The Scientific Journal of Cihan University – Sulaimaniya **PP: 61-75**
Volume (7), Issue (1), June 2023
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

Predicting Mortality COVID-19 Model and Characterize the High Risk of Dying Based on Machine Learning Models

Manal Ali Atiyah[1], Chra Ali Kamal[2]

[1,2]Database Department, Computer Scince Instuite, Sulaimani polytechnic university, Sulaimani, Iraq

Email: manal.ali@spu.edu.iq[1], chra.kamal@spu.edu.iq[2]

Abstract:

COVID-19 is a respiratory illness that the SARS-CoV-2 virus has caused. It has been declared a pandemic by the (WHO). *The surprise increase in infections and the high mortality rates have strained the public healthcare system. Thus, it is fatal to recognize the most remarkable factors for mortality prediction to improve* patient treatment strategy. We developed a predictive model using machine learning and AI to identify the health risks and predict the mortality rate of patients with COVID-19. In this study, 4.711 patients with confirmed SARS-CoV-2 infection were included the goal of this study is to develop an AI model that can help medical facilities identify which patients need immediate attention and which ones should be sent to the hospital first. It can also help them, triage patients, during times of overcrowding. various machine learning algorithms used to analyze the mortality rate of patients with COVID-19. These included Random Forests, K-Nearest Neighbors, Neural Networks, Support Vector Machines, and Decision Trees. We applied the feature selection method to identify to improve the machine learning algorithm achievement. the results showed that the Random Forests and Neural Networks approximated the best predictive model with a balanced accuracy of 85%. The model with the best performance supplies a tool to decrease mortality. We split the dataset of COVID-19 patients to analyze our model's accuracy Finally We used a confusion matrix to analyze the model's accuracy and other evaluation metrics.

Key words: COVID-19, healthcare, mortality, machine learning, pandemic, predictive model.

الملخص:

كوفيد-١٩ هو مرض تنفسي سببه فيروس سارس كوف-٢. تم إعلانه جائحة من قبل (منظمة الصحة العالمية). أدت الزيادة المفاجئة في الإصابات ومعدلات الوفيات المرتفعة إلى إجهاد نظام الرعاية الصحية العامة. وبالتالي ، من المصيري التعرف على العوامل الأكثر بروزًا للتنبؤ بالوفيات لتحسين استراتيجية علاج المريض. لقد طورنا نموذجًا تنبئيًا باستخدام التعلم الآلي(ML) والذكاء الاصطناعي (AI) لتحديد المخاطر الصحية والتنبؤ بمعدل وفيات المرضى المصابين بـ كوفيد-١٩. في هذه الدراسة ، تم تضمين٤٧١١ مريضًا مصابًا بعدوى سارس كوف-٢ المؤكدة ، والهدف من هذه الدراسة هو تطوير نموذج ذكاء اصطناعي يمكن أن يساعد المرافق الطبية في تحديد المرضى الذين يحتاجون إلى عناية فورية وأيهم يجب إرساله إلى المستشفى أولاً. يمكن أن يساعدهم أيضًا في فرز المرضى خلال أوقات الازدحام. تم إستخدام مختلف خوارزميات التعلم الآلي لتحليل معدل وفيات المرضى المصابين بـ كوفيد-١٩. وشملت هذه الغابات العشوائية ، والجيران K-الأقرب ، والشبكات العصبية ، وآلات ناقلات الدعم ، وأشجار القرار. طبقنا طريقة اختيار الميزة لتحديد تحسين إنجاز خوارزمية التعلم الآلي. أظهرت النتائج أن الغابات العشوائية والشبكات العصبية تقارب أفضل نموذج تنبؤي بدقة متوازنة تبلغ ٨٥٪. يوفر النموذج ذو الأداء الأفضل أداة لتقليل معدل الوفيات. قمنا بتقسيم مجموعة البيانات الخاصة بمرضى كوفيد-١٩ لتحليل دقة نموذجنا وأخيرًا استخدمنا مصفوفة الارتباك لتحليل دقة النموذج ومقاييس التقييم الأخرى.

DOI: http://dx.doi.org/10.25098/7.1.4

الكلمات المفتاحية: كوفيد-١٩ ، الصحة البدنية، نسبة الموت، التعلم الالي، الجائحة، النموذج التوقيعي.

پوخته:

کۆڤید-١٩ نەخۆشییەکی کۆئەندامی هەناسەیە کە بەهۆی ڤایرۆسی سارس-کۆڤ-٢ دروست دەبێت وە لە لایەن ڕێکخراوی تەندروستی جیهانییەوە بە پەتا ناسێنراوە. زیادبوونی لەناکاوی تووشبوون و بەرزبوونەوەی ڕێژەی مردن، فشاری بۆ سیستەمی چاودێری تەندروستی گشتیی دروست کردووە بۆیە زۆر گرنگە زانینی دیارترین هۆکارەکان بۆ پێشبینیکردنی مردن لەپێناوباشترکردنی ستراتیژی چارەسەرکردنی توشبووەکان. ئێمە مۆدێلێکی پێشبینیکراومان پەرەپێدا بە بەکارهێنانی فێربوونی ئامێر(ML) و زیرەکی دەستکرد (AI) بۆ دەستنیشانکردنی مەترسییە تەندروستییەکان و پێشبینیکردنی ڕێژەی مردنی نەخۆشانی تووشبوو بە کۆڤید-19. لەم توێژینەوەیەدا، ٤٧١١ نەخۆش کە تووشبوونیان بە نەخۆشی سارس-کۆڤ-٢ پشتڕاستکراوەتەوە، بەشداربوون وە ئامانج لەم توێژینەوەیە بریتییە لە پەرەپێدانی مۆدێلێکی زیرەکی دەستکرد کە بتوانێت یارمەتی دامەزراوە پزیشکییەکان بدات بڕیاربدەن کام نەخۆش پێویستی بە چاودێرییەکی خێرا هەیە و کامیان پێویستە سەرەتا بنێردرێتە نەخۆشخانە. هەروەها دەتوانێت یارمەتیان بدات لە کاتی سەرقاڵیدا نەخۆشەکان جیابکەنەوە. جۆرەها ئەلگۆریتمی فێربوونی ئامێری بەکارهێنراون بۆ شیکردنەوەی ڕێژەی مردنی ئەو نەخۆشانەی کە تووشی کۆڤید-١٩ بوون. لەوانە دارستانە هەڕەمەکییەکان، دراوسێکانی نزیکترین K، تۆڕە دەمارییەکان، ئامێری ڤێکتەری پشتگیری و دارەکانی بڕیاردان. ئێمە شێوازی هەڵبژاردنی تایبەتمەندیمان بەکارهێنا بۆ دیاریکردن و باشتربوونی دەستکەوتی ئەلگۆریتمەکانی فێربوونی ئامێر. ئەنجامەکان دەریدەخەن کە دارستانە هەڕەمەکییەکان و تۆڕە دەمارییەکان باشترین مۆدێلی پێشبینیکردنن بە وردبینییەکی هاوسەنگ بەڕێژەی ٨٥٪. باشترین مۆدێلی ڕاهێنراومان وەکو ئامرازێک بەکارهێنا بۆ کەمکردنەوەی مردن وە کۆمەڵەی داتاکانی نەخۆشانی کۆڤید-١٩مان دابەش کرد بۆ شیکردنەوەی وردی مۆدێلەکەمان و لە کۆتاییدا ماتریکسی سەرلێشێواویمان بەکارهێنا بۆ شیکردنەوەی وردی مۆدێلەکە و پێوەرەکانی تری هەڵسەنگاندن.

کلیلە وشە: کۆڤید-١٩ ، چاودێری تەندروستی، ڕێژەی مردن، فێربوونی ئامێر، پەتا، مۆدیلی پێشبینی.

1. INTRODUCTION

COVID-19 is the last coronavirus family that can be prevalent from animals to humans. In December 2019, COVID-19 was introduced in China. As of September 2022, around 620 million cases of COVID-19 have been reported in various countries the mortality rate was 1% of the total closed cases while the Serious or Critical cases recorded 39,930 with a rate of 0.3 of the total active cases [1]. the announcement of COVID-19 as a pandemic resulted from increasing fear of the terrible expansion and riskiness of the virus. It is recognized via its presence as a public health anxiety that has spread over the world [2]. Countries around the world have been affected a lot by this virus, which has resulted in some precautionary measures Such as general closures, curfews, and travel restrictions. Although the signs of the disease are often mild, some injuries were dangerous and sometimes fatal [3]. However, it’s assumed that the real infected numbers were much more in order to the lack of tasting and the likely reduction of notification. Furthermore, currently, there aren’t settled prediction biomarkers to exactly predict which patients are at pressing risk of death or require immediate go-up care, making resource distribution hard [4]. Due to the late diagnosis and the unknown nature of the virus, many infected patients did not receive the proper treatment. To improve the quality of care for these patients, researchers have been developing new methods to screen them for various clinical features [5].

Machine learning (ML) is one of the most paramount offshoots of artificial intelligence (AI) used to produce systems that are able to canon examples and recover with no need to be explicitly programmed [6].ML is modifying healthcare by giving away human and people's health through

DOI: http://dx.doi.org/10.25098/7.1.4

various computational benefits. It gives a share in observing diseased patients, disease type analysis, determining and making prescriptions of a medicine, supplying patient-healthcare, reducing clinical mistakes, predicting outcomes curative decision making, disclosing sepsis, and high-risk emergencies in sick people [7]. Many research papers have been published in covid -19 by using ML. Models of AI, big data platforms, and nature-inspired algorithms can be used for issue detection and approach tracing of COVID-19 [8]. This paper aims to develop and compare some of the different machine learning algorithms to predict mortality in COVID-19 patients. The system offered to preprocess algorithm for data to choose new features, eject vain and excessive data elements flowed by selecting the most paramount features. After that predictive model has been developed to classify the data and predict the risk of mortality and the probability of survival. The model will be rendered as an early notification system to timely characterize at-risk patients. The remnant of this paper is illustrated as follows. Section 2 provides related research Section 3 demonstrates different methods and models will demonstrate with explaining each method by giving information about each model, data preprocessing, and feature Section 4 provides the results with several processes and measurements. Section 5 covers the discussion and conclusion.

2. RELATED WORK

[9] used a set of machine learning models to analyze the data set generated by the hospital in Wuhan, China. The data included 375 blood samples taken from patients admitted to the hospital. The authors present a case study that shows how Machine Learning models can improve the accuracy of hospital records by identifying patients who survived and died while they were in the hospital. A study found [10] analyzed the early mortality risk prediction of patients with SARS-CoV-2 using laboratory markers and demographic information. They used a combination of these tools to develop models that were able to predict the disease's severity and mortality. Five baseline models were then used to analyze the 183 individuals from Wuhan's Tongji Hospital. The researchers used Logistic Regression as their final model due to its simplicity. The Area under the ROC Curve (AUC) of their external dataset was 88%.

The sensitivity and specificity of their model were respectively 83% and 79.4%. In Pakistan, a study conducted [11] analyzed the performance of different regression models in simulating the deaths of COVID-19 patients. They found that the most accurate model was the quadratic one, which showed the best performance. They then concluded that the mortality rate would decrease by October. According to [12] the Support Vector Machines perform better than linear regression and multi-layer perceptron when it comes to predicting the COVID-19 trend in Germany, the USA, and the globe.[13] conducted a study on 3,524 COVID-19 patients, the researchers used five machine learning models to predict their mortality. The models included Logistic Regression, a Support Vector Machine, a Random Forest, gradient boosting, and a K Nearest Neighbor.

They proposed that the model could be used as an open-source decision support tool.[14] present a study The goal of this study was to develop an algorithm that could predict the mortality of infected individuals based on their clinical features. The researchers used 97 clinical features, including demographic variables, blood measurements, and anthropometric measures. with an accuracy of 81.4%) and a severity (of 72.3%). [15] The authors of this study looked into the impact of different nutrients on the number of COVID-19 deaths in different countries. They then calculated the K-means

DOI: http://dx.doi.org/10.25098/7.1.4

by country based on the distribution of energy, protein, and fat in 23 different types of food. They found that high-fat consumption was associated with higher death rates.[16] analyze the various types of lockdowns in three countries and predict the number of cases and deaths caused by COVID-19. Through the use of machine learning models and time series, such as the SVM, random forests, decision trees (DTs), K-Nearest Neighbors, polynomial regression, and Holt winter the researchers were able to identify the most likely cases and deaths. [17] The authors used various machine learning algorithms to analyze the data collected from COVID-19 patients. They were able to identify the most alarming features and symptoms of the condition. [18] contrast the machine learning algorithms executed for predicting the Intensive Care Unit (ICU) admission and mortality for COVID-19 patients. They present internally validated models by using patients presenting to the Emergency Department (ED) they conclude that ensemble-based models proceed better than the other model types at predicting mortality for COVID-19. CRP, LDH, and O2 saturation were recorded as important for ICU admission models while eGFR <60 ml/min/1.73m2, and neutrophil and lymphocyte percentages represent the most important variables for predicting mortality.

3. METHODS AND MATERIALS

3.1.source of data

the data used in this paper from a healthcare surveillance software package (Clinical Looking Glass [CLG]; Streamline Health, Atlanta, Georgia) and a check of the primary medical records, freely on Kaggle [19]. the dataset includes information on demographics, comorbidities, admission laboratory rate, admission medications, admission supplemental oxygen demand, discharge, and mortality. The original dataset includes 85 data elements from each patient, some redundant and useless data have been removed such as Age, Black, White and Asian. There isn't a missing value in this data, The default categorical features in the dataset were coded as 1 for positive and 0 for negative. The dataset should be balanced so it would be more accurate. Figure (1) shows the case diagram for the proposed study.

DOI: http://dx.doi.org/10.25098/7.1.4

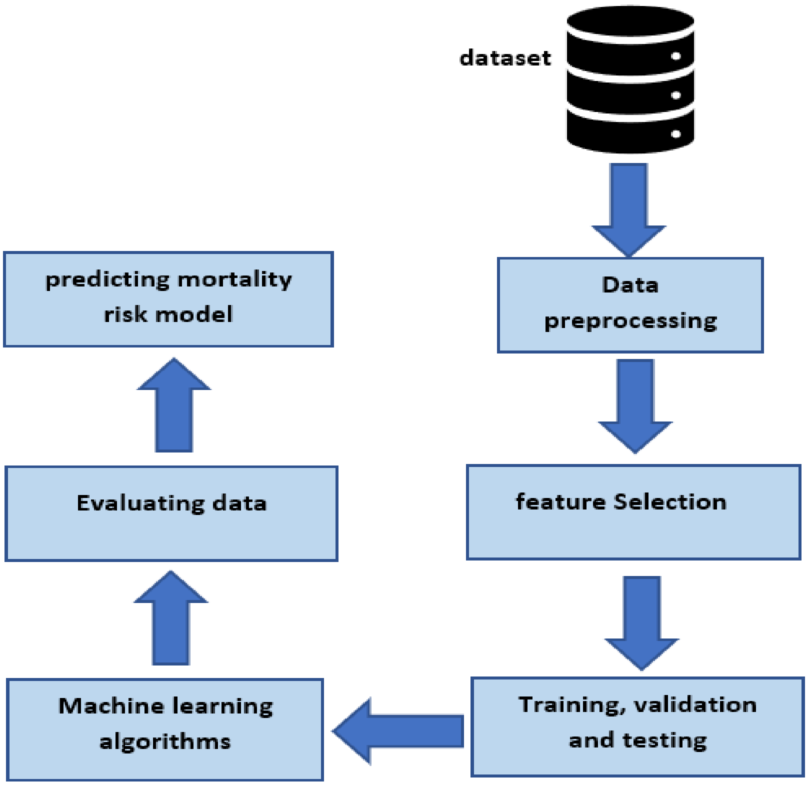


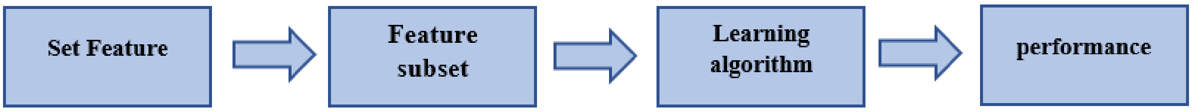
Figure 1: Framework of the proposed study.

3.2.Dataset Pre-processing

3.2.1 Feature selection

Feature selection is a fatal step in machine learning algorithms and data mining. It can help improve the efficiency of the algorithms and reduce the time it takes to process the data, it minimizes the processing time and accuracy of the denomination, there are three main classes of feature selection which it is Filtering, Wrapper, and embedded methods, The most efficient method for evaluating feature subsets is the Wrapper method, which trains a model using learning algorithms. However, it is also the most demanding, as it requires a lot of computational resources. Another approach is the Filter method, which takes a measure to assign a score to a set of features. This group is generally less expensive than the other family but does not require model training. The embedded methods, which train models using learning algorithms, perform feature selection during the model construction phase. Compared to the filter methods, the embedded methods are faster. They are also more precise and have better computational difficulty [20]. The development of intelligent analysis tools requires the use of feature selection methods. These methods can be used to enhance the performance of machine learning algorithms and reduce the number of selected features [21]. The Filter method has been applied in this study fig 2 illustrates the steps of the method used for the feature selection.20 features were used out of 85 features. The list of final features is available in Table 1.

DOI: http://dx.doi.org/10.25098/7.1.4



Figuar2: Filter method.

Table1: A feature selection list with the score of the algorithm

NO	Feature	Score
1	Severity	1075.132784
2	MAP < 70	860.068698
3	Age.1	438.294812
4	AgeScore	419.658824
5	MAP	320.882670
6	CrtnScore	254.673858
7	CrctProtein	230.067559
8	Procalciton > 0.1	220.153989
9	BUN > 30	207.063661
10	C-Reactive Prot > 10	187.751790
11	BUN	186.729807
12	D-dimer	149.145630
13	Procalcitonin	143.670952
14	D-Dimer > 3	126.428686
15	INR > 1.2	126.428686
16	IL6 > 150	117.882268
17	Troponin > 0.1	116.433243
18	LOS_Y	109.770169
19	O2 Sat < 94	106.813724
20	AST > 40	101.479708

3.2.2 Correlation Heat Map

This paper used Pearson’s correlation to find the correlation between the features and patient mortality. This method was used to test whether there is a strong relationship between the independent and dependent variables. The correlation coefficient is a statistical technique that measures the strength of the relationship between various variables. It can be used to determine the significance of a particular attribute in the data. According to the concept of this analytical technique, the relationship between the target and dependent variables can be determined by taking into account the strength of the correlation between the two. Table 2. Shows Severity, MAP < 70 display a high positive correlation with patient mortality. While O2 Sat < 94, AST > 40, CrctProtein shows a low correlation with patient mortality. Fig. 3 concludes that Severity, MAP < 70, Age.1, AgeScore, CrtnScore, CrctProtein, and Procalciton > 0.1 are the most important features that identify patient sufferers. The findings support the literature's assumption that COVID-19 is associated with an increased risk of death among older individuals or suffering from other comorbidities.

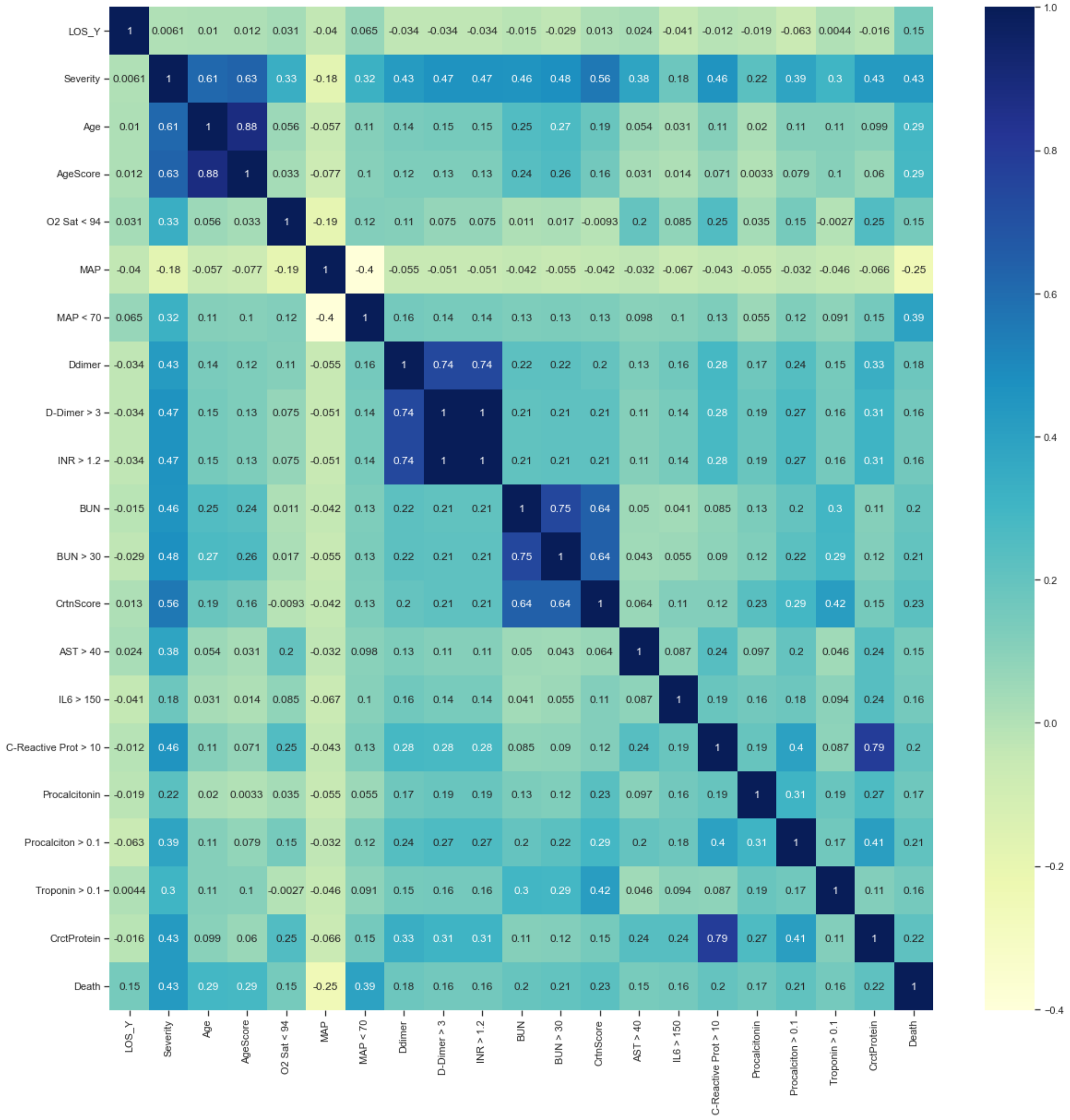


Figure 3: Correlation Heatmap shows the impact of features on variable death.

Table 2: calculate the correlation between variables and Death.

NO.	Features	Death correlation
1	Severity	0.431134
2	MAP < 70	0.392984
3	Age.1	0.291806
4	AgeScore	0.286053
5	CrtnScore	0.226512
6	CrctProtein	0.215827
7	Procalciton > 0.1	0.211338
8	BUN > 30	0.205231
9	C-Reactive Prot > 10	0.195811
10	BUN	0.195298
11	D-dimer	0.175214
12	Procalcitonin	0.172066
13	D-Dimer > 3	0.161698
14	INR > 1.2	0.161698
15	IL6 > 150	0.156275
16	Troponin > 0.1	0.155335
17	LOS_Y	0.150929
18	O2 Sat < 94	0.148929
19	AST > 40	0.145243
20	Creatinine	0.144933

3.3 Analytics Algorithms

After selecting the most desirable feature subset, various machine learning algorithms have developed a predictive model. Including Neural Networks, Decision Trees, Support Vector Machines (SVM), K-Nearest Neighbor (KNN), and Random Forst Classifiers.

3.3.1 Neural Networks

Machine learning algorithms are inspired by the dynamics and architecture of the brain's networks of neurons. These algorithms use highly idealized neuron models, which are fundamental to the principle of artificial neural networks. For instance, a neural network can learn to recognize structures in a data set. It can then create a training set containing input patterns and target values that represent the properties of the patterns it learns.

DOI: http://dx.doi.org/10.25098/7.1.4

3.3.2 Decision tree

Data-driven learning (DTs) is utilized for both regression and classification purposes [22] They can predict the values of a dependent variable by learning simple decision rules based on the data featured [23]. A DTs algorithm starts at the root node and goes through multiple split or internal nodes until it reaches the leaf. Data points should go to the right node if the binary tree goes there. Left until these points are in the appropriate leaves and should be left until the learning process is over. After the learning process is completed and the data has been tested, we can analyze the algorithm with unknown features (Khan et al.,2020).

3.3.3 K-Nearest Neighbors

The KNN algorithm is considered one of the most important algorithms in machine learning and data mining [24]. K-nearest Neighbors (KNN) is a learner model that takes into account the nearest values of various data sets. It performs time-series analysis to determine the Nearest Neighbours by examining the data set's closest past values. The KNN is a statistical measure that shows the smallest similarity between past and present cases [25].

3.3.4 Support Vector Machines

The Support Vector Machine is a type of machine-learning technique that can be used for classification and regression purposes. It has excellent generalization capabilities and is especially suited for small data sets [26]. It can be used to solve large-dimensional problems that are related to structural risk minimization. It was developed using the statistical learning theory to solve local and nonlinear minimum sequence problems. The system call can then use the sequence's short system frequency to convert it into a call with a specified length of vectors [27].

3.3.5 Random Forest

Random Forest is a type of supervised learning algorithm that is commonly used for solving various types of problems. It can be used in both classification and regression. This method is usually combined with other models to improve its overall accuracy.

3.4 K-fold Cross-Validation Technique

Cross-Validation is a process that involves identifying the model that will interpret the results of a statistical analysis in an independent data set. A resampling process known as cross-validation is performed on a model to test its correctness. It involves taking various chunks of data and testing them again and again. This is often utilized in situations where the goal is prediction. The user wants to know how well the model can perform in practice. The grid search algorithm was used to determine the optimal hyper parameters for some models The hyper parameters selected were those that gain the best performance values. A tenfold cross-validation procedure the method was used for evaluating the performance and the general error of total classification models, it’s widely used in data mining and machine learning due to its different approach from the conventional split-case method. It helps minimize the error rate and helps increase the number of data used for both validation and training. In addition, it guarantees protection against arbitrary experiment theory by not overfitting or overlapping the data between the validation and test [28].

DOI: http://dx.doi.org/10.25098/7.1.4

3.5 Framework Tools

For coding, the program python was used as a programming language, it's associated with Python machine learning packages libraries. numpy, pandas and scikit-learn library were used. Anaconda was used as an interpreter to simplify the deployment of the package. Jupyter notebooks were used to proceed with code lines in python.

4. RESULTS

4.1 Metrics Evaluation

The evaluation metrics used for ensemble models are designed to measure the reliability of the method. Some of these include precision, recall, F1-score, AUC, specificity, sensitivity, and confusion matrix.

ACC: The percentage of patients who are correctly predicted to die in the whole dataset is computed using the following equation.

$$ACC = (TP + TN) / (TP + TN + FP + FN)$$

Where TP (True positive) and TN (True negative) are the correct predictions.

FP (False positives) and FN (False negatives) are incorrect predictions.

Our goal is to decrease the number of these errors.

Precision: The percentage of patients who survived and were correctly identified by ML models is a measure of the accuracy of these models. It is computed using the following equation.

$$Precision = TP / (TP + FP)$$

Sensitivity/ Recall: The percentage of patients who survived and were correctly categorized by the classifiers is known as recall. This is because the model with a good recall has very few false negatives.

$$Sensitivity/ Recall = TP / (TP + FN)$$

F1-score: This metric is used to measure both recall and precision. It focuses on the cases of false negatives and false positives. A higher F1 score indicates that the models correctly classify these cases. The equation used to calculate it is

$$F1\ Score = 2 * (Precision * Recall) / (Precision + Recall)$$

Specificity: The specificity of a model or algorithm is the ability to predict the true negative of a given category. This is also referred to as the true negative rate in literature. The exact formula for this is shown below.

DOI: http://dx.doi.org/10.25098/7.1.4

Specificity =TN / (TN + FP)

AUC: The relationship between the true positive rate (TPR) and the false positive rate (FPR) is explained in the ROC curve. The AUC shows how well the ensemble can distinguish between the binary classes known as Dead and Survived.

Confusion Matrix: The concept of a confusion matrix is a representation of the various instances that are correctly classified by a model. The values of the columns indicate the correct prediction (the true positives and the true negatives), while the other values show the misclassified ones (the false positives and the false negatives).

4.2 Evaluation Model

With the help of advanced prediction techniques, COVID-19 patients can receive the most effective and personalized care possible. This can be done through the use of specialized equipment and monitoring. In this study, some algorithms such as K-Nearest Neighbors, Neural Networks, Random Forests, Support Vector Machines, and Decision Trees were used as state-of-path classifiers since they are used to award better results, the models were evaluated using a 10-fold Cross-Validation method. Data were split into training (80%) and testing (20%) using the train-test-split method. grid search was used to find the best hyperparameters for the models. A Confusion Matrix was used to represent the performance of algorithms.in addition, it can help identify the model misclassifies and provide insight into its performance. Figure.4 shows the Confusion Matrix of the Random Forest algorithm. It is distinguished that most of the cases have been specified accurately. the accuracy, precision, and other metrics were calculated by using the Confusion Matrix percentage of correct classifications. The result of these percentages is summarized in Table 3. The Random Forest AUC was shown in Figure.5 The Random Forests recorded the best results among all models. the percentage of correct classification which was made by algorithms can be known as Accuracy. The accuracy that results from Random Forests and 85%, and Neural Networks was 85%, K-Nearest Neighbours, Support Vector Machine, and Decision Trees displayed an accuracy of 83%,82%, and 84% respectively. The recall obtained by The Random Forests, K-Nearest Neighbors, Neural Networks, Support Vector Machines, and Decision Trees models was 96%,97%,94%,98%, and 97% respectively. The F1 scores showed by The Random Forests, K-nearest Neighbors, Neural Networks, Support Vector Machines, and Decision Trees models were 91%,89%,90%,89%, and 90% respectively. The AUC obtained by the Random Forest, K-nearest Neighbors, Neural Networks, Support Vector Machines, and Decision Trees models were 87%,82%,87%,85% and 82% respectively. The higher the AUC range means the better the rendering of the models.

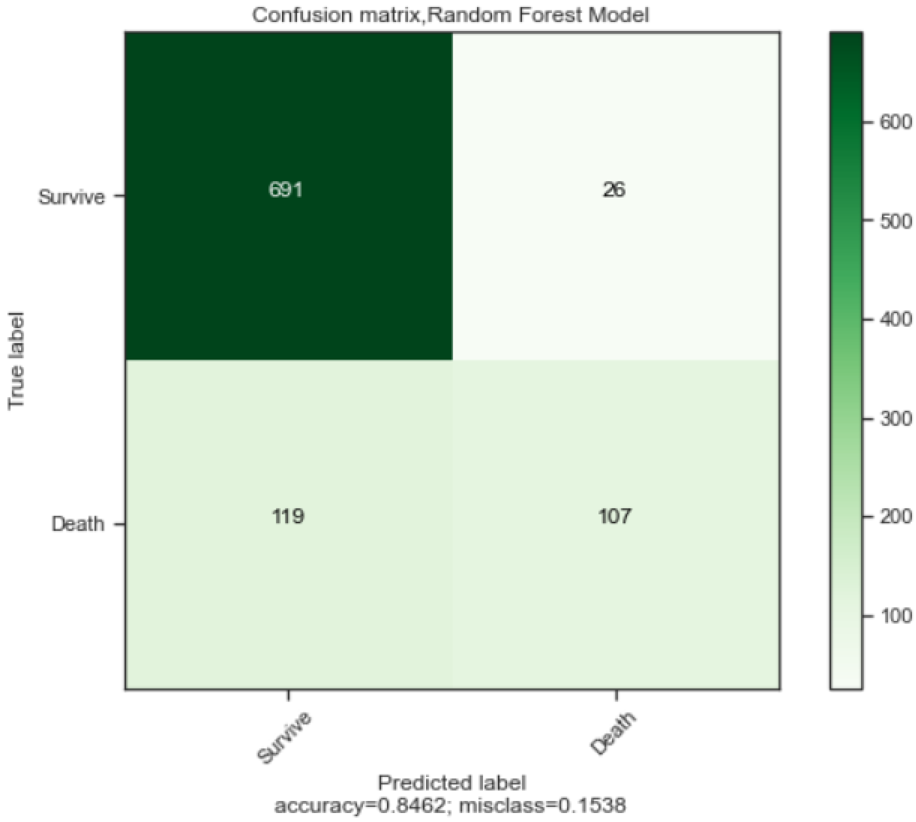


Figure 4: Random Forest Confusion Matrix for mortality prediction.

Table 3: Predictive performance of machine learning algorithms.

Model	Accuracy	Precision	Recall	F1-score	AUC
Random Forest	85%	85%	96%	91%	87%
K-Nearest Neighbors	83%	83%	97%	89%	82%
Neural Networks	85%	87%	94%	90%	87%
Support Vector Machines	82%	82%	98%	89%	85%
Decision Trees	84%	84%	97%	90%	82%

DOI: http://dx.doi.org/10.25098/7.1.4

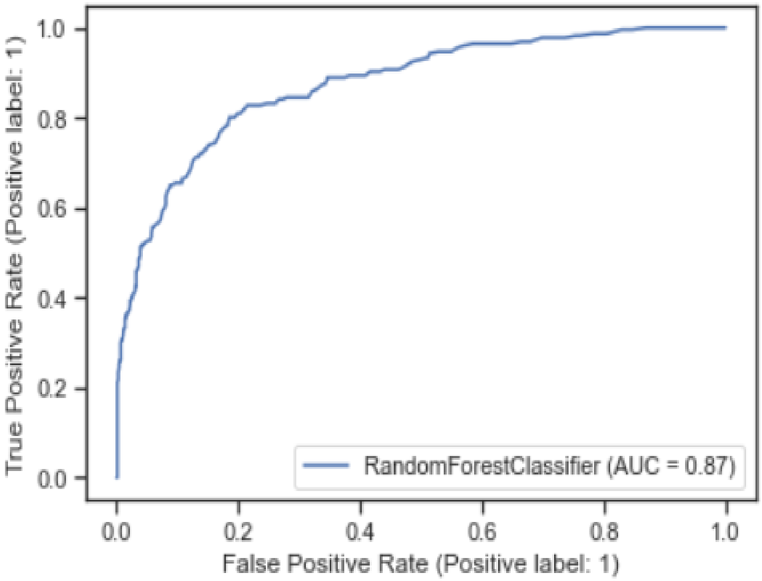


Figure: Random Forest AUC for mortality prediction.

5. DISCUSSION

This study aimed to analyze the data collected from COVID-19 patients and develop a machine learning algorithm that could predict their mortality rate. We performed a variety of metrics to evaluate the developed models. The evaluation results indicate that the models can accurately predict the mortality rate of COVID-19 patients. Other studies have also shown that the prediction of mortality rate using blood samples and clinical data can be performed.While, we performed a study on demographics, comorbidities, and the admission laboratory rate. Our study was able to achieve an accuracy of 85% of the Random Forests model and Neural Networks.

6. CONCLUSION

This study aimed to develop a predictive algorithm to help medical facilities and hospitals maximize the number of COVID-19 survivors. This tool would allow them to make more accurate and timely decisions regarding the treatment of these patients. Our algorithm can accurately predict the mortality rate of patients with COVID-19 using their physiological conditions, demographic information, and pre-exit conditions. It can help hospitals and caregivers identify those who need immediate attention and those who should be triaged before other patients arrive. The current study could eventually help improve the healthcare system’s response to an outbreak or pandemic.

DOI: http://dx.doi.org/10.25098/7.1.4

REFERENCE

[1] Worldometers. Covid-19 Coronavirus Pandemic [Online]. Available: https://www.worldometers.info/coronavirus/?utm_campaign=homeAdvegas1 Accessed:2022, September 26.

[2] R. Zagrouba, M.A. Khan, A.U. Rahman, M.A. Saleem, et al. (2021)." Modelling and simulation of covid-19 outbreak prediction using supervised machine learning " *Computers, Materials, & Continua*, vol.66, issue 3, pp.2397-2407.

[3] N. Alballa, I. Al-Turaiki. (2021)." Machine learning approaches in COVID-19 diagnosis, mortality, and severity risk prediction: A review" *Informatic in Medicine Unlocked*, vol.24, pp. 100564.

[4] W. Hou, Z. Zhao, A. Chen, H. Li, et al. (2021)" Machining learning predicts the need for escalated care and mortality in COVID-19 patients from clinical variables " *International journal of medical sciences*, vol.18.issue 8, pp. 1739-1745.

[5] C. Gazzaruso, E. Paolozzi, C. Valenti, M. Brocchetta, et al. (2020). " Association between antithrombin and mortality in patients with COVID-19. A possible link with obesity " *Nutrition, metabolism, and cardiovascular diseases*, vol.30, issue 11, pp. 1914-1919.

[6] S. Dargan, M. Kumar, M.R. Ayyagari, G. Kumar. (2020). "A Survey of Deep Learning and Its Applications: A New Paradigm to Machine Learning" *Archives of Computational Methods in Engineering*, Vol. 27, issue 4, pp.1071–1092.

[7] S. Quazi, (2022). " Artificial intelligence and machine learning in precision and genomic medicine ", *Med Oncol*, vol. 39, issue 8. pp.120.

[8] I.E. Agbehadji, B.O. Awuzie, A. B. Ngowi, R.C. Millham. (2020). "Review of Big Data Analytics, Artificial Intelligence and Nature-Inspired Computing Models towards Accurate Detection of COVID-19 Pandemic Cases and Contact Tracing." *International Journal of Environmental Research and Public Health*, vol. 17, issue 15, pp.5330.

[9] M.Smith, F. Alvarez. (2021). " Identifying mortality factors from Machine Learning using Shapley values - a case of COVID19" *Expert Syst Appl*, vol. 176, pp. 114832.

[10] C.Hu, Z. Liu, Y. Jiang, O. Shi, et al. (2020). "Early prediction of mortality risk among patients with severe COVID-19, using machine learning " *International Journal of Epidemiology*, vol. 49, issue 6, pp.1918-1929.

[11] M.Daniyal, R. O. Ogundokun, K. Abid, M. D. Khan, et al. (2020)."Predictive modeling of COVID-19 death cases in Pakistan" *Infectious Disease Modelling*, *5*, 897–904.

[12] R.M. Arias Velásquez, J.V. Mejía Lara. (2020). "Forecast and evaluation of COVID-19 spreading in USA with reduced-space Gaussian process regression" *Chaos, solitons, and fractals*, vol. 136, pp.109924.

[13] A.K. Das, S. Mishra, S. Saraswathy Gopalan. (2020). " Predicting CoVID-19 community mortality risk using machine learning and development of an online prognostic tool" *Peer J*, vol. 8, pp. 10083.

[14] KC. Wong KC, Y. Xiang, L. Yin, HC.So.(2020) "Uncovering Clinical Risk Factors and Predicting Severe COVID-19 Cases Using UK Biobank Data: Machine Learning Approach" *JMIR Public Health Surveill*. vol.7, issue 9. pp. e29544.

DOI: http://dx.doi.org/10.25098/7.1.4

[15] MT. García-Ordás, N. Arias, C. Benavides, García-Olalla, et al. (2020). "Evaluation of Country Dietary Habits Using Machine Learning Techniques in Relation to Deaths from COVID-19" *Healthcare (Basel, Switzerland),* vol.8, issue 4, pp. 371.

[16] T. Saba, I. Abunadi, MN. Shahzad, AR. Khan. (2021)." Machine learning techniques to detect and forecast the daily total COVID-19 infected and deaths cases under different lockdown types ", *Microscopy research and technique,* vol.84, issue 7, pp. 1462–1474.

[17] M.Pourhomayoun, M. Shakibi. (2021). "Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making" *Smart Health (Amst),* vol.20, pp.100178.

[18] S. Subudhi, A. Verma, A.B. Patel, c. Hardin, et al. (2021). "Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19" *NPJ Digit Med,* vol.4, issue 1, pp.87.

[19] Kaggle. (2022, January 20). Mortality risk clinical data of covid19 patients. [Online]. Available:https://www.kaggle.com/datasets/harshwalia/mortality-risk-clinincal-data-of-covid19-patients. Accessed:2022, March 19.

[20] B. Abdollahzadeh, F.S. Gharehchopogh. (2021). *" A multi-objective optimization algorithm for feature selection problems"* Eng. with Comput, vol.38, issue suppl 1, pp. 1845–1863.

[21] A.A. Ewees, L. Abualigah, D. Yousri, Z.Y. Algamal, et al. (2022). "Improved Slime Mould Algorithm based on Firefly Algorithm for feature selection: a case study on QSAR model" Eng. with Comput, vol.38, issue (Suppl 3), pp. 2407–2421.

[22] T. Saba, (2021). " Computer vision for microscopic skin cancer diagnosis using handcrafted and non-handcrafted features " *Microscopy research and technique,* vol. 84, issue 6, pp. 1272-1283.

[23] M.A. Khan, M. Sharif, T. Akram, M. Raza, et al. (2020)." Hand-crafted and deep convolutional neural network features fusion and selection strategy: An application to intelligent human action recognition "Applied *Soft Computing, vol. 87,* pp.105986.

[24] H. Wang, p. Xu, J. Zhao. (2021)."Improved KNN Algorithm Based on Preprocessing of Center in Smart Cities " *Complexity,* vol. 2021, pp.

[25] R. G. da Silva, M. Ribeiro, V. C. Mariani, L. Coelho. (2020). ."Forecasting Brazilian and American COVID-19 cases based on artificial intelligence coupled with climatic exogenous variables " *Chaos, solitons, and fractals,* vol.139, pp.110027.

[26] S. A. Khan, M. Nazir, M. A. Khan, T. Saba, et al. (2019). " Lungs nodule detection framework from computed tomography images using support vector machine" *Microsc Res Tech,* vol.82, issue 8, pp.1256–1266.

[27] B. Gaye, D. Zhang, A. Wulamu. (2021). " Improvement of Support Vector Machine Algorithm in Big Data Background ", *Mathematical Problems in Engineering,* vol.2021, issue 1 pp. 1-9.

DOI: http://dx.doi.org/10.25098/7.1.4