International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 14, No.1 (Aug-2023)

http://dx.doi.org/10.12785/ijcds/140143

Prediction of Need for ICU Admission and Mortality of Covid-19 patients using Machine Learning: A Comparative Analysis

Uzmat Ul Nisa[1] and Mohammad Ahsan Chishti[2]

[1]*Department of Information Technology, National Institute of Technology, Srinagar, India*
[2]*Department of Computer Science & Engineering, National Institute of Technology, Srinagar, India*

Received 02 Jan. 2023, Revised 09 Jul. 2023, Accepted 12 Jul. 2023, Published 01 Aug. 2023

Abstract: The novel coronavirus disease (COVID-19) has caused severe damage worldwide, affecting the lives of millions of people and destroying the global infrastructure and health systems. The timely prediction of a patient's mortality risk can facilitate the health care systems to learn about the patients that are going to become severe and offer timely medical care to those patients, thereby reducing mortality and the burden on the health systems. This will also ensure the optimal allocation of resources in hospitals. Machine Learning can prove very helpful in this prediction of mortality. We have evaluated five different machine learning algorithms to predict the need for ICU admission and mortality of Covid-19 patients using two different datasets and identified the most significant features. This identification of significant features among an array of available features helps identify the patients at higher risk of severity and mortality. We have also compared the significant predictors of mortality from two datasets from the US and Mexico to analyze the effect of the infection on different populations. It was found that Random Forest achieves the best performance in the classification task, followed by Logistic Regression. Therefore, Random Forest's predictive model can be helpful for clinicians to prioritize patients appropriately.

Keywords: Covid-19, Mortality, ICU admission, Machine learning, Random Forest

1. INTRODUCTION

The upsurge of the new pandemic, Covid-19, has led to devastating effects on people worldwide. The Covid-19 disease, which has been attributed to the newly identified virus, SARS nCoV2 (Severe Acute Respiratory Syndrome novel Coronavirus-2), is a highly transmissible disease. The first reports of the disease came from China on December 31, 2019. Since then, the disease has caused havoc globally, from people's lives to the global economy. The rapid spread and high severity of the disease prompted the World Health Organization (WHO) to declare a pandemic on March 11, 2020 [1]. As of September 6, 2022, the number of confirmed cases has risen to more than 603 million, with more than 6.4 million deaths worldwide [2]. The causative agent, SARSnCoV2, has shown variations over time, like other viruses. These variations may have little impact on the speed of its spread and how severe the disease can be [3]. However, the phenotype of Covid-19 ranges from mild or asymptomatic and intermittent recovery to systemic collapse or death [4], [5], [6], [7]. Therefore, accurate and rapid diagnosis of severe cases is an important task. However, the accurate prediction of the outcome of the patients is quite challenging when the phenotype has such a range of clinical manifestations. This problem poses a considerable challenge to the prognosis and proper triage of patients when admitted to the hospital. The RT-PCR is the most reliable diagnostic test available for detecting Covid-19. However, this test is cumbersome as it requires considerable human resources and may take hours to days to get the results [8]. In this regard, many researchers have explored machine learning models to detect Covid-19 using medical images of the suspected individuals. They have employed deep learning models to detect the Covid-19 disease from medical images such as X-rays and CT scans [9], [10], [11], [12], [13]. Although they provide high-quality detection for Covid19, they lack the ability to distinguish the severity levels of the patients. During the peak of the pandemic waves, hospitals have been facing a deficit in crucial resources like care equipment, oxygen beds, and other necessities. The rapid surge of COVID-positive patients, particularly in countries with moderate to low levels of income, has left health systems overburdened and in dire need of additional resources [14], [15], [16], [17]. This can be averted by prioritizing the patients with a higher risk of severity or mortality so that these patients receive immediate medical aid or hospitalization, while

E-mail address: uzmatulnisa@gmail.com, ahsan@nitsri.ac.in

those with a lower severity or mortality risk are treated on an outpatient or self-quarantine basis [18], [19]. For this purpose, there should be a proper prognostic model capable of predicting the severity of illness and risk of mortality to ensure optimal resource allocation and stratified triage of the patients. In addition, early and accurate identification of the patient's feature variables that are most responsible for advancing to severity is required to ensure patient prioritization. The paper is organized as follows in the ensuing sections: section 2 discusses the related work. In section 3, we have presented the datasets, the models, and the evaluation metrics that we have used. Section 4 discusses the data preprocessing and classification parameters. Section 5 presents the experimental details, while section 6 outlines the corresponding results. The paper is concluded in section 7.

2. RELATED WORK

The diagnosis and prognosis of diverse medical conditions and ailments have become increasingly reliant on machine learning applications, which have now become indispensable clinical tools. Predicting the severity or mortality of Covid-19 patients is mainly a classification task, therefore, supervised machine learning algorithms are used in this problem. The term 'severity' has different interpretations, where some studies link it to the need for intensive care [20], and some relate it to death [21]. At the same time, some define it according to the specifications from the national health board [22].

Some researchers performed meta-analyses on Covid-19 positive patients and reported that people with hypertension [23], [24], diabetes [25], [26], and cerebrovascular and cardiovascular diseases [27] are at more risk of severity or mortality. Some studies such as [28] and [29] performed systematic analyses on the clinical characteristics of Covid positive patients and found various biomarkers such as elevated levels of troponin, Interlukin, and LDH, and depressed levels of total lymphocytes and albumin are indicators of high severity or mortality. Several studies collected patient information such as demographics, clinical data, underlying comorbidities, laboratory test results, and medical images. They used various machine learning models to predict the severity or mortality of the patients and identify the significant severity or mortality predictors. Tingting Dan et al. [20] collected the relevant information of 733 patients and used SVM to predict the need for ICU, death even after being admitted to ICU, and length of ICU stay in case of survivors. They achieved an accuracy of 92% for mortality prediction. In their study, Darapaneni et al. [30] employed Logistic Regressor (LR), Decision Trees (DT), Random Forest (RF), and Ensemble classifiers to accomplish two objectives. The first objective was to predict confirmed cases among suspected ones based on their clinical records, while the second objective was to predict the ward (general, semi-ICU, or ICU) where positive cases identified in the first task would be admitted. The RF model used the Ginni score to identify the predictor variables for the high severity of the disease and achieved the highest testing accuracy of 94.8%. However, the data collected did not contain essential features such as D-dimer and potassium levels of the suspected individuals. Li X et al. [31] employed deep learning that predicted the requirement for ICU admission and mortality of Covid-19 patients and obtained 85.3% accuracy and 75% sensitivity. They used the Gini feature of importance to find the significant predictors of mortality. Ezz M et al. [32] employed Extreme Gradient Boosting to predict the need for ICU admission and achieved 97% accuracy and 96% sensitivity. El-Shafeiy E et al. [33] selected features from the collected data using Quick Redundant Feature Selection (QRFS) technique and trained the quantum neural network to predict the severity of Covid-positive patients. Jianhong K et al. [34] employed Artificial Neural Networks to predict different severity levels of Covid-19 patients once admitted to the hospital. According to them, low albumin and high globulin levels, and blood urea nitrogen are the main risk factors for higher severity. Di Castelnuovo A et al. [35] were able to obtain 83.4% accuracy and 95.25% sensitivity using Random Forest for predicting the mortality in Covid positive patients. Using Permutation Feature Importance, they found eGFR, CRP, and age are the main mortality predictors.

The authors of [36] collected data of 10,237 patients and evaluated different machine learning models such as KNN, SVM (rbf and linear), Lasso, and RF to predict Covid-19 mortality. They obtained the highest accuracy of 91.9% and the highest sensitivity of 92% using linear-SVM. They used Lasso and RF separately to find the significant mortality predictors. According to Lasso and RF, old age, Diabetes Mellitus and cancer and age, infection route, and hypertension are the main mortality predictors, respectively. In their work, Han et al. [21] introduced a neural network system called the Broad Learning System, which is designed to predict mortality in Covid-19 patients and achieved 94.64% accuracy and 94.5% sensitivity. Chowdhury et al. [37] collected data from 375 patients to build LR classfier to assess the risk of death due to Covid-19. The authors of [38] also used various machine learning models such as GBDT and LR for mortality prediction. They used LR to find the significant mortality predictors. Elham Jamshidi et al. [39] used LR and RF models for this purpose. The authors of [40] used SVM, LR, and XGBoost to predict mortality.

Table I summarizes the works of different studies that have used machine learning to prognosticate the probability of death in Covid-infected patients. These studies have employed different supervised machine learning models. However, the most frequently used include RF, SVM, XGBoost, and LR. These studies have been conducted in different nations for different periods. In most of these studies, the size of the dataset is very small, with only 250 - 500 data records. In addition, the available data is typically imbalanced, with a more substantial number of patients who have recovered than those who have died. While certain studies have produced models with acceptable accuracy,

TABLE I. Summary of Machine Learning Models Used for Mortality Prediction of Covid-19 Patients

Work	Dataset Modality	Number of Cases	Algorithm used	Techniques to identify significant predictor variables	Mortality Predictor Variables	Performance
[35]	Demographics, Laboratory tests, clinical notes	3894	RF	Permutation Feature Imortance (PFI)	Glomerular Filtration rate(e-GFR), C-Reactive Protein(CRP), Age	Accuracy=83.4% Sensitivity=95.2% Specificity=30.8% F1 score=90.4%
[31]	Demographics, Chronic comorbidities, laboratory tests	1022	Deep learning model (5 hidden layers)	Ginni feature of importance	Age, LDH, Oxygen saturation (SpO2), CRP, Procalcitonin, Cardiac Troponin	Accuracy= 85.3% Sensitivity=75.0% Specificity=87.2% F1 score=61.6%
[20]	Demographics, Clinical and laboratory test results	733	SVM (kernel-poly)	Recursive Feature Elimination	Lymphocyte Absolute value, D-dimer, Albumin, Respiratory rate, LDH, Adenosine deaminase, Direct Bilirubin (DB)	Accuracy=92% AUC=0.98
[36]	Demographics and medical information	10,237	LASSO, Linear SVM, RBF-SVM, RF, KNN	LASSO and RF	Lasso: Old age, Diabetes Mellitus(DM), Cancer RF: Age, Infection Route, Hypertension	Accuracy: LASSO=91.1% Linear SVM=91.9% RBF-SVM=70.2% RF=65.4% KNN=85.4% Sensitivity: Lasso=90.7% Linear SVM=92% Specificity: Lasso=91.4% Linear SVM=91.8%
[21]	Demographics, laboratory test results, symptoms	375	Broad Learning system	NA	NA	Accuracy=94.64% Sensitivity=94.50% Specificity=94.80% AUC=98.84%
[38]	Clinical, demographic, laboratory, radiological data	2924	GBDT, LR, LR-5	Logistic Regression	LDH, BUN, Lymphocyte(%), Age, Interlukin	Accuracy: GBDT =88.9% LR=86.8% LR-5=88.7% Sensitivity: GBDT=89.9% LR=87.8% LR-5=89.8% Specificity: GBDT=78.8% LR=76.9% LR-5=77.1%
[41]	Epidemiological, demographic, clinical, laboratory data	485	XGBoost	XGBoost	Lymphocyte, LDH, and CRP	Accuracy =90%
[39]	Laboratory indicators, demographics	263	LR and RF	RF	BUN, Creatinine, albumin, gender, age, Red cell distribution width(RDW),INR (International Normalized Ratio)	Sensitivity: RF=70% LR=65% Specificity: RF=75% LR=70%
[40]	Demographics, Clinical	3841	SVM, Logistic Regression, XGBoost	Recursive Feature Elimination, XGBoost	Age, SpO2, type of patient encounter	AUC=0.91
[37]	Demographics, clinical, laboratory tests	375	Logistic Regression	XGBoost	LDH, Neutrophils(%), Lymphocyte(%), CRP, Age	AUC=0.991

they have not been able to achieve higher sensitivity [31] [39]. Achieving high sensitivity in predicting mortality is essential as predicting a low risk of death (false negative) for a patient who is actually at a higher risk of death can lead to unfavorable results.

3. MATERIALS AND METHODS

A. Data

In this study, we have worked on two datasets, one from Xiaoran Li et al. [31], obtained from Stony Brook University Hospital, NY, USA, and another dataset from [42], released by the Mexican government. The US dataset [31] consists of patient records with demographic information such as age, gender, nationality; comorbidities such as hypertension, asthma, COPD; clinical notes such as fever, sore throat, chest pain, sputum; and laboratory tests details like LDH, ALT, D-dimer, Lymphocytes, procalcitonin. The US dataset contains two separate datasets; a) US dataset_1: for predicting the need for ICU admission (containing 1106 records and 43 features (target label included)), and b) US dataset_2: for predicting the mortality (containing 1020 records and 43 features (target label included)). In these datasets, nearly 43% are women and 57% are men; about 29% of men and 19% of women are admitted to ICU. About 27% of the patients are diabetic and 48% are having hypertension. The Mexican dataset [42], accessed on 23 February 2021, consists of 10,48,575 patient records containing patients' demographic and comorbidities information. The demographic information includes age, gender, residence, and nationality; and comorbidities include diabetes, hypertension, obesity, renal chronic disease. This dataset consists of 40 features (including columns of 'icu' and 'death'). In this dataset, nearly 12% of patients are diabetic and 16% are having hypertension.

B. Methodology

To account for missing values in the US dataset, the predictive mean modeling technique was implemented, whereas for the Mexican dataset, columns or features with missing data exceeding 20% were removed due to the dataset's substantial size. We have divided the problem into three parts:

- To predict the need for ICU admission and identify the significant predictors for ICU admission
- To predict the mortality of Covid patients and identify the significant predictors of mortality, and
- Perform a comparative analysis of the predictor variables of mortality in the two datasets

We aim to categorize patients based on their need for ICU and determine their likelihood of mortality. To achieve this, we have opted for five supervised learning models. These algorithms are simple to implement and are robust in nature. Our inclination is towards machine learning models instead of deep learning for prediction and classification, primarily to decrease computational complexities. Additionally, machine learning models provide us with the capability to obtain insights into our data, such as feature importance, using algorithms like Random Forest.

1) K-Nearest Neighbor (KNN) [43] [44]: It uses the concept of feature similarity to predict the category of a new data point. KNN [45] being a non-parametric algorithm does not make any hypothesis on the training data and therefore does not derive any pattern or fit a curve in it. It stores the training data, and at the time of prediction, it finds the k closest neighbors and predicts the class of a new data point on the majority voting principle. Fitriyadi et al. [46] have used KNN to predict the degree of Covid-19 dissemination and [47] have implemented KNN to predict the status of infected patients.
2) Logistic Regression (LR) [48] [49] [50]: The algorithm is a predictive analysis tool that operates on probabilities. Logistic Regression (LR) is its basic form and deals with predicting binary outcomes, i.e., either 0 or 1. By employing a logistic function, it maps the input data to these two values. LR is effective when data is linearly separable, but can result in overfitting if the dataset has more features than records.
3) Support Vector Machine (SVM) [51] [52] [53] [54]: SVM represents distinct classes in a multi-dimensional space using a hyperplane. It maps data points onto an n-dimensional plane, with n signifying the attribute count in the dataset, and each coordinate on the plane representing a feature of the data point. The algorithm endeavors to identify a hyperplane that effectively separates the two classes and assigns the class of a new data point according to which side of the hyperplane it belongs to. It works pretty well with a higher number of features.
4) Decision Tree (DT) [55] [56] [57] [58]: Decision Tree is a structure resembling a tree where the internal nodes are known as decision nodes as they are used to make decisions. The leaf nodes denote the output. At each split of the tree, a decision is made based on the feature taken into consideration. The order of attributes to be taken as a root node or decision node is based on statistics that measure a particular attribute's importance using measures such as entropy, information gain, or Gini score.
5) Random Forest (RF) [59] [60] [61] [62]: It is an ensemble technique that employs a collection of decision trees to solve complex problems and provide improved model performance. The decision trees in a Random Forest take different overlapping subsets of the training dataset and, based on majority voting on the predictions of decision trees, RF provides the output. RF usually obtains higher accuracy when the number of trees is greater.
In order to determine the variables that are most predictive in the classification task, we have used

the RF algorithm that, by default, uses the Gini score [63] to find the most important variables that lead to the higher prediction accuracy [64]. The lower the Gini score, the more important the variable is.

4. MODEL EVALUATION

The dataset is partitioned into training and testing sets at a ratio of 75:25 to assess the models' performance. The classifier's true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) help us to calculate the following measures.

Accuracy: It represents the percentage of the model's total predictions that were accurate [65].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Sensitivity (or recall): This pertains to the true positive rate, which represents the proportion of correct positive predictions out of all the positive predictions made [65].

$$Sensitivity = \frac{TP}{TP + FN} \quad (2)$$

In the case of mortality prediction, it tells us what percentage of patients that actually died was predicted in the death category. High sensitivity is quite essential in mortality prediction or prediction of the need for ICU because identification of patients at higher risk is more important than reducing false positive prediction.

Specificity: It is a measure of the model's ability to predict true negatives in each available category [65].

$$Specificity = \frac{TN}{TN + FP} \quad (3)$$

It tells us what percentage of patients that did not die, were predicted by the model in the 'no_death' category.

Precision: This pertains to the ratio of positive predictions that are accurately correct [65].

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

This measure indicates the percentage of patients in the "death" category that were accurately predicted in the 'death' category.

F1 Score: It tries to find a steadiness between precision and sensitvity. It is measured as the harmonic mean of the two [66].

$$F1Score = 2 * \frac{Precision * Sensitivity}{Precision + Sensitivity} \quad (5)$$

AUC: The AUC-ROC curve is a widely used method of assessing the efficacy of a binary classifier, with the true positive rate plotted against the false positive rate. The AUC-ROC score, which is the area under this curve, provides a gauge of the classifier's capacity to differentiate between the classes. An AUC score ranging from 0 to 1 is possible, with a higher value indicating a superior classifier [65].

5. EXPERIMENTAL SETUP

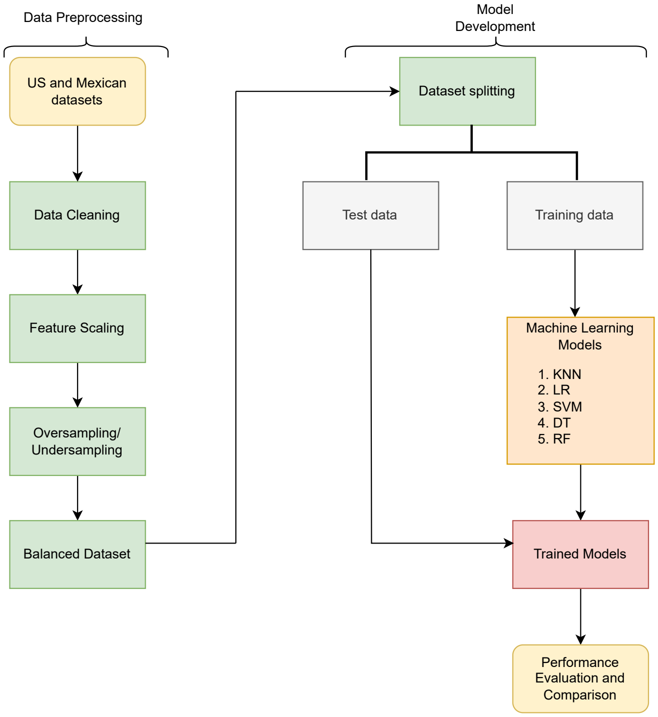
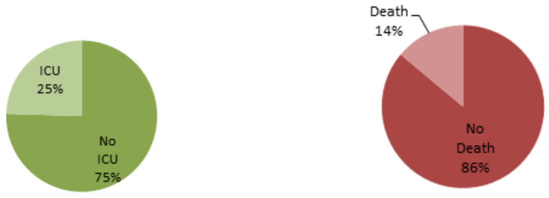


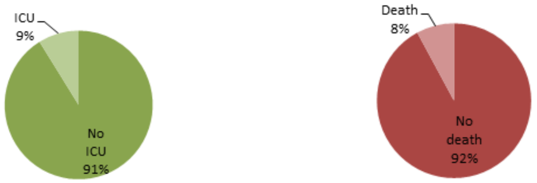
Figure 1. Proposed Processing model

Figure 1 shows our proposed processing model with different steps of data pre-processing and model development. The two US datasets do not contain any missing values as the missing values have been imputed by predictive mean modeling by [31]. The Mexican data had a lot of missing data for many columns. The columns with more than 20% missing values were removed. Moreover, the rows containing missing data were also removed. Since the dataset is quite large, therefore discarding records with missing values would not affect the model performance. This step of eliminating rows with missing values is required as many machine learning models do not support it, leading to biased results otherwise. We used feature scaling for both the Mexican and US datasets to make all the data values present in the same range. For this purpose, all the data features are scaled using min_max normalization of data to rescale all the data values in the range of [0, 1]. An imbalance was found in the distribution of categories in the "Death" and "ICU" columns in the US dataset. Figure 2a clearly shows the imbalance in the class distribution in US dataset_1 as Class 0 ("no ICU") has 875 data samples, and Class 1 ("ICU") has only 271 data samples. Similarly, in US dataset_2 (Figure 2b), category 0 ("no death") has 878 data samples, and category 1 ("death") has 142 data samples. To balance the data in both cases, we applied data oversampling to make the number of minority-class samples equal to



(a) (b)

Figure 2. US Data: Class distribution for (a) ICU admission (b) Mortality



(a) (b)

Figure 3. Mexican Data: Class distribution for (a) ICU admission (b) Mortality

the majority-class samples. Therefore, after oversampling, class 0 and class 1 each consist of 875 samples in the US dataset_1 and 878 samples in US dataset_2. Table II represents the number of data samples before and after sampling of the US dataset.

From Figure 3, we can depict that the Mexican dataset also has a class imbalance in both cases. Hence, we performed the over and undersampling of the dataset. After data cleaning and keeping the ’icu’ feature as the target label in the Mexican dataset, the dataset contains 166796 records with 28 features, including the target label. It consists of 152010 data samples of class 0 (”no ICU”) and 14786 data samples of class 1 (”ICU”). After oversampling, both classes consist of 152010 data records. Table III represents the number of data samples before and after sampling of the Mexican dataset. In all cases of oversampling, we have used Random Oversampler (in python) [67]. RandomOverSampler uses a naive oversampling technique that randomly duplicates the data records in the minority dataset so that both the classes of the dataset contain an equal number of data records. While keeping the ’death’ feature as the target label, the Mexican dataset contains 833457 records with 29 columns, including the target label. It consists of 767544 data samples of class 0 (”no death”) and 65913 samples of class 1 (”death”). In this case, the size of the dataset is quite large, we undersample the dataset to reduce the computational time complexities. We performed undersampling of class 0 using Near_miss (version 3) undersampler [68] so that each class contains 65913 samples of data. Near_Miss UnderSampler uses the knn technique to eliminate the data points from the large class distribution. It calculates the distance between the closer data points from the two class distributions and eliminated the closer data points from the majority class.

A. Classification

In each experiment, we divided the data into a training set of 75% and a test set of 25%. To optimize the classification performance, we employed hyperparameter tuning with 10-fold cross-validation using GridSearchCV for each machine learning model. By using a grid of parameters, GridSearchCV identified the optimal parameters for a given model and training dataset.

1) Experiment 1

1) Need for ICU prediction (US dataset): To determine the model with the optimal performance, we employed five distinct machine learning models. KNN gives the best results when k is set to 1. The optimal performance for LR is achieved by setting the value of C to 10, max_iter to 100, and using the L1 penalty. SVM shows the best performance with the rbf kernel. Decision Trees show the best performance with default parameters, and RF shows the best performance with n_est set to 19 and all other parameters set to default values.
2) Mortality prediction (US dataset): To determine the model with the optimal performance, we employed five distinct machine learning models. KNN shows its best performance when K is set to 1; logistic regression shows the best performance with parameters C =100, the penalty set to L2, and max_iter=100. SVM shows the best results with kernel ’rbf’, the Decision tree shows the best results with default parameters, and random forest gives the best performance when ’n_est’ is set to 19, and all other parameters are default.
We have used the Random Forest algorithm to find the significant features for each case, that is, for ICU admission and mortality prediction.

2) Experiment 2

1) Need for ICU prediction (Mexican dataset): The best performance by these models is achieved when K is set to 1 in KNN; C=0.01, max_iter =100 and penalty=L2 for LR model; rbf kernel in SVM; default parameters in Decision Trees and in Random Forest, n_est=19, and all other parameters are default.
2) Mortality prediction (Mexican dataset): For mortality prediction, the best performance is achieved when k=25 in KNN; default parameters in all other models.
We have used the Random Forest algorithm to find the significant features for each case, that is, for the prediction of the need for ICU admission and mortality prediction.

TABLE II. Data Sampling in the US dataset

	ICU		Mortality	
	Class 0	Class 1	Class 0	Class 1
Before Sampling	875	271	878	142
After Sampling	875	875	878	878 (oversampling)

TABLE III. Data Sampling in the Mexican dataset

	ICU		Mortality	
	Class 0	Class 1	Class 0	Class 1
Before Sampling	152010	14786	833457	767544 (oversampling)
After Sampling	152010	152010	65913 (undersampling)	65913 (undersampling)

3) Experiment 3

1) We conducted a comparative examination of mortality prediction between US and Mexican datasets. Our analysis involved utilizing comparable input features that only contained information concerning the demographic and underlying comorbidities of Covid-positive patients. To identify the crucial predictors of mortality in both datasets, we employed the Random Forest algorithm.

6. RESULTS AND DISCUSSION

TABLE IV. US data: Confusion Matrices for ICU admission prediction

	TP	TN	FP	FN
KNN	156	195	52	15
LR	140	118	68	92
SVM	158	199	50	11
DT	148	198	60	12
RF	165	194	43	16

TABLE V. US data: Confusion Matrices for mortality prediction

	TP	TN	FP	FN
KNN	192	216	31	0
LR	174	177	49	39
SVM	201	216	22	0
DT	190	215	33	1
RF	200	214	23	2

TABLE VI. Mexican data: Confusion Matrices for ICU admission prediction

	TP	TN	FP	FN
KNN	35396	37830	2679	100
LR	27619	27063	10456	10867
SVM	29486	27832	8589	10098
DT	35572	37921	2503	9
RF	36374	37928	1701	2

TABLE VII. Mexican data: Confusion Matrices for mortality prediction

	TP	TN	FP	FN
KNN	10790	9490	5622	7055
LR	9854	10961	6558	5584
SVM	8776	11553	7636	4992
DT	9666	9675	6746	6870
RF	10428	10102	5984	6443

A. Experiment 1: US data: Need for ICU admission and Mortality prediction

Tables IV and V represent the confusion matrices obtained in the classification task of ICU and mortality prediction respectively in the case of US data. Tables VIII and IX show the results of the prediction of the need for ICU admission and mortality, respectively, in the case of US data. Since the dataset used is the same, we have compared the results of the models used in our paper with [31], as shown in tables VIII and IX. Li et al. [31] utilized a deep learning model to anticipate both the requirement for ICU admission and death in patients infected with Covid. Their risk score model facilitated the evaluation of the possibility of the need for ICU and mortality based on specific clinical factors of an individual. The Random Forest model used in our paper has outperformed the deep learning model [31] and other models used in this paper. The RF model achieves an accuracy of 89% and sensitivity of 93% for ICU prediction and 97% accuracy and 100% sensitivity for mortality prediction. We have plotted various features against their Ginni values to depict their significance in the prediction process. Figures 4 and 5 show the various significant features for ICU admission and mortality prediction, respectively, in order of their

TABLE VIII. US data: Performance indices for icu admission prediction

	Accuracy	Sensitivity	Specificity	Precision	F1 Score	AUC
KNN	84	93	75	79	85	0.84
LR	62	57	67.3	64	60	0.69
SVM	85	95	75.9	80	87	0.91
DT	84	95	73.55	78	86	0.82
RF	**89**	**93**	**85**	**86**	**89**	**0.96**
Deep Learning [31]	72	76	71	43	55	0.73

TABLE IX. US data: performance indices for mortality prediction

	Accuracy	Sensitivity	Specificity	Precision	F1 Score	AUC
KNN	93	100	86	87	93	0.93
LR	80	82	78	78	80	0.87
SVM	95	100	90	91	95	0.97
DT	92	100	85	87	93	0.93
RF	**97**	**100**	**95**	**95**	**98**	**1**
Deep Learning [31]	85	75	87	52	62	0.84

TABLE X. Mexican data: performance indices for icu admission prediction

	Accuracy	Sensitivity	Specificity	Precision	F1 Score
KNN	96	100	93	93	96
LR	72	72	72	72	72
SVM	76	73	78	77	75
DT	97	100	94	94	97
RF	**98**	**100**	**96**	**96**	**98**

TABLE XI. Mexican data: performance indices for mortality prediction

	Accuracy	Sensitivity	Specificity	Precision	F1 Score
KNN	62	57	66	63	60
LR	**63**	**66**	**60**	**62**	**64**
SVM	62	70	54	61	65
DT	59	59	58	59	59
RF	62	61	63	63	62

significance. We find that the main five significant predictors of the need for ICU admission are Lactate Dehydrogenase (LDH), Procalcitonin, C-Reactive Protein (CRP), Ferritin, and Oxygen Saturation (SpO2). The top five significant predictors of mortality are Age, Procalcitonin, D.dimer, LDH, and CRP. The higher levels of LDH are indicative of severe tissue damage [29]. Elevated levels of CRP and ferritin show inflammation in the patient's body [29], and elevated levels of procalcitonin usually are associated with a high bacterial or viral infection. Low values of SpO2 indicate less oxygenated blood in the body. Higher levels of D.dimer indicate high blood clots in the body. Therefore, the changes in the normal levels of these biomarkers indicate the severity level of the Covid-positive patients. Moreover, from Figure 5, we can see that older Covid-infected people are at more risk of death.

B. *Experiment 2: Mexican data: Need for ICU admission and Mortality prediction*

Tables VI and VII represent the confusion matrices obtained in the classification task of ICU and mortality prediction respectively in the case of Mexican data. The results obtained from different models for the need for ICU and mortality prediction are shown in tables X and XI, respectively. We get the best results from the Random Forest algorithm with 98% accuracy and 100% sensitivity for ICU

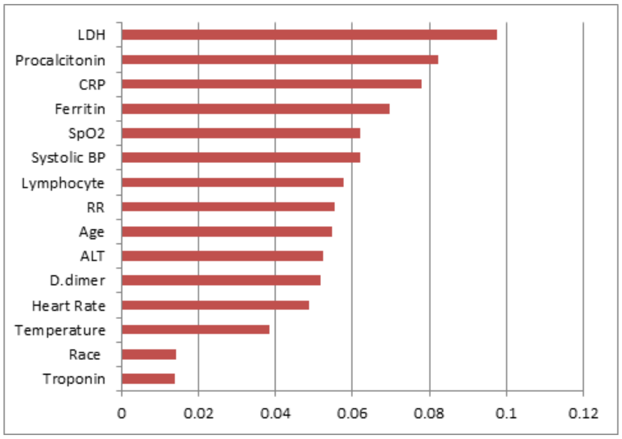


Figure 4. Significant features for ICU prediction (US data)

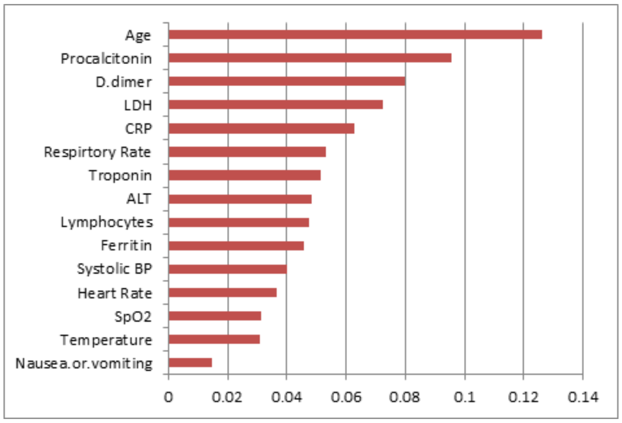


Figure 5. Significant features for Mortality prediction (US data)

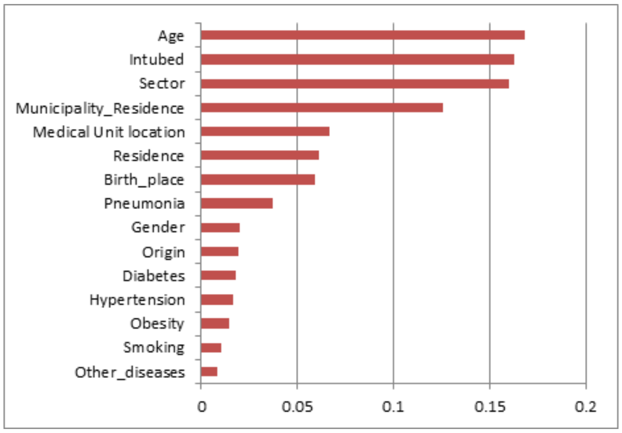


Figure 6. Significant features for ICU prediction (Mexican data)

prediction, and the LR model gives the best results for mortality prediction with 63% accuracy and 66% sensitivity. It can be seen from table XI that all the models show comparable results in mortality prediction, however, LR shows higher sensitivity and specificity than other models. It may be noted that RF shows the best performance except in the case of mortality prediction in Mexican data. This is due to the reason that the data preprocessing and undersampling done before classification has changed the dataset properties. The dataset has become more linearly separable and hence LR performs slightly better than the RF model. As per the RF algorithm, Age, Intubed, Sector, Municipality_residence, Medical unit location, and patient residence are the top five significant predictors among the provided input features for ICU admission. The top mortality predictors obtained from the RF algorithm include Age, Municipality_residence, Birth_place, Medical unit location, and residence of the patient. We have plotted various features against their Ginni values to depict their significance in the prediction process. The various significant predictors of the need for ICU admission and mortality prediction in the case of Mexican data are shown in Figures 6 and 7, respectively. It can be observed that the residence information of the patient is more significant than the comorbidities information for the prediction of ICU admission and mortality. This indicates that people from one place are at more risk than people from other places. Possible reasons may include more availability of medical assistance, sanitization, and awareness about the Covid-19 disease in some places than the other ones.

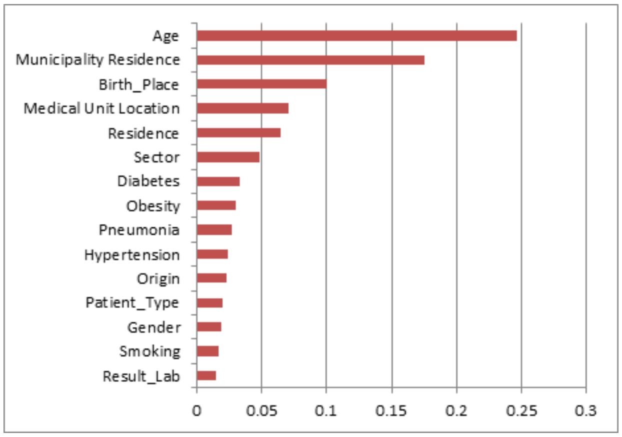


Figure 7. Significant features for Mortality prediction (Mexican data)

C. Experiment 3

In order to know whether the same features are responsible for mortality in the two datasets, we compared the significant features of the two datasets. To make a comparison, we employed identical feature types in both datasets that included patient demographic information and underlying comorbidities. Figure 8 and Figure 9 present the relationship between these features and their corresponding Ginni values in the case of the US and Mexican datasets respectively. The features with higher Ginni values are mainly responsible for prediction. Based on Figure 9, we can infer that age, ethnicity, gender, diabetes, and hypertension are the top five predictors of mortality in the US data. In the case of Mexican data, age, pneumonia, hypertension, diabetes, and gender are the top five significant mortality predictors

(Figure 9). This indicates that the types of patients more vulnerable to Covid-19 are similar with respect to the underlying comorbidities. However, the order of significance may vary; for example, in the US, diabetic people are at more risk of death due to Covid-19 than people with hypertension, but it is vice-versa in the case of Mexican data.

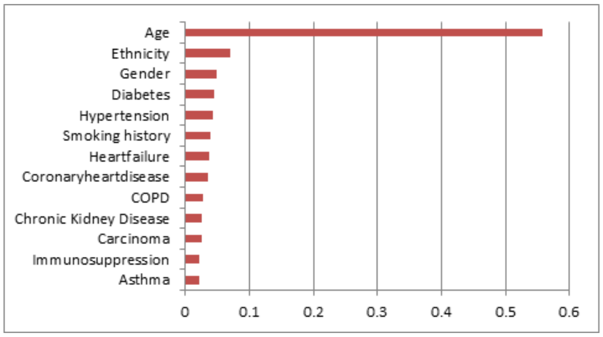


Figure 8. Significant features for mortality prediction using demographic and comorbidities features from US data

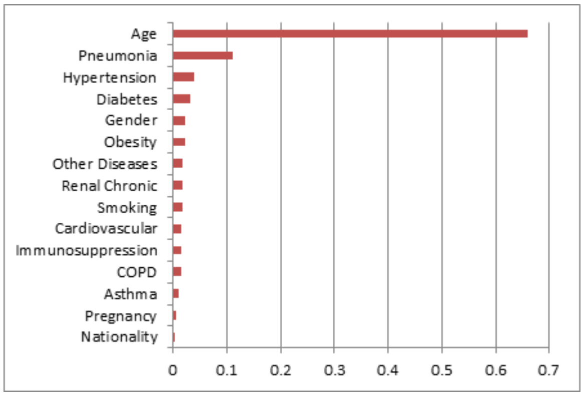


Figure 9. Significant features for mortality prediction using demographic and comorbidities features from Mexican data

7. CONCLUSION AND FUTURE WORK

The Covid-19 pandemic has had a catastrophic impact worldwide, with hospitals and healthcare systems struggling to cope with the surge in cases during the peak of the pandemic. Shortages of essential supplies and overwhelming patient numbers have been major challenges. Having a reliable prognostic model that can precisely forecast the risk of ICU admission necessity and death in Covid-19 patients is critical to avoid similar circumstances in the future and to facilitate appropriate triage. This study highlights the efficacy of machine learning models in achieving these objectives. The findings suggest that among all the algorithms evaluated, Random Forest outperformed the rest.

From the Mexican dataset, we can deduce that although the demographic and comorbidities information of the Covid-19 patients only gives good results for the need for ICU prediction, it does not provide good results for mortality prediction. Therefore, more data about Covid-19 patients, such as clinical notes and laboratory tests, is essential for mortality prediction. The US dataset shows promising results as we obtained 97% accuracy and 100% sensitivity for mortality prediction, which is higher than other previous studies. Comparing the demographic and comorbidities features of the two datasets, we find that the significant predictors of mortality are similar. The type of Covid-19 patients that are more vulnerable does not differ much in the US and Mexican data. We have found that patients that are older or have hypertension or diabetes are more vulnerable to Covid-19.

We build a model based on patients' demographics and comorbidities information only to predict the need for ICU admission and mortality risk. Other similar studies can compare the performance with our model based on Mexican data. Further studies can be done to collect the patients' data affected by different variants of the virus. More research can be done to find if the significant mortality predictors are the same in the case of different variants of the virus. Furthermore, we can extend the study to other nations and analyze the effect of the virus on their people. A more generalized model that is trained from data containing information on patients from different nations can be built, and significant predictors of mortality can be analyzed on a more general population. In addition to this, we can have image data along with the current datasets and develop a prediction model that can use combined data for the prediction task. It would be interesting to find if the mixed data gives better performance than individual types of data.

References

[1] "Coronavirus (covid-19) events as they happen,," https://www.who.int/emergencies/diseases/novel-coronavirus-2019/events-as-they-happen, [Online; Accessed: 2022-09-06].

[2] "Coronavirus disease (covid-19)," https://www.who.int/emergencies/diseases/novel-coronavirus-2019, [Online; Accessed: 2022-09-06].

[3] "Tracking sars-cov-2 variants," https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/, [Online; Accessed: 2022-09-06].

[4] N. Chen, M. Zhou, X. Dong, J. Qu, F. Gong, Y. Han, Y. Qiu, J. Wang, Y. Liu, Y. Wei *et al.*, "Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in wuhan, china: a descriptive study," *The lancet*, vol. 395, no. 10223, pp. 507–513, 2020.

[5] S. Richardson, J. S. Hirsch, M. Narasimhan, J. M. Crawford, T. McGinn, K. W. Davidson, D. P. Barnaby, L. B. Becker, J. D. Chelico, S. L. Cohen *et al.*, "Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with covid-19 in the new york city area," *Jama*, vol. 323, no. 20, pp. 2052–2059, 2020.

[6] O. L. Aiyegbusi, S. E. Hughes, G. Turner, S. C. Rivera, C. McMullan, J. S. Chandan, S. Haroon, G. Price, E. H. Davies, K. Nirantharakumar *et al.*, "Symptoms, complications and management of long covid: a review," *Journal of the Royal Society of Medicine*, vol. 114, no. 9, pp. 428–442, 2021.

[7] E. Elibol, “Otolaryngological symptoms in covid-19,” *European Archives of Oto-Rhino-Laryngology*, vol. 278, pp. 1233–1236, 2021.

[8] M. N. Esbin, O. N. Whitney, S. Chong, A. Maurer, X. Darzacq, and R. Tjian, “Overcoming the bottleneck to widespread testing: a rapid review of nucleic acid testing approaches for covid-19 detection,” *Rna*, vol. 26, no. 7, pp. 771–783, 2020.

[9] M. Nour, Z. Cömert, and K. Polat, “A novel medical diagnosis model for covid-19 infection detection based on deep features and bayesian optimization,” *Applied Soft Computing*, vol. 97, p. 106580, 2020.

[10] M. K. Pandit, S. A. Banday, R. Naaz, and M. A. Chishti, “Automatic detection of covid-19 from chest radiographs using deep learning,” *Radiography*, vol. 27, no. 2, pp. 483–489, 2021.

[11] S. Minaee, R. Kafieh, M. Sonka, S. Yazdani, and G. J. Soufi, “Deep-covid: Predicting covid-19 from chest x-ray images using deep transfer learning,” *Medical image analysis*, vol. 65, p. 101794, 2020.

[12] A. G. Dastider, M. R. Subah, F. Sadik, T. Mahmud, and S. A. Fattah, “Rescovnet: A deep learning-based architecture for covid-19 detection from chest ct scan images,” in *2020 IEEE REGION 10 CONFERENCE (TENCON)*. IEEE, 2020, pp. 57–60.

[13] M. Alazab, A. Awajan, A. Mesleh, A. Abraham, V. Jatana, and S. Alhyari, “Covid-19 prediction and detection using deep learning,” *International Journal of Computer Information Systems and Industrial Management Applications*, vol. 12, no. June, pp. 168–181, 2020.

[14] C.-L. Bong, C. Brasher, E. Chikumba, R. McDougall, J. Mellin-Olsen, and A. Enright, “The covid-19 pandemic: effects on low-and middle-income countries,” *Anesthesia and analgesia*, 2020.

[15] P. G. Walker, C. Whittaker, O. J. Watson, M. Baguelin, P. Winskill, A. Hamlet, B. A. Djafaara, Z. Cucunubá, D. Olivera Mesa, W. Green *et al.*, “The impact of covid-19 and strategies for mitigation and suppression in low-and middle-income countries,” *Science*, vol. 369, no. 6502, pp. 413–422, 2020.

[16] J. P. Figueroa, M. E. Bottazzi, P. Hotez, C. Batista, O. Ergonul, S. Gilbert, M. Gursel, M. Hassanain, J. H. Kim, B. Lall *et al.*, “Urgent needs of low-income and middle-income countries for covid-19 vaccines and therapeutics,” *The Lancet*, vol. 397, no. 10274, pp. 562–564, 2021.

[17] L. Kola, B. A. Kohrt, C. Hanlon, J. A. Naslund, S. Sikander, M. Balaji, C. Benjet, E. Y. L. Cheung, J. Eaton, P. Gonsalves *et al.*, “Covid-19 mental health impact and responses in low-income and middle-income countries: reimagining global mental health,” *The Lancet Psychiatry*, vol. 8, no. 6, pp. 535–550, 2021.

[18] Q. Wang, X. Wang, and H. Lin, “The role of triage in the prevention and control of covid-19,” *Infection Control & Hospital Epidemiology*, vol. 41, no. 7, pp. 772–776, 2020.

[19] B. Herreros, P. Gella, and D. R. De Asua, “Triage during the covid-19 epidemic in spain: better and worse ethical arguments,” *Journal of medical ethics*, vol. 46, no. 7, pp. 455–458, 2020.

[20] T. Dan, Y. Li, Z. Zhu, X. Chen, W. Quan, Y. Hu, G. Tao, L. Zhu, J. Zhu, Y. Jin *et al.*, “Machine learning to predict icu admission, icu mortality and survivors’ length of stay among covid-19 patients: toward optimal allocation of icu resources,” in *2020 IEEE international conference on bioinformatics and biomedicine (BIBM)*. IEEE, 2020, pp. 555–561.

[21] R. Han, Z. Liu, C. Philip Chen, L. Xu, and G. Peng, “Mortality prediction for covid-19 patients via broad learning system,” in *2020 7th International Conference on Information, Cybernetics, and Computational Social Systems (ICCSS)*. IEEE, 2020, pp. 837–842.

[22] Y. Chen, L. Ouyang, F. S. Bao, Q. Li, L. Han, B. Zhu, M. Xu, J. Liu, Y. Ge, and S. Chen, “An interpretable machine learning framework for accurate severe vs non-severe covid-19 clinical type classification,” *medRxiv*, pp. 2020–05, 2020.

[23] G. Lippi, J. Wong, B. M. Henry *et al.*, “Hypertension and its severity or mortality in coronavirus disease 2019 (covid-19): a pooled analysis,” *Pol Arch Intern Med*, vol. 130, no. 4, pp. 304–309, 2020.

[24] R. Pranata, M. A. Lim, I. Huang, S. B. Raharjo, and A. A. Lukito, “Hypertension is associated with increased mortality and severity of disease in covid-19 pneumonia: a systematic review, meta-analysis and meta-regression,” *Journal of the renin-angiotensin-aldosterone system: JRAAS*, vol. 21, no. 2, 2020.

[25] A. Mantovani, C. D. Byrne, M.-H. Zheng, and G. Targher, “Diabetes as a risk factor for greater covid-19 severity and in-hospital death: a meta-analysis of observational studies,” *Nutrition, Metabolism and Cardiovascular Diseases*, vol. 30, no. 8, pp. 1236–1248, 2020.

[26] A. Kumar, A. Arora, P. Sharma, S. A. Anikhindi, N. Bansal, V. Singla, S. Khare, and A. Srivastava, “Is diabetes mellitus associated with mortality and severity of covid-19? a meta-analysis,” *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 14, no. 4, pp. 535–545, 2020.

[27] R. Pranata, I. Huang, M. A. Lim, E. J. Wahjoepramono, and J. July, “Impact of cerebrovascular and cardiovascular diseases on mortality and severity of covid-19–systematic review, meta-analysis, and meta-regression,” *Journal of stroke and cerebrovascular diseases*, vol. 29, no. 8, p. 104949, 2020.

[28] W. Tian, W. Jiang, J. Yao, C. J. Nicholson, R. H. Li, H. H. Sigurslid, L. Wooster, J. I. Rotter, X. Guo, and R. Malhotra, “Predictors of mortality in hospitalized covid-19 patients: a systematic review and meta-analysis,” *Journal of medical virology*, vol. 92, no. 10, pp. 1875–1883, 2020.

[29] D. Wang, B. Hu, C. Hu, F. Zhu, X. Liu, J. Zhang, B. Wang, H. Xiang, Z. Cheng, Y. Xiong *et al.*, “Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in wuhan, china,” *jama*, vol. 323, no. 11, pp. 1061–1069, 2020.

[30] N. Darapaneni, A. Singh, A. Paduri, A. Ranjith, A. Kumar, D. Dixit, and S. Khan, “A machine learning approach to predicting covid-19 cases amongst suspected cases and their category of admission,” in *2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS)*. IEEE, 2020, pp. 375–380.

[31] X. Li, P. Ge, J. Zhu, H. Li, J. Graham, A. Singer, P. S. Richman, and T. Q. Duong, “Deep learning prediction of likelihood of icu admission and mortality in covid-19 patients using clinical variables,” *PeerJ*, vol. 8, p. e10337, 2020.

[32] M. Ezz, M. K. Elbashir, and H. Shabana, “Predicting the need for icu admission in covid-19 patients using xgboost,” *Computers, Materials and Continua*, pp. 2077–2092, 2021.

[33] E. El-Shafeiy, A. E. Hassanien, K. M. Sallam, and A. Abohany, "Approach for training quantum neural network to predict severity of covid-19 in patients," *Computers, Materials, & Continua*, pp. 1745–1755, 2021.

[34] J. Kang, T. Chen, H. Luo, Y. Luo, G. Du, and M. Jiming-Yang, "Machine learning predictive model for severe covid-19," *Infection, Genetics and Evolution*, vol. 90, p. 104737, 2021.

[35] A. Di Castelnuovo, M. Bonaccio, S. Costanzo, A. Gialluisi, A. Antinori, N. Berselli, L. Blandi, R. Bruno, R. Cauda, G. Guaraldi *et al.*, "Common cardiovascular risk factors and in-hospital mortality in 3,894 patients with covid-19: survival analysis and machine learning-based findings from the multicentre italian corist study," *Nutrition, Metabolism and Cardiovascular Diseases*, vol. 30, no. 11, pp. 1899–1913, 2020.

[36] C. An, H. Lim, D.-W. Kim, J. H. Chang, Y. J. Choi, and S. W. Kim, "Machine learning prediction for mortality of patients diagnosed with covid-19: a nationwide korean cohort study," *Scientific reports*, vol. 10, no. 1, p. 18716, 2020.

[37] M. E. Chowdhury, T. Rahman, A. Khandakar, S. Al-Madeed, S. M. Zughaier, S. A. Doi, H. Hassen, and M. T. Islam, "An early warning tool for predicting mortality risk of covid-19 patients using machine learning," *Cognitive Computation*, pp. 1–16, 2021.

[38] S. Li, Y. Lin, T. Zhu, M. Fan, S. Xu, W. Qiu, C. Chen, L. Li, Y. Wang, J. Yan *et al.*, "Development and external evaluation of predictions models for mortality of covid-19 patients using machine learning method," *Neural Computing and Applications*, pp. 1–10, 2021.

[39] E. Jamshidi, A. Asgary, N. Tavakoli, A. Zali, H. Esmaily, S. H. Jamaldini, A. Daaee, A. Babajani, M. A. S. Kashi, M. Jamshidi *et al.*, "Using machine learning to predict mortality for covid-19 patients on day zero in the icu," *medRxiv*, pp. 2021–02, 2021.

[40] A. S. Yadaw, Y.-c. Li, S. Bose, R. Iyengar, S. Bunyavanich, and G. Pandey, "Clinical features of covid-19 mortality: development and validation of a clinical prediction model," *The Lancet Digital Health*, vol. 2, no. 10, pp. e516–e525, 2020.

[41] L. Yan, H.-T. Zhang, J. Goncalves, Y. Xiao, M. Wang, Y. Guo, C. Sun, X. Tang, L. Jing, M. Zhang *et al.*, "An interpretable mortality prediction model for covid-19 patients," *Nature machine intelligence*, vol. 2, no. 5, pp. 283–288, 2020.

[42] "Open data general directorate of epidemiology — ministry of health — government," https://www.gob.mx/salud/documentos/datos-abiertos-152127, [Online; Accessed: 2021-08-06].

[43] K. Taunk, S. De, S. Verma, and A. Swetapadma, "A brief review of nearest neighbor algorithm for learning and classification," in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*. IEEE, 2019, pp. 1255–1260.

[44] K. M. Leung, "k-nearest neighbor algorithm for classification," *Polytechnic University Department of Computer Science/Finance and Risk Engineering*, 2007.

[45] "K-nearest neighbors (knn) — explained — by soner yıldırım — towards data science," https://towardsdatascience.com/k-nearest-neighbors-knn-explained-cbc31849a7e3, [Online; Accessed: 2021-08-11].

[46] F. Fitriyadi and M. Muqorobin, "Prediction system for the spread of corona virus in central java with k-nearest neighbor (knn) method," *International Journal of Computer and Information System (IJCIS)*, vol. 2, no. 3, pp. 80–85, 2021.

[47] R. A. Jaleel, I. M. Burhan, and A. M. Jalookh, "A proposed model for prediction of covid-19 depend on k-nearest neighbors classifier: iraq case study," in *2021 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*. IEEE, 2021, pp. 1–6.

[48] X. Zou, Y. Hu, Z. Tian, and K. Shen, "Logistic regression model optimization and case analysis," in *2019 IEEE 7th international conference on computer science and network technology (ICCSNT)*. IEEE, 2019, pp. 135–139.

[49] J. Feng, H. Xu, S. Mannor, and S. Yan, "Robust logistic regression and classification," *Advances in neural information processing systems*, vol. 27, 2014.

[50] "Logistic regression classifier. how it works (part-1) — by caglar subasi — towards data science," https://towardsdatascience.com/logistic-regression-classifier-8583e0c3cf9, [Online; Accessed: 2021-08-11].

[51] D. M. Abdullah and A. M. Abdulazeez, "Machine learning applications based on svm classification a review," *Qubahan Academic Journal*, vol. 1, no. 2, pp. 81–90, 2021.

[52] S. Yue, P. Li, and P. Hao, "Svm classification: Its contents and challenges," *Applied Mathematics-A Journal of Chinese Universities*, vol. 18, pp. 332–342, 2003.

[53] A. Patle and D. S. Chouhan, "Svm kernel functions for classification," in *2013 International Conference on Advances in Technology and Engineering (ICATE)*. IEEE, 2013, pp. 1–9.

[54] "Support vector machine — introduction to machine learning algorithms — by rohith gandhi — towards data science," https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47, [Online; Accessed: 2021-08-11].

[55] B. Charbuty and A. Abdulazeez, "Classification based on decision tree algorithm for machine learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 20–28, 2021.

[56] "Decision trees in machine learning — by prashant gupta — towards data science," https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052, [Online; Accessed: 2021-08-11].

[57] C. Jin, L. De-Lin, and M. Fen-Xiang, "An improved id3 decision tree algorithm," in *2009 4th international conference on computer science & Education*. IEEE, 2009, pp. 127–130.

[58] H. H. Patel and P. Prajapati, "Study and analysis of decision tree based classification algorithms," *International Journal of Computer Sciences and Engineering*, vol. 6, no. 10, pp. 74–78, 2018.

[59] Y. Liu, Y. Wang, and J. Zhang, "New machine learning algorithm: Random forest," in *Information Computing and Applications: Third International Conference, ICICA 2012, Chengde, China, September 14-16, 2012. Proceedings 3*. Springer, 2012, pp. 246–252.

[60] "Understanding random forest. how the algorithm works and why it is… — by tony yiu — towards data science," https://towardsdatascience.com/understanding-random-forest-58381e0602d2, [Online; Accessed: 2021-08-11].

[61] M. Schonlau and R. Y. Zou, “The random forest algorithm for statistical learning,” *The Stata Journal*, vol. 20, no. 1, pp. 3–29, 2020.

[62] C. M. Yeşilkanat, “Spatio-temporal estimation of the daily cases of covid-19 in worldwide using random forest machine learning algorithm,” *Chaos, Solitons & Fractals*, vol. 140, p. 110210, 2020.

[63] S. Nembrini, I. R. König, and M. N. Wright, “The revival of the gini importance?” *Bioinformatics*, vol. 34, no. 21, pp. 3711–3718, 2018.

[64] “Random forest for feature importance — by z_ai — towards data science,” https://towardsdatascience.com/random-forest-for-feature-importance-ea90852b8fc5, [Online; Accessed: 2021-08-11].

[65] A. Tharwat, “Classification assessment methods,” *Applied computing and informatics*, vol. 17, no. 1, pp. 168–192, 2021.

[66] D. Chicco and G. Jurman, “The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation,” *BMC genomics*, vol. 21, pp. 1–13, 2020.

[67] “Random oversampler,” https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.RandomOverSampler.html, [Online; Accessed: 2021-08-11].

[68] “Nearmiss undersampler,” https://imbalanced-learn.org/dev/references/generated/imblearn.under_sampling.NearMiss.html, [Online; Accessed: 2021-08-11].

Uzmat Ul Nisa received her B.Tech in Information Technology from National Institute of Technology, Srinagar. She received her M.Tech in Information Technology from the Central University of Kashmir. Her research interests include machine learning, deep learning, and their applications in various fields.

Mohammad Ahsan Chishti has done his Doctor of Philosophy (Ph.D.) from National Institute of Technology Srinagar. He has completed a Bachelor of Engineering (B.E.) and M.S. in Computer and Information Engineering (MSCIE) from International Islamic University Malaysia. Presently he is working as an Associate Professor and Head of the Department of Computer Science & Engineering, National Institute of Technology Srinagar. He has more than 150 research publications to his credit and 14 patents with two granted International Patents. He has successfully completed a number of sponsored research projects. He has been awarded “IEI Young Engineers Award 2015-2016” in the field of Computer Engineering for the year 2015-16 by the Institution of Engineers (India) and “Young Scientist Award 2009-2010” from the Department of Science & Technology, Government of Jammu and Kashmir for the year 2009-2010. He has guided 8 research scholars for the award of Ph.D. in Engineering and also he is presently guiding a number of Ph.D., Master of Technology (M.Tech), and Bachelor of Technology (B. Tech) research projects and his research area includes Artificial Intelligence, Machine Learning, Internet of Things, and Next Generation Networks. He is a Senior Member-Institute of Electrical and Electronics Engineers (IEEE) and a Member of several other societies like IEI, CSI, and IETE apart from being a member of other technical societies.