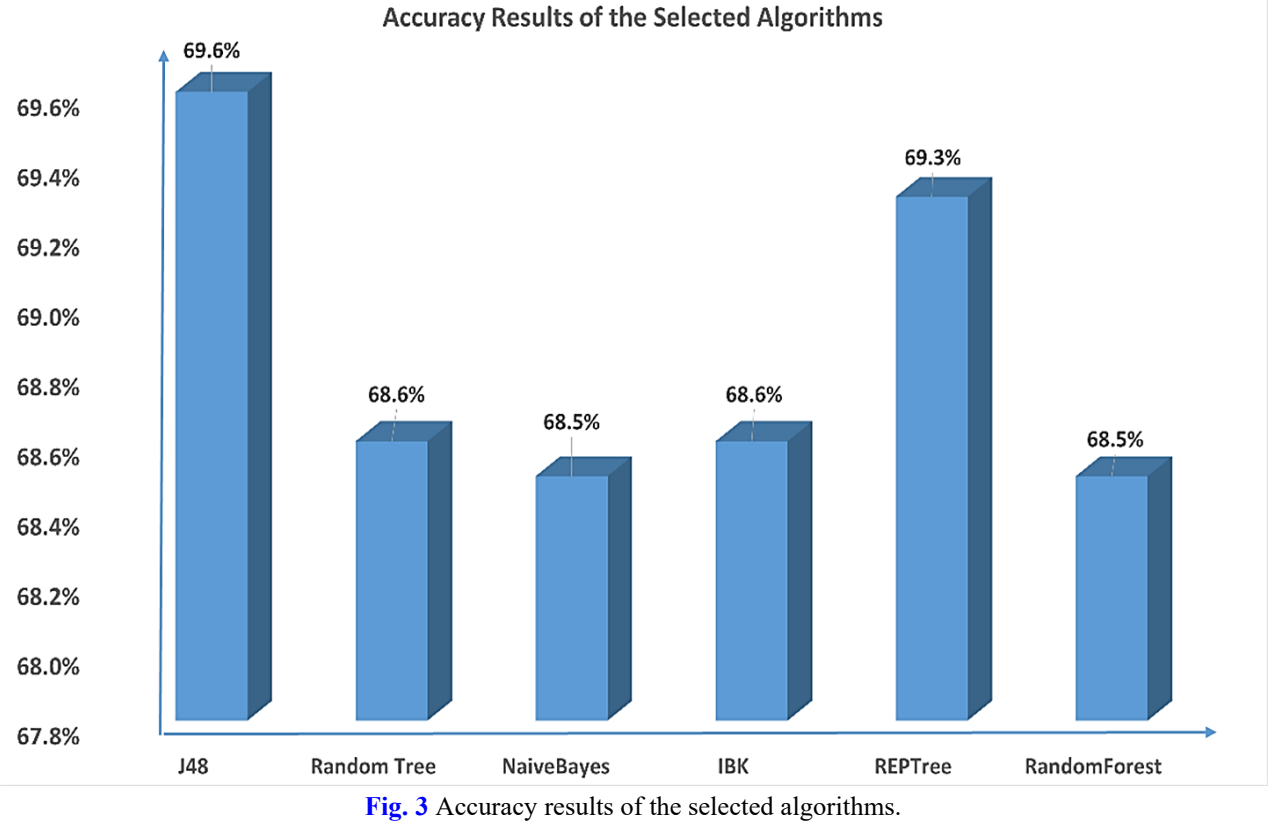


**Fig. 3** Accuracy results of the selected algorithms.


© Engineered Science Publisher LLC 2024

Figure 4 depicts a comparison among the chosen algorithms concerning precision and recall, highlighting J48 as the most optimal algorithm. This implies that when predicting death cases, the model accurately predicted these cases 68.3% of the time. Regarding recall, the model correctly identified 69.9% of all death and recovered cases. Given the trade-off between precision and recall, it's challenging to maximize both simultaneously since high precision results in low recall and vice versa. Therefore, it is advisable to consider the F-Measure score, where a higher value indicates a better-performing model. Once again, in the comparison of selected algorithms, the F-measure reaffirms that the J48 algorithm stands out as the most optimal, boasting a value of 67.2%.

The other effectiveness measurements are presented in Table 2. The ROC Curve, a crucial technique for assessing both model and algorithm effectiveness, indicates that J48 did not achieve the highest results in this particular measure. Nevertheless, its results are acceptable as it is approaching one. Based on its value, there is a 67.9% chance that the model is capable of differentiating between death and recovered

**Table 2.** Comprehensive precision of the algorithms.

| Algorithm | Class | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area |
|---|---|---|---|---|---|---|---|---|---|
| J48 | No | 0.879 | 0.637 | 0.716 | 0.879 | 0.789 | 0.287 | 0.671 | 0.758 |
| | Yes | 0.363 | 0.121 | 0.623 | 0.363 | 0.459 | 0.287 | 0.671 | 0.534 |
| | Avg. | 0.696 | 0.454 | 0.683 | 0.696 | 0.672 | 0.287 | 0.671 | 0.679 |
| Random Tree | No | 0.87 | 0.65 | 0.709 | 0.87 | 0.781 | 0.26 | 0.704 | 0.788 |
| | Yes | 0.35 | 0.13 | 0.597 | 0.35 | 0.442 | 0.26 | 0.704 | 0.541 |
| | Avg. | 0.686 | 0.465 | 0.669 | 0.686 | 0.661 | 0.26 | 0.704 | 0.701 |
| NaiveBayes | No | 0.822 | 0.579 | 0.721 | 0.822 | 0.768 | 0.263 | 0.714 | 0.816 |
| | Yes | 0.421 | 0.178 | 0.564 | 0.421 | 0.482 | 0.263 | 0.714 | 0.559 |
| | Avg. | 0.68 | 0.437 | 0.665 | 0.68 | 0.667 | 0.263 | 0.714 | 0.725 |
| IBK | No | 0.87 | 0.65 | 0.709 | 0.87 | 0.781 | 0.259 | 0.704 | 0.791 |
| | Yes | 0.35 | 0.13 | 0.597 | 0.35 | 0.441 | 0.259 | 0.704 | 0.542 |
| | Avg. | 0.686 | 0.466 | 0.669 | 0.686 | 0.661 | 0.259 | 0.704 | 0.703 |
| REPTree | No | 0.875 | 0.638 | 0.714 | 0.875 | 0.786 | 0.278 | 0.716 | 0.81 |
| | Yes | 0.362 | 0.125 | 0.613 | 0.362 | 0.455 | 0.278 | 0.716 | 0.565 |
| | Avg. | 0.693 | 0.456 | 0.678 | 0.693 | 0.669 | 0.278 | 0.716 | 0.723 |
| RandomForest | No | 0.861 | 0.634 | 0.712 | 0.861 | 0.78 | 0.262 | 0.708 | 0.802 |
| | Yes | 0.366 | 0.139 | 0.591 | 0.366 | 0.452 | 0.262 | 0.708 | 0.548 |
| | Avg. | 0.686 | 0.459 | 0.669 | 0.686 | 0.663 | 0.262 | 0.708 | 0.712 |

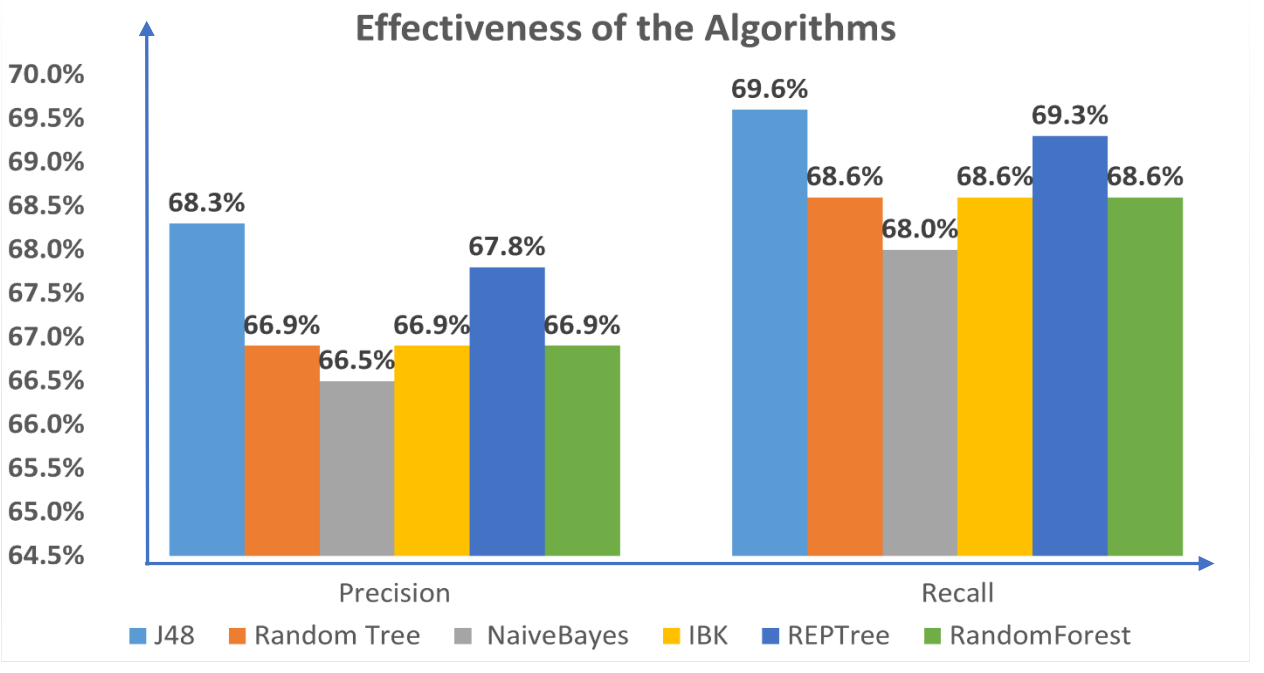


**Fig. 4** Precision and recall of the selected algorithms.

© Engineered Science Publisher LLC 2024

patients. Thus, based on the experimental results analysis, the J48 algorithm is the most optimal among the selected algorithms.

While models for diagnosing and prognosing COVID-19 have been developed, as mentioned earlier, there remains a deficiency in predicting and identifying the factors that lead to mortality.[54,55] This study addresses the limitations of prior researchers, identify the causes of death due to COVID-19 based on the obtained dataset and provide guidelines for future diagnosis.

Ahmed *et al*.[56] proposed a decision model that utilizes predictive analytics and machine learning to forecast the requirement for ICU care among COVID-19 patients. The model integrates feature selection, metaheuristic optimization, and machine learning techniques, utilizing both existing electronic health record (EHR) data and post-initial assessment information from emergency department (ED) patients. The research assesses three feature selection methods and employs optimized XGBoost to achieve the best-performing model, achieving an AUC of 89.2%. The experimental results suggest that this approach can enhance hospital patient flow efficiency by accurately predicting ICU care needs early on, facilitating improved resource planning and coordination across hospital units. This may lead to reductions in ED boarding delays and overcrowding, while also optimizing the utilization of regional healthcare resources.[56]

Sun L. *et al*.[22] conducted research in machine learning to early diagnosis Covid19 patients by predicting critical cases before symptoms appear. Their research utilized a dataset comprising 336 cases and 36 distinct parameters, including factors such as age, gender, hypertension, diabetes, and coronary disease. The target class in their study had two labels, with 3100 cases categorized as non-critical and 26 cases as critical. The researchers employed the Support Vector Machine (SVM) algorithm to predict the patients' status, and they assessed the accuracy of their classification using the Area Under the ROC Curve (AUC) measurement. According to the authors, their developed model demonstrated a high level of effectiveness, achieving an accuracy rate of 97.6%. However, it's worth noting that the dataset they used was relatively small, and they did not compare their results with those of other classification algorithms to determine the optimal classification accuracy. Additionally, the dataset suffered from an imbalance, as the non-critical cases dominated the overall classification accuracy. Therefore, it is crucial to consider other metrics such as Precision and Recall. de Holanda *et al*.[57] investigated the progression of COVID-19 in patients, focusing on hospitalization and mortality. Using 14 machine learning algorithms, predictive models were developed for both outcomes. Dataset was from patients with suspected COVID-19 was used to train the models. The study emphasized the influence of the number of vaccine doses on disease progression and underscored the importance of creating prediction models based solely on demographic data, vaccination history, symptoms, and comorbidities, without requiring clinical data. Based on the presented results, the Gradient Boosting (GB) algorithm outperformed others, achieving an 83% accuracy and an AUC of 0.89 for predicting mortality, and a 71% accuracy and an AUC of 0.75 for predicting hospitalization. The study revealed that age, shortness of breath, lack of vaccination, and the duration of symptoms are critical factors linked to disease progression.

In a separate study conducted by Wu *et al*.[23], a machine learning algorithm was deployed to expedite the diagnosis of COVID-19 patients. Their dataset included 253 cases and 49 parameters, encompassing variables like age, gender, pneumonia, tuberculosis, and lung cancer. They utilized the random forest algorithm to predict the target class and assessed the classification results using 10-fold cross-validation. Multiple metrics were used to evaluate classification accuracy, including AUC, Matthews correlation coefficient (MCC), and total accuracy (ACC). Their model proved to be highly effective, streamlining the process of laboratory blood tests and expediting treatment for infected patients. The primary limitation of this study was the reliance on a single algorithm, with no clear justification provided for not exploring other potentially effective algorithms.

Yang *et al*.[58] employed machine learning technique to conduct a comprehensive analysis to investigate how COVID-19 vaccinations affect carbon market pricing, using data on daily mass vaccinations and the EU carbon market. This study expands the understanding of how COVID-19 impacts the carbon market. The analysis suggests that COVID-19 vaccinations have a greater impact on predicting the volatility of the carbon market than its return. Specifically, it has been found that as the number of daily vaccinations increases, the market's volatility increases at a slower rate, while the return grows at a lower rate. This indicates that the predictive influence of COVID-19 vaccinations varies significantly between carbon market return and volatility. These findings contribute to the existing literature on the carbon market and provide insights into how vaccination programs influence market dynamics.

Yadav M. *et al*.[24] employed SVM algorithm in tracing the development of the Covid19 pandemic. Besides, Pearson's Correlation was used to identify the association between the Covid19 virus and weather conditions. The obtained results

 © Engineered Science Publisher LLC 2024

were compared with other well-known regression models including Polynomial Regression and Simple Linear Regression. The author claims that the obtained experimental results are promising in terms of efficiency and accuracy. Nevertheless, no details about the used data is provided. Besides, the SVM algorithm is not always the absolute winner in terms of the accuracy and when it scored the first position, it won with slightly better accuracy than the other compared algorithms.

A research was conducted for a purpose to find the effect of climate-related factors and population density on the spread of Covid19.[21] Virus optimization algorithm (VOA), adaptive network-based fuzzy inference system (ANFIS) and linear regression (LR) were tested in this research. The researchers tested the algorithms using the datasets from U.S. countries only. The result showed that the factors which are related to population density have more impact on the model which have proved the fact that social distancing reduce the spread of Covid19. With reference to the climates factors, this research proved that the increase in humidity leads to increase the infection rate while the increase in the maximum temperature leads to a reduction in the infection rate. In term of the accuracy, VOA recorded better accuracy than NFIS and LR. Also, VOA showed strong correlation between the variables. Like to above studies, this research has more or less the same limitations such as the size of the dataset is not mentioned and the dataset is limited to only one-month data.

In a study conducted by South Korean scientists,[59] researchers aimed to evaluate the predictive capabilities of machine learning regarding the mortality of COVID-19 patients using sociodemographic and specific medical data. Their dataset encompassed 10,237 COVID-19 patients, with 2.2% of them died, 75.9% experiencing a full recovery, and 21.9% remaining in isolation. The study employed five distinct machine learning algorithms: Linear SVM, K-nearest Neighbors, Random Forest (RF), Least Absolute Shrinkage and Selection operator (LASSO), and SVM with Radial Basis Function Kernel. The results of the experiments revealed that LASSO and Linear SVM achieved the highest sensitivities, measuring at 90.7% and 92%, respectively, along with the highest specificities, which were 91.4% and 91.8%, respectively. Additionally, Cox proportional hazards regression analysis was utilized in this research. The findings demonstrated a strong association between male patients and increased mortality risk, particularly among those aged above 70, those with moderate or severe disability, those displaying symptoms, those residing in nursing homes, and those with comorbidities like diabetes mellitus, chronic lung disease, or asthma.

This study focuses on a broad spectrum of parameters with the goal of identifying those that make significant contributions to the mortality of COVID-19 patients. These findings will help reduce the risk of death by enhancing diagnosis and expediting COVID-19 treatment protocols. To fulfill the objective, a dataset comprising 67,300 patients diagnosed with COVID-19 was utilized. Among them, 43,447 patients recovered from the disease, while 23,853 cases resulted in death. Evaluation of patients was conducted considering key parameters such as age, history of past illness, admission to the ICU, intubation and invasive ventilation, diabetes, cardiovascular conditions, obesity, chronic renal issues, and other clinical features. Statistically, among the given samples, 44796 patients with Covid19 pandemic have pneumonia disease, 426 pregnant women, 64378 have other historical diseases, and 5529 have the habit of smoking. Around 6335 patients have a drop in the oxygen; thus, there were intubed with the lungs ventilator pumps to smooth the oxygen in and Co2 out of the lungs. About 5442 patients had died due to the drop in drop in the oxygen in their blood cell. Although these patients were intubed, they could not have recovered as they have the infection of pneumonia.

As the J48 is the most optimal classification algorithm, Fig. 5 displays the decision tree model created using J48. In the beginning, the tree is at its root node and has the label "intubed" pointing hence it is beginning with the classification for whether or not a patient is intubated. Starting from that point, the tree has two main mono-components each of which is influenced by the patient's condition and attributes already. The left branch relates to patients who have and have not been intubated through close investigation of reasons that might cause pneumonia, and then an age stratification of higher risk patients and more conditions that are preexisting such as chronic renal issues, diabetes, and chronic obstructive pulmonary disease (COPD). Contrary to that, the left branch considers intubation and examines the same question as far as the status of pneumonia is concerned and proceeds to analysis by age and the same their health problems.

Conversely, as we move to the end of each branch, we get an increasing detail and use of more specific conditions and thresholds to narrow down the classification into more and more specific cases—which can also account for other factors that influence the classification—which is represented by the leaf nodes. These nodes, marked by numbers, mostly stand for the count of patients in each category as well as the degree model's prediction accuracy or divergence for these groups. This interactive leaves the path of the decision tree provides a good picture of the way the patient health factors and interventions are connected in medical settings leading to

© Engineered Science Publisher LLC 2024

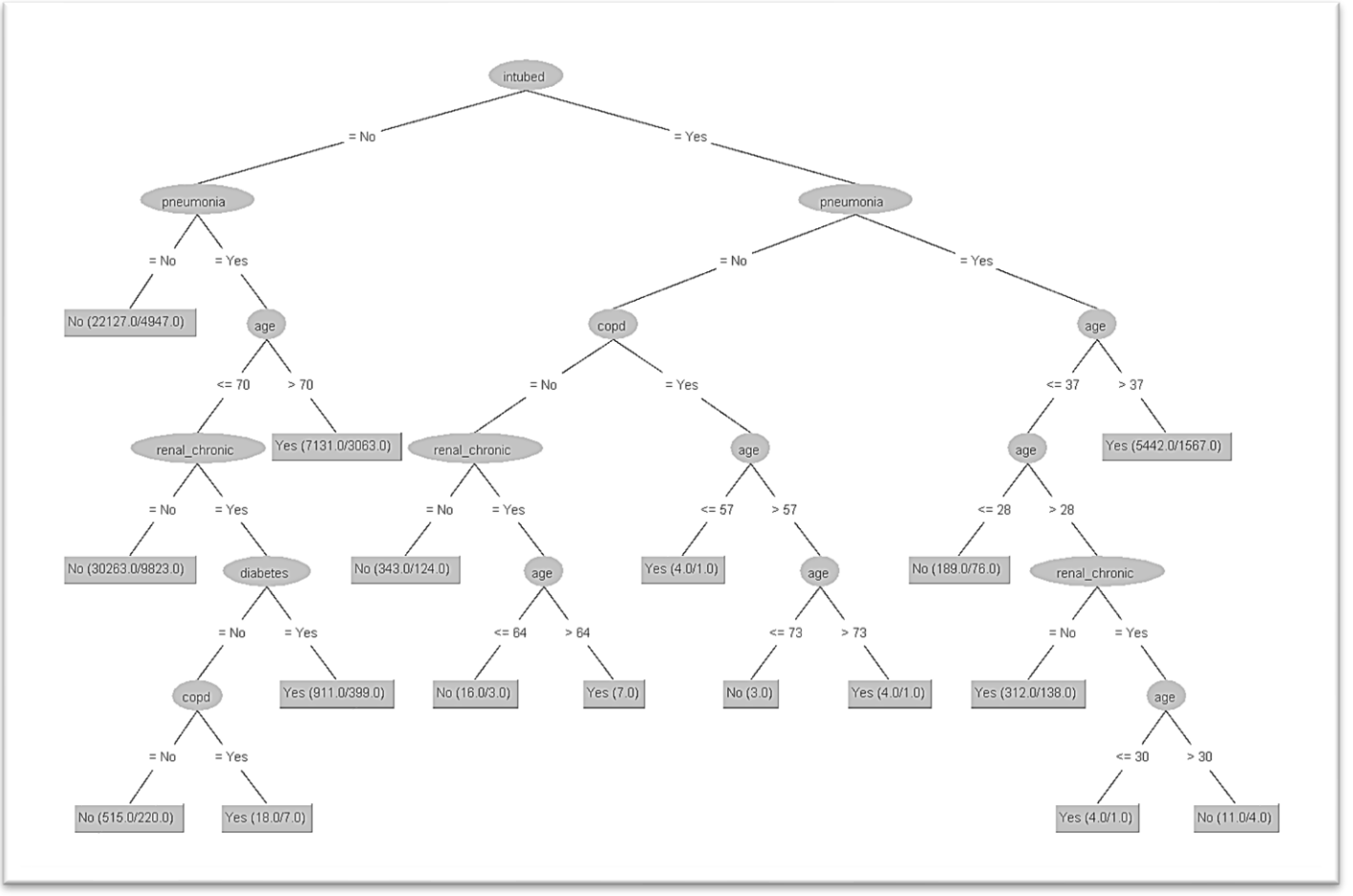


**Fig. 5** Decision tree of the J48 algorithm.

comprehensive understanding of patient management in the process.

Among 7009 patients who were more than 37 of age, had pneumonia and needed intubation, 78% of them had died. This finding confirmed the results in Refs. [60-62]. Pneumonia and invasive ventilation were also important factors for younger ages (under 37 years old), especially patients who were older than 28 of age, among whom 69% had died. These patients were not suffering from chronic renal disease or diabetes. The experimental result also shows that when patients did not have pneumonia but still required invasive ventilation, they were still at risk of the worse outcome. Among 517 patients who were pneumonia free but underwent intubation, 377 (71%) of them had died, although they were not suffering chronic diseases. Taken together, the analysis clearly suggests that invasive ventilation, pneumonia and age are among the top prognosis predictors.

When then analyzed another set of patients who did not required invasive ventilation. In this group, pneumonia was a good predictor for better outcome. Among 27074 patients who didn't require intubation and had no pneumonia, only 4947 (18%) had died. However, once again, suffering from pneumonia and age can lead to worse prognosis even when patients didn't require invasive ventilation. In this group, patient older than 70 years old were at higher risk of dying (70%). In contrast, younger age was associated with better outcome. Among 35207 patients who were less than 64 years old and had pneumonia, only 8277 (23.5%) had died. Our results support the fins of An C. *et al.*[20]; Biswas M. *et al.*[63]; Ho F.K. *et al.*[64]; Romero S. K. *et al.*[65] who found that older age is related directly with mortality risk.

Chronic diseases such as diabetes and hypertension can increase the risk for bad prognosis. Like the case in Refs. [66-74], it has been shown that such chronic diseases can affect COVID-19 disease outcome especially for older patients. Concerning the group of COVID-19 patients (7031) who were older than 64 of age, among them, 4058 patients were free of diabetes and hypertension, 1066 patients were suffering from diabetes and 1907 patients were suffering from both diabetes and hypertension. The analysis shows that not having both diabetes and hypertension increases the likelihood of survival in COVID-19 patients. Out of 4058 patients, only 1298 patients (32%) had died. By analyzing 1066 patients who suffered diabetes only, similar result was obtained. Among

 © Engineered Science Publisher LLC 2024

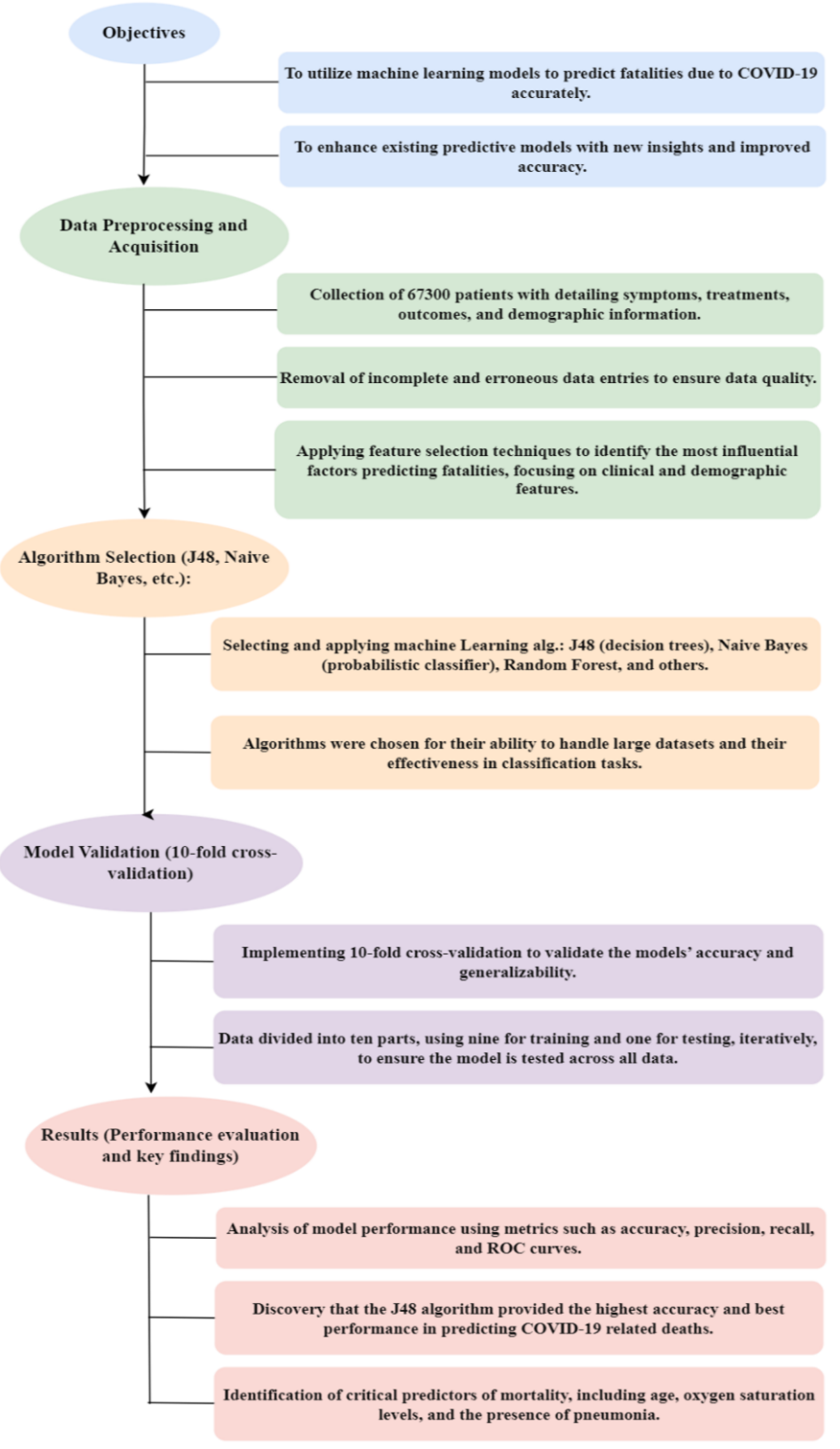


**Fig. 6** Depiction of the main steps of the machine learning-based approach for predicting COVID-19 fatalities.

them, only 31% had died. On the other hand, among 1907 patients who suffered from both diabetes and hypertension, 67% of them had died. These results clearly indicate that having more than chronic disease, significantly increase the risk of the bad prognosis

Generally, there is a positive relationship between the age of the patients and the death.[7,20,64,65] Thus, older people are at more risk than the young patients. Besides, there is a 10% risk of death for patients who have other diseases. This parameter is a bit vague as it is not clear what are these diseases. The hypertension is another parameter that contributes positively to the death by more than 10%. According to the study, 44% of the infected patients who have the history of high blood pressure (hypertension) had died. Whereas this percentage

© Engineered Science Publisher LLC 2024

dropped to 31% in case there was no history of such a disease. Patients who have the history of the COPD are at high risk of death as 48% of them had died. However, around 1250 patients who have such a disease have recovered from the Covid19 which give a positive sign of hope in the ability of the human body in tackling such a pandemic. Thus, as mentioned above, the light need be spotted in understanding the difference between these two classes by addressing other blood parameters to reveal the secret.

Figure 6 is the depiction of the structured research procedure which this study used to forecast the number of deaths from COVID-19 with machine leaning techniques.

The scheme flows with the objective of the study, which is to use machine learning in order to be able to make the right predictions regarding deaths cases of pandemic. For the study purposes the data was carefully collected from the Mexican Government's health database containing more than 67,300 patients. This data represents a variety of cases including diabetes, hypertension, Aids, asthma, less energy, and psychological stress. Therefore, subsequent steps of the preparation process were all about data preprocessing which was mainly data cleansing and feature selection, to give the information its integrity and relevance for the studies. Among various algorithms implemented, J48 and Naive Bayes ranked among the leaders, whereas Random Forest showed slight inferiority only in the case of death prediction. Models were validated through 10-fold cross-validation which allows ensuring predictive models to be sufficiently robust. The results part of this work analyses the performance of these algorithms, especially focusing on how the J48 algorithm was superior in its ability to predict fatalities more accurately using parameters such as age and oxygen saturation levels. Thus, the flowchart reiterates that findings would be implemented to solve the vulnerabilities of public health responses and to implement innovation in healthcare practice and highlights the key findings as well as the impact of the study. The study revealed that some parameters such as oxygen saturation, pneumonia, and age are the leading predictors for predicting mortality. This step-by-step regime not only made the study process run smoothly but also added precision to the research conclusions that predicted the COVID-19-related mortality.

## 4. Implications

This study offers a dual contribution, encompassing both theoretical and practical aspects. The theoretical contribution is particularly noteworthy, as it introduces several pivotal practical implications. The research findings have the potential to play a vital role in aiding decision-makers within healthcare institutions, providing them with the tools needed to comprehensively assess the risks, health-related benefits, and the cost-effectiveness associated with managing COVID-19. These insights empower decision-makers to take proactive measures to safeguard patients' lives, suggesting that treatment protocols have to be restructured to lessen the threat of mortality.

From a practical standpoint, the model developed in this study is dynamic and adaptable, offering the potential for use in diagnosing future diseases. This adaptability and versatility can significantly expedite the diagnostic process, ultimately reducing the incidence of fatal cases. The framework developed here extends its utility beyond COVID-19, showcasing its potential as a valuable resource in the ongoing battle against emerging diseases, thereby bolstering our capacity to respond swiftly and effectively to healthcare crises.

## 5. Limitations and future work

It is important to mention that some non-technical limitations played the role of the main obstacles to this research including but not limited to the availability of blood tests parameters that are related to the immune system approach. The study shows, a large number of the afflicted individuals had recovered from the epidemic without needing any additional care. Although admission to the ICU has an equal opportunity of death and recovery, it is highly recommended to consider other blood parameters to highlight the related antibodies. Consequently, future work should consider obtaining more details about the antibodies. This will support in conducting intensive experiments on machine learning algorithms to detect symptoms and immune responses in suspected cases across various age and blood groups. The comprehensive experiments will help identify and create self-therapies with the highest possible accuracy.

## 6. Conclusion

This study revealed that the pregnancy status, diabetes, hypertension, renal chronic and COPD are the main contributors to the death of the patients. Although admission to the ICU has an equal opportunity of death and recovery, it is highly recommended to consider these parameters to minimize the death by speeding up the initial medical checkup and then employing the appropriate medical intervention. It is found that the patients have better chance of recovery if they do not suffer from the hypertension. Pregnant patients require extra care and attention as the study shows that 93% of them have recovered from Covid19 whereas only 7% had died. In contrast, the number of death of non-pregnant is 37% which is 5 times higher than the pregnant patients. There might be a secret in the medicines which are taken by pregnant women. It has been highlighted that the gender, asthma, inmspur and

 © Engineered Science Publisher LLC 2024

obesity parameters do not have significant effect on identifying the reason behind the death due to COVID-19. The feature selection confirmed this finding as these parameters scored low values via both the Information Gain and the Correlation Algorithms. Concerning the patients who have history of smoking or have suffered from cardiovascular or have Renal Chronic, there is no clear evidence that these parameters are risky. The study found that there is somehow an equal percentage between the death and recovery due to these three parameters. More details about the antibodies need to be provided to distinguish between the two classes and extract the hidden knowledge. Although some of the findings in this study have been reported and become known to the medical staff, this study went beyond the limitation of the other studies by re-structuring the treatment protocol. Accordingly, cases are classified and prioritized systematically to minimize the death and ease the treatments rapidly.

## Conflict of Interest

There is no conflict of interest.

## Supporting Information

Not applicable.

## References

[1] WHO. Coronavirus disease (COVID-19) Pandemic, 2023.
[2] W. C.-. Dashboard, WHO Coronavirus (COVID-19) Dashboard Available, 2023.
[3] S. Lei, F. Jiang, W. Su, C. Chen, J. Chen, W. Mei, L.-Y. Zhan, Y. Jia, L. Zhang, D. Liu, Z.-Y. Xia, Z. Xia, Clinical characteristics and outcomes of patients undergoing surgeries during the incubation period of COVID-19 infection, *EClinicalMedicine*, 2020, **21**, 100331, doi: 10.1016/j.eclinm.2020.100331.
[4] I. Rahimi, F. Chen, A. H. Gandomi, A review on COVID-19 forecasting models, *Neural Computing and Applications*, 2023, **35**, 23671-23681, doi: 10.1007/s00521-020-05626-8.
[5] S. Ji, S. Xiao, H. Wang, H. Lei, Increasing contributions of airborne route in SARS-CoV-2 omicron variant transmission compared with the ancestral strain, *Building and Environment*, 2022, **221**, 109328, doi: 10.1016/j.buildenv.2022.109328.
[6] V. S. Salian, J. A. Wright, P. T. Vedell, S. Nair, C. Li, M. Kandimalla, X. Tang, E. M. Carmona Porquera, K. R. Kalari, K. K. Kandimalla, COVID-19 transmission, current treatment, and future therapeutic strategies, *Molecular Pharmaceutics*, 2021, **18**, 754-771, doi: 10.1021/acs.molpharmaceut.0c00608.
[7] WHO, Coronavirus disease (COVID-19), 2023.
[8] W. Dhouib, J. Maatoug, I. Ayouni, N. Zammit, R. Ghammem, S. Ben Fredj, H. Ghannem, The incubation period during the pandemic of COVID-19: a systematic review and meta-analysis, *Systematic Reviews*, 2021, **10**, 101, doi: 10.1186/s13643-021-01648-y.
[9] Daniel, T. W. Cenggoro, B. Pardamean, A systematic literature review of machine learning application in COVID-19 medical image classification, *Procedia Computer Science*, 2023, **216**, 749-756, doi: 10.1016/j.procs.2022.12.192.
[10] G. Badiola-Zabala, J. M. Lopez-Guede, J. Estevez, M. Graña, Machine learning first response to COVID-19: a systematic literature review of clinical decision assistance approaches during pandemic years from 2020 to 2022, *Electronics*, 2024, **13**, 1005, doi: 10.3390/electronics13061005.
[11] R. Prince, Z. Niu, Z. Y. Khan, M. Emmanuel, N. Patrick, COVID-19 detection from chest X-ray images using CLAHE-YCrCb, LBP, and machine learning algorithms, *BMC Bioinformatics*, 2024, **25**, 28, doi: 10.1186/s12859-023-05427-5.
[12] M. H. Ryalat, O. Dorgham, S. Tedmori, Z. Al-Rahamneh, N. Al-Najdawi, S. Mirjalili, Harris Hawks optimization for COVID-19 diagnosis based on multi-threshold image segmentation, *Neural Computing and Applications*, 2023, **35**, 6855-6873, doi: 10.1007/s00521-022-08078-4.
[13] NIH. (2021). How COVID-19 variants evade immune response. National Institues of Health Research Matters, 2023.
[14] W. Meter. (2023). Worldometer COVID-19 Data, 2023.
[15] O. Dorgham, M. Abu Naser, M. H. Ryalat, A. Hyari, N. Al-Najdawi, S. Mirjalili, U-NetCTS: U-Net deep neural network for fully automatic segmentation of 3D CT DICOM volume, *Smart Health*, 2022, **26**, 100304, doi: 10.1016/j.smhl.2022.100304.
[16] O. Dorgham, M. Fisher, S. Laycock, Accelerated generation of digitally reconstructed radiographs using parallel processing, *Proceedings Medical Image Understanding and Analysis*, 2009, 14-15.
[17] P. Chadha, G. N. Singh, Classification rules and genetic algorithms in data mining, *Global Journal of Computer Science and Technology Software and Data Engineering*, 2012, 12.
[18] D. Sindhu, S. Sindhu, Biological computers: their application in gene mining and protein engineering, *International Journal of Technical Research*, 2015, **4**, 15-21.
[19] S. Sindhu, D. Sindhu, Data Mining and Gene Expression Analysis in Bioinformatics, (IJCSMC) *International Journal of Computer Science and Mobile Computing*, 2017, **6**, 72-83.
[20] C. An, H. Lim, D.-W. Kim, J. H. Chang, Y. J. Choi, S. W. Kim, Machine learning prediction for mortality of patients diagnosed with COVID-19: a nationwide Korean cohort study, *Scientific Reports*, 2020, **10**, 18716, doi: 10.1038/s41598-020-75767-2.
[21] A. Behnood, E. Mohammadi Golafshani, S. M. Hosseini, Determinants of the infection rate of the COVID-19 in the U.S. using ANFIS and virus optimization algorithm (VOA), *Chaos, Solitons & Fractals*, 2020, **139**, 110051, doi: 10.1016/j.chaos.2020.110051.
[22] L. Sun, F. Song, N. Shi, F. Liu, S. Li, P. Li, W. Zhang, X. Jiang, Y. Zhang, L. Sun, X. Chen, Y. Shi, Combination of four clinical indicators predicts the severe/critical symptom of patients infected COVID-19, *Journal of Clinical Virology*, 2020, **128**, 104431, doi: 10.1016/j.jcv.2020.104431.
[23] J. Wu, P. Zhang, L. Zhang, W. Meng, J. Li, C. Tong, Y. Li, J. Cai, Z. Yang, J. Zhu, M. Zhao, H. Huang, X. Xie, S. Li, Rapid and accurate identification of COVID-19 infection through machine learning based on clinical available blood test results, 2020, 2020-2024, doi: 10.1101/2020.04.02.20051136.

© Engineered Science Publisher LLC 2024

[24] M. Yadav, M. Perumal, M. Srinivas, Analysis on novel coronavirus (COVID-19) using machine learning methods, *Chaos, Solitons & Fractals*, 2020, **139**, 110050, doi: 10.1016/j.chaos.2020.110050.
[25] J. Al Shaqsi, O. Drogham, S. Aburass, Advanced machine learning based exploration for predicting pandemic fatality: Oman dataset, *Informatics in Medicine Unlocked*, 2023, **43**, 101393, doi: 10.1016/j.imu.2023.101393.
[26] T. Chandrasekhar, K. Thangavel, E. Elayaraja, Effective clustering algorithms for gene expression data, 2012.
[27] M. Molla, M. Waddell, D. Page, J. Shavlik, Using machine learning to design and interpret gene-expression microarrays, *AI Magazine*, 2004, **25**, 23-44.
[28] H. Pirim, B. Ekşioğlu, A. D. Perkins, Ç. Yüceer, Clustering of high throughput gene expression data, *Computers & Operations Research*, 2012, **39**, 3046-3061, doi: 10.1016/j.cor.2012.03.008.
[29] M. Smolkin, D. Ghosh, Cluster stability scores for microarray data in cancer studies, *BMC Bioinformatics*, 2003, **4**, 36, doi: 10.1186/1471-2105-4-36.
[30] K. Y. Yeung, M. Medvedovic, R. E. Bumgarner, Clustering gene-expression data with repeated measurements, *Genome Biology*, 2003, **4**, R34, doi: 10.1186/gb-2003-4-5-r34.
[31] D. M. Mason, S. Friedensohn, C. R. Weber, C. Jordi, B. Wagner, S. M. Meng, P. Gainza, B. Correia, S. Reddy, Deep learning enables therapeutic antibody optimization in mammalian cells by deciphering high-dimensional protein sequence space, 2019.
[32] A. M. U. D. Khanday, S. T. Rabani, Q. R. Khan, N. Rouf, M. Mohi Ud Din, Machine learning based approaches for detecting COVID-19 using clinical text data, *International Journal of Information Technology*, 2020, **12**, 731-739, doi: 10.1007/s41870-020-00495-9.
[33] J. Al Shaqsi, M. Borghan, O. Drogham, S. Al Whahaibi, A machine learning approach to predict the parameters of COVID-19 severity to improve the diagnosis protocol in Oman, *SN Applied Sciences*, 2023, **5**, 273, doi: 10.1007/s42452-023-05495-5.
[34] A. Altan, S. Karasu, Recognition of COVID-19 disease from X-ray images by hybrid model consisting of 2D curvelet transform, chaotic salp swarm algorithm and deep learning technique, *Chaos, Solitons & Fractals*, 2020, **140**, 110071, doi: 10.1016/j.chaos.2020.110071.
[35] Khalifa N. E. M., Taha M., G. Manogaran, M. Loey, A deep learning model and machine learning methods for the classification of potential coronavirus treatments on a single human cell, 2020.
[36] J. C. W. Debuse, B. de la Iglesia, C. M. Howard, V. J. Rayward-Smith, Building the KDD roadmap. Roy R, Industrial Knowledge Management. London: Springer, 2001.
[37] B. Sekeroglu, R. Abiyev, A. Ilhan, M. Arslan, J. B. Idoko, Systematic literature review on machine learning and student performance prediction: critical gaps and possible remedies, *Applied Sciences*, 2021, **11**, 10907, doi: 10.3390/app112210907.
[38] P. C. Sen, M. Hajra, M. Ghosh, Supervised classification algorithms in machine learning: a survey and review. J. K. Mandal, D. Bhattacharya, eds. Advances in Intelligent Systems and Computing. Singapore: Springer Singapore, 2019.
[39] S. L. Salzberg. C4.5: Programs for Machine Learning by J. Ross Quinlan., 1993, **16**, 235–240, doi: 10.1007/BF00993309
[40] A. Dhakar, B. Singh, P. Gupta, Fault diagnosis of air compressor set-up using decision tree based J48 classification algorithm, *Journal of Engineering Research*, 2023, doi: 10.1016/j.jer.2023.09.028.
[41] A. Z. A. Magdacy Jerjes, A. Y. Dawod, M. F. Abdulqader, Detect malicious web pages using naive Bayesian algorithm to detect cyber threats, *Wireless Personal Communications*, 2023, doi: 10.1007/s11277-023-10713-9.
[42] R. R. Chowdhury, A. C. Idris, P. E. Abas, Identifying SH-IoT devices from network traffic characteristics using random forest classifier, *Wireless Networks*, 2024, **30**, 405-419, doi: 10.1007/s11276-023-03478-3.
[43] S. K. David, M. Rafiullah, K. Siddiqui, Comparison of different machine learning techniques to predict diabetic kidney disease, *Journal of Healthcare Engineering*, 2022, **2022**, 7378307, doi: 10.1155/2022/7378307.
[44] T. Mukherjee, COVID-19 patient pre-condition dataset, M. government, Ed., 2 ed. kaggle data repository 2020.
[45] D. A. Newman, Missing data: five practical guidelines, Sage, 2014.
[46] S. K. Arjaria, A. S. Rathore, J. S. Cherian, Kidney disease prediction using a machine learning approach: a comparative and comprehensive analysis. Demystifying Big Data, Machine Learning, and Deep Learning for Healthcare Analytics. Amsterdam: Elsevier, 2021.
[47] H. Hamla, K. Ghanem, A comparative study of filter feature selection methods on microarray data. Laouar MR, Balas VE, Lejdel B, Eom S, Boudia MA, International Conference on Computing and Information Technology. Cham: Springer, 2023.
[48] J. Brownlee, A Gentle Introduction to k-fold Cross-Validation, in tatistics, J. Brownlee, Ed., ed: Machinr Lesrning Mastery, 2018.
[49] S. Aranganayagi, K. Thangavel, Improved K-modes for categorical clustering using weighted dissimilarity measure, *International Journal of Computational Intelligence*, 2009, **5**, 729-735, doi: 10.5281/zenodo.1070405.
[50] Z. He, S. Deng, X. Xu, Approximation algorithms for K-modes clustering, Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006.
[51] Z. Huang, Extensions to the k-means algorithm for clustering large data sets with categorical values, *Data Mining and Knowledge Discovery*, 1998, **2**, 283-304, doi: 10.1023/A:1009769707641.
[52] Khan S. S., S. Kant, Computation of initial modes for k-modes clustering algorithm using evidence accumulation, in International Joint Conference on Artificial Intelligence Hyderabad, India, 2007.
[53] A. Gionis, H. Mannila, P. Tsaparas, Clustering aggregation, *ACM Transactions on Knowledge Discovery from Data*, 2007, **1**, 4, doi: 10.1145/1217299.1217303.

 © Engineered Science Publisher LLC 2024

[54] N. Tzenios, M. Chahine, M. Tazanios, Better strategies for coronavirus (COVID-19) vaccination, *Special Journal of the Medical Academy and Other Life Sciences*, 2023, **1**, doi: 10.58676/sjmas.v1i2.11.
[55] A. L. dos Santos, C. Pinhati, J. Perdigão, S. Galante, L. Silva, I. Veloso, A. C. Simões e Silva, E. A. Oliveira, Machine learning algorithms to predict outcomes in children and adolescents with COVID-19: a systematic review, *Artificial Intelligence in Medicine*, 2024, **150**, 102824, doi: 10.1016/j.artmed.2024.102824.
[56] A. Ahmed, F. D. Zengul, S. Khan, K. R. Hearld, S. S. Feldman, A. G. Hall, G. N. Orewa, J. Willig, K. Kennedy, Developing a decision model to early predict ICU admission for COVID-19 patients: a machine learning approach, *Intelligence-Based Medicine*, 2024, **9**, 100136, doi: 10.1016/j.ibmed.2024.100136.
[57] W. D. de Holanda, L. C. e Silva, Á. A. de Carvalho César Sobrinho, Machine learning models for predicting hospitalization and mortality risks of COVID-19 patients, *Expert Systems with Applications*, 2024, **240**, 122670, doi: 10.1016/j.eswa.2023.122670.
[58] C. Yang, H. Zhang, F. Weng, Effects of COVID-19 vaccination programs on EU carbon price forecasts: evidence from explainable machine learning, *International Review of Financial Analysis*, 2024, **91**, 102953, doi: 10.1016/j.irfa.2023.102953.
[59] N. Chen, M. Zhou, X. Dong, J. Qu, F. Gong, Y. Han, Y. Qiu, J. Wang, Y. Liu, Y. Wei, J.-A. Xia, T. Yu, X. Zhang, L. Zhang, Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study, *The Lancet*, 2020, **395**, 507-513, doi: 10.1016/s0140-6736(20)30211-7.
[60] M. E. H. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M. A. Kadir, Z. Bin Mahbub, K. R. Islam, M. S. Khan, A. Iqbal, N. Al Emadi, M. Bin Ibne Reaz, M. T. Islam, Can AI help in screening viral and COVID-19 pneumonia? *IEEE Access*, 2020, **8**, 132665-132676, doi: 10.1109/access.2020.3010287.
[61] L. Gattinoni, D. Chiumello, P. Caironi, M. Busana, F. Romitti, L. Brazzi, L. Camporota, COVID-19 pneumonia: different respiratory treatments for different phenotypes? *Intensive Care Medicine*, 2020, **46**, 1099-1102, doi: 10.1007/s00134-020-06033-2.
[62] E. Y. P. Lee, M.-Y. Ng, P.-L. Khong, COVID-19 pneumonia: what has CT taught us? *The Lancet Infectious Diseases*, 2020, **20**, 384-385, doi: 10.1016/s1473-3099(20)30134-1.
[63] M. Biswas, S. Rahaman, T. K. Biswas, Z. Haque, B. Ibrahim, Association of sex, age, and comorbidities with mortality in COVID-19 patients: a systematic review and meta-analysis, *Intervirology*, 2021, **64**, 36-47, doi: 10.1159/000512592.
[64] F. K. Ho, F. Petermann-Rocha, S. R. Gray, B. D. Jani, S. V. Katikireddi, C. L. Niedzwiedz, H. Foster, C. E. Hastie, D. F. MacKay, J. M. R. Gill, C. O'Donnell, P. Welsh, F. Mair, N. Sattar, C. A. Celis-Morales, J. P. Pell, Is older age associated with COVID-19 mortality in the absence of other risk factors? General population cohort study of 470, 034 participants, *PLoS One*, 2020, **15**, e0241824, doi: 10.1371/journal.pone.0241824.
[65] K. Romero Starke, G. Petereit-Haack, M. Schubert, D. Kämpf, A. Schliebner, J. Hegewald, A. Seidler, The age-related risk of severe outcomes due to COVID-19 infection: a rapid review, meta-analysis, and meta-regression, *International Journal of Environmental Research and Public Health*, 2020, **17**, 5974, doi: 10.3390/ijerph17165974.
[66] Z. T. Bloomgarden, Diabetes and COVID-19, *Journal of Diabetes*, 2020, **12**, 347-348, doi: 10.1111/1753-0407.13027.
[67] L. Fang, G. Karakiulakis, M. Roth, Are patients with hypertension and diabetes mellitus at increased risk for COVID-19 infection? *The Lancet Respiratory Medicine*, 2020, **8**, e21, doi: 10.1016/s2213-2600(20)30116-8.
[68] G. Lippi, J. Wong, B. M. Henry, Hypertension and its severity or mortality in Coronavirus Disease 2019 (COVID-19): a pooled analysis, *Polish Archives of Internal Medicine*, 2020, 130, 304-309, doi: 10.20452/pamw.15272.
[69] H. Liu, S. Chen, M. Liu, H. Nie, H. Lu, Comorbid chronic diseases are strongly correlated with disease severity among COVID-19 patients: a systematic review and meta-analysis, *Aging and Disease*, 2020, **11**, 668, doi: 10.14336/ad.2020.0502.
[70] R. Muniyappa, S. Gubbi, COVID-19 pandemic, coronaviruses, and diabetes mellitus, *American Journal of Physiology-Endocrinology and Metabolism*, 2020, **318**, E736-E741, doi: 10.1152/ajpendo.00124.2020.
[71] S. Peric, T. M. Stulnig, Diabetes and COVID-19, *Wiener Klinische Wochenschrift*, 2020, **132**, 356-361, doi: 10.1007/s00508-020-01672-3.
[72] F. Rubino, S. A. Amiel, P. Zimmet, G. Alberti, S. Bornstein, R. H. Eckel, G. Mingrone, B. Boehm, M. E. Cooper, Z. Chai, S. Del Prato, L. Ji, D. Hopkins, W. H. Herman, K. Khunti, J.-C. Mbanya, E. Renard, *New-onset diabetes in covid-19, New England Journal of Medicine*, 2020, **383**, 789-790, doi: 10.1056/nejmc2018688.
[73] J. A. Shaqsi, O. Drogham, S. Aburass, Advanced machine learning based exploration for predicting pandemic fatality: Oman dataset, *Informatics in Medicine Unlocked*, 2023, **43**, 101393, doi:10.1016/j.imu.2023.101393.
[74] E. Selvin, S. P. Juraschek, Diabetes epidemiology in the COVID-19 pandemic, *Diabetes Care*, 2020, **43**, 1690-1694, doi: 10.2337/dc20-1295.

**Publisher's Note**: Engineered Science Publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.