

Forecasting the Spread of COVID-19 and ICU Requirements

Prajoy Podder¹, Aditya Khamparia², M. Rubaiyat Hossain Mondal¹(✉),

Mohammad Atikur Rahman³, Subrato Bharati¹

¹Institute of ICT, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh

²Lovely Professional University, Punjab, India

³Ranada Prasad Shaha University, Narayanganj-1400, Bangladesh
rubaiyat97@iict.buet.ac.bd

Abstract: Since December 2019, the world is fighting against coronavirus disease (COVID-19). This disease is caused by a novel coronavirus termed as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). This work focuses on the applications of machine learning algorithms in the context of COVID-19. Firstly, regression analysis is performed to model the number of confirmed cases and death cases. Our experiments show that autoregressive integrated moving average (ARIMA) can reliably model the increase in the number of confirmed cases and can predict future cases. Secondly, a number of classifiers are used to predict whether a COVID-19 patient needs to be admitted to an intensive care unit (ICU) or semi-ICU. For this, classification algorithms are applied to a dataset having 5644 samples. Using this dataset, the most significant attributes are selected using features selection by ExtraTrees classifier, and Proteina C reativa (mg/dL) is found to be the highest-ranked feature. In our experiments, random forest, logistic regression, support vector machine, XGBoost, stacking and voting classifiers are applied to the top 10 selected attributes of the dataset. Results show that random forest and hard voting classifiers achieve the highest classification accuracy values near 98%, and the highest recall value of 98% in predicting the need for admission into ICU/semi ICU units.

Keywords: COVID-19, ICU, feature selection, classification, ARIMA model

1 Introduction

In December 2019, novel coronavirus diseases (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) started in Wuhan, China. As of 29 November 2020, there have been 62,789,393 confirmed cases and 1,462,086 deaths in more than 218 countries and territories. This particular coronavirus is related to Middle East respiratory syndrome coronavirus (MERS-CoV) and severe acute respiratory syndrome coronavirus (SARS-CoV). This virus has a similarity with bat-derived coronaviruses, and it is likely that the virus has spread to human beings through an unidentified intermediate carrier [1]. There are four subtypes of coronaviruses namely α , β , γ and δ coronaviruses, some of which have been affecting humans or animals including birds, mice, pigs, dogs and cats [2-5]. SARS-CoV-2 is a form of β coronavirus, and it affects people of different severity. For instance, many infected people develop mild or moderate symptoms, while some do not have any noticeable symptoms at all. On the other hand, some infected people have serious symptoms, including difficulty in breathing, respiratory failure needing hospitalisation. Based on a report published on 24 January 2020, COVID-19 patients have some standard features, i.e. fatigue, cough, and fever, while dyspnea and diarrhoea are found to be unidentified features [6-8].

SARS-CoV-2 has a positive-sense and single-stranded Ribonucleic acid (RNA) genome [9]. Six CoVs are recognised as human-susceptible viruses out of which MERS-CoV and SARS-CoV cause severe respiratory tract infections [10]. SARS-CoV-2 has 96.2% similarity to a bat CoV RaTG13 in terms of the genome sequence. Moreover, SARS-CoV-2 shares 79.5% characteristics to SARS-CoV. In order to spread into human cells, SARS-CoV-2 uses angiotensin-converting enzyme 2 (ACE2) as a cell receptor [11]. Like other viruses, this novel coronavirus has mutated in the last six months or so since December 2019. The mutation is a normal part of a virus's life cycle, but the mutation can play an important part in the spread of the virus and in finding ways to overcome the virus infection [12-13]. SARS-CoV-2 has a mutation rate faster than MERS and SARS-CoV [14], but slower than the mutation of H7N9 avian influenza [15]. The whole world and almost every country of the world are suffering a lot due to this novel coronavirus. The number of death cases is also increasing in many countries. Hence, an effective public health action plan is vital to combat COVID-19. Appropriate forecasting of the spread of COVID-19 can assist in managing the effects of the disease.

The main theme of this work is to forecast confirmed and death cases using autoregressive (AR), moving average (MA) and autoregressive integrated moving average (ARIMA) models. ARIMA is a combination of AR and MA models. Firstly, we analyse statistical time series data that is continuously changing day by day with the spread of COVID-19. ARIMA is one of the effective and easiest models for time series data forecasting. For that reason, we use this model for future prediction and forecasting. Although the introduction of vaccines can reduce the spread of this infected disease, proper treatment will need for serious patients. Therefore, intensive care unit (ICU) and semi-ICU are required for managing serious patients. Accordingly, we can apply machine learning methods on clinical data of COVID-19 patients. Thus, we can predict the future spread and can plan for managing COVID-19 situations. A short 10-day forecasting can help the authorities to plan for quick control measures. In other words, the effects of COVID-19 can be managed if predictions can be made on the future spread of the disease and the possible requirement of ICU and semi ICU units.

The major outcomes of the paper can be summarized as follows:

- (1) Autoregressive integrated moving average (ARIMA) model is proposed and implemented for future forecasting of COVID-19 cases.
- (2) Machine learning classification is performed in order to predict the normal patients and patients that need admission to intensive care unit (ICU) or semi-ICU among the samples available in the dataset.
- (3) The top 30 correlation values between features and target variable are illustrated, and Extra trees classifier is used as a feature selection technique.
- (4) The performance of different classifiers, including stacking classifiers are compared in predicting whether patients need to ICU/semi-ICU or not.

This paper is divided into five sections and is organised as follows. Some related works dealing with regression and machine learning are described in Section 2. Future forecasting of COVID-19 cases using the ARIMA model is described in Section 3. Section 4 contains a summary of the analysed dataset and machine learning approaches using the dataset. The performance of the proposed classifiers is illustrated in Section 5. Finally, Section 6 provides a discussion on the goals achieved and the points of further research.

2 Related works

Many researchers are working on COVID-19 using machine learning. Sarker et al. [18] applied a Random Forest (RF) approach to identify the causes of mortality due to COVID-19. However, their paper did not describe any precision, recall, accuracy results, etc. They worked with a statistical dataset, and their dataset includes the clinical data of 1085 cases from 13 January to 28 February 2020. Benvenuto et al. [19] predicted the spread of the virus infection using a new ARIMA model. The work in [19] also forecasted the various parameters for the next two days. The authors in [19] used the real-time statistical dataset of John Hopkins University. Tandon et al. [23] considered the registered cases of India by deploying the ARIMA model. Ardabili et al. [20] proposed multi-layered perceptron neural network (MLPN) and ANFIS in order to predict the COVID-19 outbreak of some top affected countries such as Italy, China, Iran, USA, and Germany. A number of ARIMA models with different boundaries were considered for the datasets of Italy, Spain and France [28]. Several time series analyses including ARIMA, RF, RIDGE and SVR were reported to forecast infection cases for 6 days [29]. A study predicted confirmed cases for the next 10 days for the case of Canada, South Korea, France and the UK [30]. Yang et al. [25] considered 710 COVID-19 patients and reported the clinical outcomes of 52 one who were critically ill. Some machine learning results using some classification algorithms are provided on the diagnosis of COVID-19 patients using a subset of a COVID-19 dataset in [26]. Randhawa et al. studied the genome structure of the virus by using supervised machine learning algorithms and by applying digital signal processing [27]. They used LDA, SVM, KNN, Subspace Discriminant as classification models.

3 Regression analysis

The number of confirmed COVID-19 infected cases and death cases have reached up to 30,055,710 and 9,43,433, respectively, according to the WHO situation reports (18 September 2020) [24]. ACE 2 receptor is the entry receptor in humans of the novel coronavirus [21]. As of January 2020, some vaccines are being introduced for preventing COVID-19. A number of vaccines including Moderna's vaccine and Pfizer-BioNTech vaccines are approved by

national regulatory authorities for public use. In different countries, doctors prescribe different medicines for patients based on the symptoms and conditions of the patients. Currently, it has become a global pandemic. This disease has become a serious burden on the health care facilities of different countries. It is considered as a serious health crisis and an enemy to humanity. In this situation, preventing the spread of this infection is important. Even slowing down the spread helps healthcare systems to be ready for COVID 19 patients, as there are not enough hospital beds to support a huge number of simultaneous patients. Considering these circumstances, constructing appropriate models to predict the spread of the virus is useful. These models should not only be computationally competent but should also be realistic. Therefore, they can help policymakers, medical personals and also the general public. In particular, it can help the medical system is being prepared for the confirmed cases in the coming days and plan accordingly. As a result, the ARIMA model is employed for predicting the confirmed and death cases of novel coronavirus disease. It is shown in [22] that compared to linear regression (LR) and support vector machine (SVM), the ARIMA model can predict events more reliably [22].

The dataset used for regression analysis is collected from Kaggle repository [16]. Johns Hopkins University has prepared this dataset of the number of COVID 19 cases [17]. This dataset contains time series data on the number of confirmed cases, number of death cases and number of recovered cases. For regression analysis, ARIMA is considered, which is a fusion of auto-regressive (AR) and moving average (MA) models. It is already applied in the prediction problems of environment monitoring [30, 31], financial economy [32, 33], food safety [34], traffic system [35], etc. Seasonal ARIMA model is used to predict the daily and monthly solar radiation [36]. ARIMA model works well for predicting short time series data, and the dataset considered in this work is a short time series. Hence, ARIMA model is considered in this work. A non-seasonal ARIMA model can be classified as an "ARIMA(x,y,z)" model, where

x = the number of autoregressive terms,

y = the number of non-seasonal differences needed for stationarity, and

z = the number of lagged forecast errors in the prediction equation.

Let P is denoted as the y -th difference of P , which means:

If $y=0$: $p_t = P_t$

If $y=1$: $p_t = P_t - P_{t-1}$

If $y=2$: $p_t = (P_t - P_{t-1}) + (P_{t-1} - P_{t-2})$

The general forecasting equation is:

$$\hat{p}_t = \mu + \phi_1 p_{t-1} + \dots + \phi_x p_{t-x} - \dots - \phi_z e_{t-z} \quad (1)$$

Fig. 1 presents the number of confirmed cases in the world from 22 January 2020 to 23 September 2020 using the ARIMA model. In this case, one plot is for the training data of the actual cases, another plot is for the validation set of the actual confirmed cases, and the third one is the confirmed cases modelled by ARIMA. It can be seen from Fig. 1 that the ARIMA model fits with the actual confirmed cases except for the days since early September 2020. In the last few days, the ARIMA model has a decreasing trend as opposed to the increasing trend of the validation set. Table 1 shows the forecast of confirmed cases using the ARIMA model and for comparative analysis linear regression model is also presented. Moreover, Table 1 presents the actual number of confirmed cases and the average error with respect to ARIMA model. It can be seen from Table 1 that the projected confirmed cases using the ARIMA model are higher than linear regression, and are closer to the actual numbers. Table 1 shows that the number of predicted confirmed cases on 14 September 2020 is 2,91,71,386 using ARIMA model, where the real number of confirmed cases is 3,01,54,445.

In order to compare the ARIMA and linear regression models, root mean square error (RMSE) was calculated. The calculated value of RMSE for ARIMA is 57235, which is lower than that of linear regression being 7942350. Hence, the ARIMA model is more reliable than linear regression in predicting future confirmed cases.



Fig. 1. Forecasting of confirmed cases using the ARIMA model

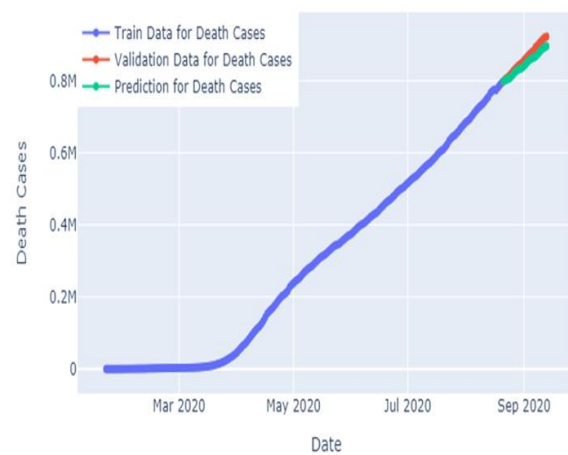


Fig. 2. Forecasting of death cases using ARIMA model

Table 1. Comparison between ARIMA and linear regression for confirmed case

Date	Linear Regression Prediction	ARIMA Model Prediction	Actual Confirmed Case	Relative error with ARIMA
14/09/20	19313002	29171386	30154445	3.3%
15/09/20	19413940	29439866	30439705	3.3%
16/09/20	19514877	29726395	30753403	3.3%
17/09/20	19615815	30023520	31071813	3.4%
18/09/20	19716753	30317078	31402132	3.5%
19/09/20	19817691	30596777	31706954	3.5%
20/09/20	19918629	30863636	31969394	3.5%
21/09/20	20019567	31128920	32210989	3.4%
22/09/20	20120505	31405581	32492282	3.3%
23/09/20	20221443	31698711	32813778	3.4%

Fig. 2 presents the number of death cases in the world from 22 January 2020 to 23 September 2020 using the ARIMA model. Similar to Fig. 1, Fig. 2 has one plot for the training data of the actual cases, another plot for the validation set of the actual confirmed cases, and a third one for the confirmed cases modelled by ARIMA. It can be seen from Fig. 2 that the ARIMA model successfully fits with the actual death cases. Table 2 presents the predicted number of deaths with the ARIMA model and the actual number of death cases. Table 2 illustrates that on 14 September 2020, the number of predicted death cases is 8,99,261 using ARIMA Model of order (2,2,3), where the real number of death cases is 9,44,300. So, from Table 1 and Table 2, the predicted confirmed and death cases are quite close to the real confirmed and death cases for ARIMA model. In Table 1, the average error between the real confirmed cases and the cases modelled by ARIMA is 3.395%, whereas the average error between the real death cases and the cases modelled by ARIMA is 4.83% as shown in Table 2.

Table 2. Future ten days' death case using ARIMA model

Date	ARIMA Model Death Forecast	Actual Death Case	Relative Error
14/09/20	899261	944300	4.8%
15/09/20	904047	950300	4.9%
16/09/20	910123	956312	4.8%
17/09/20	916391	961725	4.7%
18/09/20	921650	967177	4.7%

Date	ARIMA Model Death Forecast	Actual Death Case	Relative Error
19/09/20	925494	972384	4.8%
20/09/20	928593	976379	4.9%
21/09/20	932170	980181	4.9%
22/09/20	937080	985553	4.9%
23/09/20	943179	991591	4.9%

Table 3(a). Comparison of different regression models in predicting confirmed cases in India

Date	PR Model	AR Model	MA Model	ARIMA Model	Actual Case	Relative Error with AR Model
14/09/20	5615478	4718586	4674258	4686230	4926914	4.2%
15/09/20	5752073	4795516	4749208	4762345	5018034	4.4%
16/09/20	5891329	4872784	4824492	4838793	5115893	4.8%
17/09/20	6033285	4950393	4900110	4915572	5212686	5.0%
18/09/20	6177979	5028341	4976062	4992682	5305475	5.2%
19/09/20	6325451	5106629	505234	5070125	5398230	5.4%
20/09/20	6475741	5185256	5128968	5147898	5485612	5.5%
21/09/20	6628890	5264223	5205923	5226003	5560105	5.3%
22/09/20	6784936	5343530	5283211	5304439	5640496	5.3%
23/09/20	6943923	5423176	5360834	5383207	5730184	5.4%

Table 3(b). Different regression models in predicting confirmed cases in the USA

Date	PR Model	AR Model	MA Model	ARIMA Model	Actual Case	Relative Error with MA Model
14/09/20	8327614	6690583	6734119	6639293	6859862	1.8%
15/09/20	8445839	6738009	6783289	6680210	6897047	1.6%
16/09/20	8565635	6785630	6832669	6723885	6938003	1.5%
17/09/20	8687018	6833455	6882258	6771436	6985072	1.5%
18/09/20	8810006	6881487	6932057	6821323	7037177	1.5%
19/09/20	8934613	6929727	6982065	6870524	7081424	1.4%
20/09/20	9060856	6978170	7032283	6916736	7115267	1.2%
21/09/20	9188752	7026815	7082711	6960009	7152205	0.9%
22/09/20	9318316	7075661	7133348	7002646	7188463	0.8%
23/09/20	9449565	7124709	7184195	7047496	7230704	0.6%

Table 3 presents the comparative analysis of different regression models such as polynomial, AR, MA, ARIMA, etc., to predict the confirmed cases of next ten days (14 September 2020-23 September 2020) for India and the USA. For this case, the input data has been collected up to 13 September 2020. Table 3(a) is for the case of India, whereas Table 3(b) is for the case of the USA. The comparison between the real values and the predicted values are visible in Table 3(a) and Table 3(b). Table 3(a) shows that the predicted confirmed cases using AR model are comparatively closer to the real number of confirmed cases. Table 3(a) presents that the number of confirmed cases on 14 September 2020 is 49,26,914 in India, whereas the predicted confirmed cases using AR Model is approximately 47,18,586. Moreover, Table 3(b) shows that on 14 September 2020, the predicted confirm cases in USA using MA model is 67,34,119, whereas the actual confirmed case is 68,59,862. In Table 3(a), the average error between the real confirmed cases and the cases modelled by AR for India is 5.05%, whereas the average error between the real confirmed cases and the cases modelled by MA for the USA is 1.28% as shown in Table 3(b). It can be seen that the error rate in the USA is smaller than that of India. This can be due to the fact that preventive measure and proper maintenance of health guidelines and lockdown were better maintained in the USA.

In this paper, experiments are performed to classify patients who need admission to ICU or semi ICU and patients who do not need such admission. The classification is done using the samples available in the dataset publicly available Kaggle repository [37]. 5644 samples are contained in this dataset with 111 attributes provided by Hospital Israelita Albert Einstein, Brazil. Anonymised data samples collected by RT-PCR and additional laboratory tests are included during a visit to the hospital. The classification tasks in this paper are performed using sci-kit-learn machine learning library of Python programming language. The overall research work is performed using the processes shown in Fig 4. The main stages of the processing are the data pre-processing, the feature selection, and the classification using machine learning algorithms.

3.1 Data pre-processing

Data pre-processing refers to a data mining method which transfers data into a reasonable format. Several columns are filtered in this stage. Columns are filtered with the percentage of null is greater than or equal to 99%. Several unimportant columns such as 'Adenovirus', 'Bordetella pertussis', 'Metapneumovirus', 'Chlamydomphila pneumoniae', 'Inf A H1N1 2009', 'Urine - Urobilinogen', 'Urine - Crystals', 'Urine – Aspect' are also deleted for balancing the imbalanced dataset. After filtering, only 67 columns are taken into consideration.

The dataset considered in this work is imbalanced. There are a number of missing data. Median imputation process is used here for replacing the missing data. In statistical analysis, imputation is a way to avoid missing values in the dataset. Otherwise, missing values can generate a problem for evaluating data.

3.2 Correlation between features and the target variable

After executing several types of data processing, the correlation between the target variable ward_semi_intensive and the features are studied. The variable ward_semi_intensive contains the data related to the requirement of a general ward or a ICU/semi ICT. It is the summation of the patients admitted to a regular ward, semi-ICU and ICU. The top 30 correlation values between features and target value are illustrated in Table 4. Next, a correlation matrix is plotted among the features. Since the correlation matrix containing as high as 30 features will be difficult to view, only a portion will be displayed. Fig. 3 shows a correlation matrix that illustrates the correlation coefficient between different features. For clarity, only the correlation of 9 features is shown that have correlation values of 0.75 or higher. The matrix summarises the data as an input for advanced analyses. The line of 1s going from the top left to the bottom right is the main diagonal, which shows that each variable always perfectly correlates with itself. A correlation matrix is ideal for comparing the measurement for each pair of dimension values. This matrix is symmetrical, with the same correlation is shown above the main diagonal being a mirror image of those below the main diagonal.

Features	Urine_Bile_pigments_absent	Urine_Ketone_Bodies_absent	Urine_Hyaline_cylinders_absent	Urine_Yeast_absent	Urine_Protein_absent	Urine_Esterase_absent	Urine_Granular_cylinders_absent	Urine_Hemoglobin_present	Urine_color_yellow
Urine_Bile_pigments_absent	1	0.90	0.97	0.99	0.92	0.92	0.99	0.88	0.99
Urine_Ketone_Bodies_absent	0.90	1	0.88	0.89	0.82	0.83	0.88	0.75	0.89
Urine_Hyaline_cylinders_absent	0.97	0.88	1	0.98	0.92	0.91	0.99	0.85	0.96
Urine_Yeast_absent	0.99	0.89	0.98	1	0.92	0.92	0.99	0.87	0.99
Urine_Protein_absent	0.92	0.82	0.92	0.92	1	0.88	0.92	0.86	0.91
Urine_Esterase_absent	0.92	0.83	0.91	0.92	0.88	1	0.91	0.82	0.91
Urine_Granular_cylinders_absent	0.99	0.88	0.99	0.99	0.92	0.91	1	0.88	0.98
Urine_Hemoglobin_present	0.88	0.75	0.85	0.87	0.86	0.82	0.88	1	0.86
Urine_color_yellow	0.99	0.88	0.96	0.99	0.91	0.91	0.98	0.86	1

Fig. 3. Correlation matrix of features

Several classification algorithms such as RF, LR, XGB, SVM are applied to the features in order to classify the data samples. Furthermore, stacking classifiers are also considered in this work. In this case, stacking 1 means the stack of RF, XGB and LR, while stacking 2 means the stack of RF, SVM and LR.

Table 4: Correlation between features and the target variable

Features	Value	Features	Value
Proteina C reactiva mg/dL	0.285	Serum Glucose	0.149
Rods	0.282	Aspartate transaminase	0.148
Metamyelocytes	0.227	Red blood cell distribution width (RDW)	0.137
Respiratory Syncytial Virus	0.189	Urine_Protein_absent	0.125
Urine_color_yellow	0.180	SARS-Cov-2 exam result	0.122
Urine_Bile_pigments_absent	0.178	Hb saturation	0.114
Urine_Yeast_absent	0.177	pO ₂ (venous blood gas analysis)	0.114
Urine_Ketone_Bodies_absent	0.171	Alkaline phosphatase	0.111
Urine_Granular_cylinders_absent	0.169	International normalized ratio (INR)	0.102
Urine_Hyaline_cylinders_absent	0.163	Segmented	0.097
Lactic Dehydrogenase	0.157	Gamma-glutamyl transferase	0.091
Urine_Esterase_absent	0.155	Rhinovirus/Enterovirus	0.085
Myelocytes	0.153	Promyelocytes	0.076
Neutrophils	0.149	Urine - Leukocytes	0.068

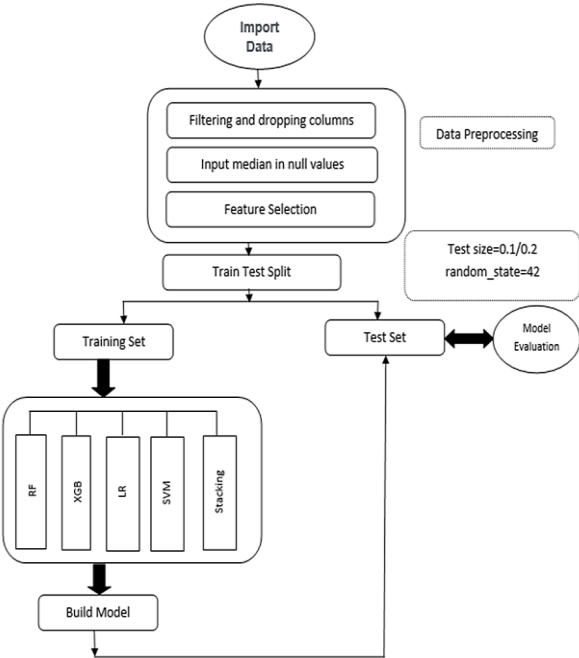


Fig. 4. Workflow diagram

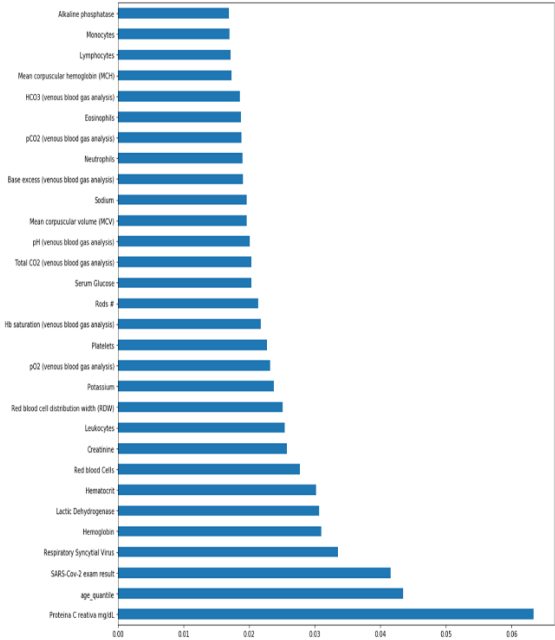


Fig. 5. Top 30 feature selection using ExtraTrees classifier

3.3 Feature Importance using ExtraTrees Classifier

In order to find the features that impact the target variable, feature importance is used in this work. With feature importance, we obtain a score against each of the features. In this context, the importance of a feature or attribute becomes greater for higher values of feature scores. Fig. 5 shows the top features selected using ExtraTrees classifier. It can be seen that Proteina C reactiva (mg/dL) is the highest-ranked feature.

3.4 Classification algorithms

Machine learning classifiers are successfully applied to classify normal (negative) cases and positive (having disease) cases for the case of many diseases [37-50]. Several classifiers such as RF [44], LR [43], SVM, XGBoost (XGB) are implemented in this work. These classifiers are often used in cancer disease prediction such as breast cancer, lung cancer, etc. [39, 42], prediction of spinal abnormalities [37] and hepatitis disease prediction [38]. Therefore, these algorithms are applied to the dataset. Next, two stacking classifiers such as stacking 1 (RF, XGB and LR) and stacking 2 (RF, SVM and LR) are proposed. Note that the stacking classifier is a type of hybrid classifier. As a result, the performance of the base classifier and the hybrid classifier are observed together. Moreover, two majority voting ensembles hard voting and soft voting are considered. For both hard and soft voting, the classification is done by combining the individual algorithms: RF, LR and SVM.

4 Result Analysis

This section discusses the effectiveness of different machine learning classifiers in classifying suspected patients who need ICU/semi-ICU admission, and who need not.

4.1 Performance Metrics

Similar to other biomedical data analytics, analysing COVID 19 data samples requires a number of metrics. Total classification accuracy alone is not good enough to find the effectiveness of a classifier. In the context of COVID-19

patients' data, accurate diagnosis, as well as no incorrect prediction, is vital. Hence, a number of metrics are considered for proper evaluation of the classifiers. These are training and testing accuracy, precision, recall, and F1-score. Besides, the receiver operating characteristic (ROC) curve, the area under the ROC curve (AUC), and precision-recall curves are considered too. The ROC curve is effective as it indicates the trade-off between the true positive rate (TPR) and the false positive rate (FPR). A ROC curve close to (0, 1) coordinate or near the upper left corner indicates the classifier is reliable. Precision-recall curves can be effective to evaluate the classifiers for the imbalance dataset used in this paper.

4.2 Performance Evaluation

In this section, a number of classifiers are applied on the top 10 features of the dataset to classify the suspected patients as needing ICU/semi-ICU or no ICU admission. We evaluated the effectiveness of several classifiers including RF, LR, SVM, XGB, stacking 1, stacking 2, AdaBoost, Bagging LR, hard voting and soft voting algorithms in classifying the requirement of ICU/semi-ICU or no ICU requirement.

We have used a number of evaluation metrics, including recall, precision, F1-score, AUC, specificity, etc. The most important evaluation metrics are recall and precision, while precision relates to depict the relevant results which is in the form of percentage. Moreover, recall relates to the total relevant results of correctly classified algorithm. The combination of recall and precision is F-score. It is well-defined as the harmonic mean of the model's precision and recall. The performance of models with its summary is depicted by area under the curve (AUC) when the dataset is imbalanced. Conversely, the probability is reflected by the specificity of the test.

Table 5. Accuracy and other performance metrics of the classifiers for 10% testing data

Classifiers	Weighted Average Precision	Weighted Average Recall	Miss rate/ FNR	Specificity	Weighted Average F1 score	Accuracy	AUC
RF	97%	98%	2%	97.85%	97%	97.69%	91%
LR	96%	90%	10%	98.62%	93%	90.09%	58%
SVM	97%	97%	3%	97.50%	97%	97.35%	86%
XGB	97%	98%	2%	97.68%	97%	97.52%	89%
Stacking 1	96%	97%	3%	97.67%	96%	96.99%	74%
Stacking 2	97%	97%	3%	97.85%	97%	97.35%	80%
AdaBoost	97%	97%	3%	99.27%	97%	97.35%	79%
Bagging LR	96%	97%	3%	99.82%	96%	96.99%	74%
Voting(RF, LR, SVM) Hard	97%	98%	2%	99.82%	98%	97.89%	91%
Voting (RF, LR, SVM) Soft	97%	97%	3%	99.45%	97%	97.35%	80%

Table 5 describes the results of RF, LR, SVM, XGB, stacking 1, stacking 2, voting hard and voting soft in terms of precision, recall, FNR, specificity, AUC, F1 score and accuracy for the case of 10% testing and 90% training data. Based on this training testing split, we compare the classifier results for the prediction of the ICU admitted COVID-19 patients. The classification accuracy of RF, LR, SVM, XGB, stacking 1, stacking 2, voting hard and voting soft algorithms are 97%, 90%, 97%, 97%, 96%, 97%, 97%, and 97%, respectively. Among the classifiers, hard voting has the highest classification accuracy, followed by RF algorithm. It can be seen from Table 5 that voting hard and RF algorithms provide the highest recall, precision and AUC values. The 98% recall value of RF and hard voting indicate that 98% of the COVID-19 patients requiring ICU admission are correctly classified with only a 2% of miss rate. Hence, it can be seen that RF and voting hard algorithms perform well when the dataset is divided into 10% testing and 90% training samples.

Table 6 describes the results of RF, LR, SVM, XGB, stacking 1, stacking 2, hard voting and soft voting in terms of precision, recall, FNR, specificity, AUC, F1 score and accuracy for the case of 20% testing and 80% training data. It can be seen from Table 6 that AdaBoost provides the highest accuracy and recall value for the case where testing data is 20% of the total data samples. It can also be seen that RF and hard voting have acceptable accuracy, recall and AUC values. However, XGB has the highest AUC values of 96%, followed by hard voting having an AUC value of 95%.

Hence, from Table 6, it can be seen that XGB, RF and voting hard algorithms perform well when the dataset is divided into 20% testing and 80% training samples.

Next, the AUC values are computed for different classifiers: RF, LR, SVM, XGB, stacking 1, stacking 2, AdaBoost, bagging LR, hard voting and soft voting. For the case of 10% testing samples, LR has the lowest AUC value of 0.58 or 58%. RF and hard voting have the highest AUC values being 0.91 or 91%. On the other hand, XGB has an AUC value of 89%. Hence, RF and hard voting algorithms can perform well in classifying patients requiring ICU/semi-ICU and normal patients when the dataset is divided into 10% testing and 90% training data.

Table 6. Accuracy and other performance metrics of the classifiers for 20% testing data

Classifier	Weighted Average Precision	Weighted Average Recall	Miss rate/ FNR	Specificity	Weighted Average F1 score	Accuracy	AUC
RF	97%	97%	3%	97.58%	97%	97.43%	92%
LR	96%	90%	10%	98.79%	92%	89.81%	60%
SVM	94%	96%	4%	96.63%	95%	96.46%	65%
XGB	98%	98%	2%	98.11%	98%	97.96%	96%
Stacking 1	98%	98%	2%	98.11%	98%	97.97%	92%
Stacking 2	97%	97%	3%	97.93%	97%	97.61%	89%
AdaBoost	98%	98%	2%	99.73%	98%	98.14%	93%
Bagging LR	96%	97%	3%	99.63%	96%	96.81%	80%
Voting (RF, LR, SVM) Hard	98%	98%	2%	99.91%	97%	97.61%	95%
Voting (RF, LR, SVM) Soft	98%	98%	2%	99.73%	98%	97.87%	92%

5 Conclusion

This work applies machine learning algorithms in describing two aspects of COVID-19. Firstly, the ARIMA model is shown to successfully model the increase in the number of confirmed COVID-19 cases in the world. Moreover, the model also works for the case of number of confirmed cases of two individual countries India and the USA, which are badly affected by the virus. This ARIMA model can also predict the number of confirmed cases in the future. However, the future confirmed cases depend on a number of factors, including how lockdown or social distancing, and personal hygiene are maintained. Secondly, this paper applies machine learning algorithms on a dataset to predict the most important attributes that may lead to the prediction of a patient's need for ICU or semi ICU admission. RF and hard voting classifiers are found to perform the best in classifying patients who need ICU/semi-ICU admission and who need no such admission. However, the results of the most important attribute and the predictions depend on the reliability of the dataset. The effectiveness of the ARIMA algorithm in modelling the spread of COVID-19 cases may vary for different data samples. Moreover, the effectiveness of the classifiers to predict ICU admission may also vary when different datasets are taken into consideration. In future, a number of other regression models along with ARIMA should be applied to different COVID-19 spread datasets including the second wave dataset of Europe. Furthermore, the reliability of the classifiers in classifying ICU or semi ICU admission has to be validated for more COVID-19 datasets.

6 Acknowledgment

Most of this work has been carried out at Institute of Information and Communication Technology (IICT), Bangladesh University of Engineering and Technology (BUET), Dhaka-1205, Bangladesh. Hence, the authors would like to thank IICT, BUET for its support.

7 References

- [1] Zhu N, Zhang D, Wang W, et al. (2020). China Novel Coronavirus Investigating and Research Team. A novel coronavirus from patients with pneumonia in China, 2019. *N Engl J Med.*, 382(8):727-733. doi:10.1056/NEJMoa2001017
- [2] Mailles A, Blanckaert K, Chaud P, van der Werf S, Lina B, Caro V, et al. (2013). First cases of Middle East respiratory syndrome Coronavirus (MERS-CoV) infections in France, investigations and implications for the prevention of human-to-human transmission, *Euro Surveill.*, 18:20502.
- [3] Buchholz U, Müller MA, Nitsche A, Sanewski A, Wevering N, Bauer-Balci T, et al. (2012). Contact investigation of a case of human novel coronavirus infection treated in a German hospital, October-November 2012. *Euro Surveill.*, 18:20406.
- [4] Saif LJ. (2004). Animal coronaviruses: what can they teach us about the severe acute respiratory syndrome? *Rev Sci Tech.*, 23:643–660.
- [5] Gwaltney JM Jr. (1985). Virology and immunology of the common cold. *Rhinology*, 23:265.
- [6] Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. (2020). A Novel Coronavirus from Patients with Pneumonia in China, 2019. *New England Journal of Medicine*, 382:727-733
- [7] Subrato Bharati (2020). How Artificial Intelligence Impacts Businesses in the Period of Pandemics?, *Journal of the International Academy for Case Studies*, 26(5): 1-2.
- [8] Chan JF-W, Yuan S, Kok K-H, To KK-W, Chu H, Yang J, et al. (2020). A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *The Lancet*, 395: 514-523.
- [9] Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. (2020). A novel coronavirus from patients with pneumonia in China, 2019. *N Engl J Med.*, 382(8): 727–33.
- [10] Yin Y, Wunderink RG. (2018). MERS, SARS and other coronaviruses as causes of pneumonia. *Respirology*, 23(2):130–7.
- [11] Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, et al. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*. <https://doi.org/10.1038/s41586-020-2012-7>.
- [12] Wu A, Peng Y, Huang B, Ding X, Wang X, Niu P, et al. (2020). Genome composition and divergence of the novel coronavirus (2019-nCoV) originating in China. *Cell Host Microbe*. doi:10.1016/j.chom.2020.02.001.
- [13] Angeletti S, Benvenuto D, Bianchi M, Giovanetti M, Pascarella S, Ciccozzi M. (2020). COVID-2019: the role of the nsp2 and nsp3 in its pathogenesis. *J Med Virol*. <https://doi.org/10.1002/jmv.25719>.
- [14] Zhang L, Shen FM, Chen F, Lin Z. (2020). Origin and evolution of the 2019 novel coronavirus. *Clin Infect Dis.*, 71(15):882-883. <https://doi.org/10.1093/cid/ciaa112>.
- [15] Wu D, Zou S, Bai T, Li J, Zhao X, Yang L, et al. (2015). Poultry farms as a source of avian influenza A (H7N9) virus reassortment and human infection. *Sci Rep.*, 5:7630.
- [16] Novel Corona Virus 2019 Dataset, <https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset> (last accessed on 4 May, 2020).
- [17] COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University, <https://github.com/CSSEGISandData/COVID-19> (last accessed on 4 May, 2020).
- [18] Sarkar J, Chakrabarti P. (2020). A Machine Learning Model Reveals Older Age and Delayed Hospitalisation as Predictors of Mortality in Patients with COVID-19. *medRxiv*. DOI: 10.1101/2020.03.25.20043331.
- [19] Domenico Benvenuto, Marta Giovanetti, Lazzaro Vassallo, Silvia Angeletti, Massimo Ciccozzi, Application of the ARIMA model on the COVID-2019 epidemic dataset, *Data in Brief*, Volume 29, 2020, 105340, ISSN 2352-3409, <https://doi.org/10.1016/j.dib.2020.105340>.
- [20] Ardabili, S. F. et al. (2020). Covid-19 outbreak prediction with machine learning. *Algorithms*, 13(10), 249.
- [21] Muhammad Adnan Shereen, Suliman Khan, Abeer Kazmi, Nadia Bashir, Rabeea Siddique, "COVID-19 infection: Origin, transmission, and characteristics of human coronaviruses", *Journal of Advanced Research*, ISSN 2090-1232, Volume 24, 2020, Pages 91-98, doi:10.1016/j.jare.2020.03.005.
- [22] Zhang, Y.; Yang, H.; Cui, H.; Chen, Q. Comparison of the Ability of ARIMA, WNN and SVM Models for Drought Forecasting in the Sanjiang Plain, China. *Nat. Resour. Res.* 2019, 29, 1447.

- [23] Tandon, Hiteshi, Prabhat Ranjan, Tanmoy Chakraborty and Vandana Suhag. "Coronavirus (COVID-19): ARIMA based time-series analysis to forecast near future", eprint arXiv: 2004.07859, April 2020.
- [24] Coronavirus disease (COVID-2019) situation reports, Situation report – 104, <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports> (last accessed on 6 May, 2020)
- [25] Yang, Xiaobo, Yu Yuan, Xu Jiqian, Shu Huaqing, Xia Jia'an, Liu Hong, et. al. (2020). Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. *The Lancet Respiratory Medicine*. 10.1016/S2213-2600(20)30079-5.
- [26] Andre Filipe de Moraes Batista, Joao Luiz Miraglia, Thiago Henrique Rizzi Donato, Alexandre Dias Porto Chiavegatto Filho, COVID-19 diagnosis prediction in emergency care patients: a machine learning approach. medRxiv. 2020.04.04.20052092; doi: <https://doi.org/10.1101/2020.04.04.20052092>.
- [27] Randhawa GS, Soltysiak MPM, El Roz H, de Souza CPE, Hill KA, Kari L (2020) Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study. *PLoS ONE* 15(4): e0232391. <https://doi.org/10.1371/journal.pone.0232391>
- [28] Benvenuto D, Giovanetti M, Vassallo L, Angeletti S, Ciccozzi M. Data in brief application of the ARIMA model on the COVID-2019 epidemic dataset. *Data Br.* 2020. <https://doi.org/10.1016/j.dib.2020.105340>.
- [29] Mhdm R, Silva RG, Mariani VC, Coelho LS. (2020). Short-term forecasting COVID-19 cumulative confirmed cases: perspectives for Brazil. *Chaos Solitons Fractals*. 135:109853. <https://doi.org/10.1016/j.chaos.2020.109853>.
- [30] Chakraborty T, Ghosh I. (2020). Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: a data-driven analysis. *Chaos Solitons Fractals*. 135: 109850. <https://doi.org/10.1016/j.chaos.2020.109850>.
- [31] Wang, Y., Wang, C., Shi, C., & Xiao, B. (2018). Short-term cloud coverage prediction using the ARIMA time series model. *Remote Sensing Letters*, 9(3), 274-283.
- [32] Rahimi, Z. H., Khashei, M. (2018). A least squares-based parallel hybridization of statistical and intelligent models for time series forecasting. *Computers & Industrial Engineering*, 118, 44-53.
- [33] Khashei, M., & Hajirahimi, Z. (2019). A comparative study of series arima/mlp hybrid models for stock price forecasting. *Communications in Statistics-Simulation and Computation*, 48(9), 2625-2640.
- [34] Chen, L., & Xu, H. (2012). Autoregressive integrated moving average model in food poisoning prediction in Hunan province," *Journal of Central South University*, vol. 37, no. 2, pp. 142–146.
- [35] Xu, D. W., Wang, Y. D., Jia, L. M., Qin, Y., & Dong, H. H. (2017). Real-time road traffic state prediction based on ARIMA and Kalman filter. *Frontiers of Information Technology & Electronic Engineering*, 18(2), 287-302.
- [36] Alsharif, M.H., Younes, M.K., Kim, J. (2019). Time Series ARIMA Model for Prediction of Daily and Monthly Average Global Solar Radiation: The Case Study of Seoul, South Korea. *Symmetry*, 11, 240.
- [37] Raihan-Al-Masud M., Mondal MRH (2020) Data-driven diagnosis of spinal abnormalities using feature selection and machine learning algorithms. *PLoS ONE* 15(2): e0228422. <https://doi.org/10.1371/journal.pone.0228422>
- [38] Subrato Bharati, Prajoy Podder, and M. Rubaiyat Hossain Mondal. Diagnosis of Polycystic Ovary Syndrome Using Machine Learning Algorithms. 2020 IEEE Region 10 Symposium (TENSYP). IEEE, 2020.
- [39] Bharati S., Podder P., Mondal R., Mahmood A., Raihan-Al-Masud M. (2020). Comparative Performance Analysis of Different Classification Algorithm for the Purpose of Prediction of Lung Cancer. In: Abraham A., Cherukuri A., Melin P., Gandhi N. (eds) *Intelligent Systems Design and Applications*. 2018. *Advances in Intelligent Systems and Computing*. 941: 447-457.
- [40] Mondal MRH, Bharati S, Podder P, Podder P. (2020). Data analytics for novel coronavirus disease. *Informatics in Medicine Unlocked*. 20:100374.
- [41] Bharati S, Podder P, Mondal MRH. (2020). Hybrid deep learning for detecting lung diseases from X-ray images. *Informatics in Medicine Unlocked*. 20:100391.
- [42] Khamparia, A., Bharati, S., Podder, P. et al. (2021). Diagnosis of breast cancer based on modern mammography using hybrid transfer learning. *Multidim Syst Sign Process*. <https://doi.org/10.1007/s11045-020-00756-7>.
- [43] Bharati S., Robel M.R.A., Rahman M.A., Podder P., Gandhi N. (2021) Comparative Performance Exploration and Prediction of Fibrosis, Malign Lymph, Metastases, Normal Lymphogram Using Machine Learning Method. In: Abraham A., Panda M., Pradhan S., Garcia-Hernandez L., Ma K. (eds) *Innovations in Bio-Inspired Computing and Applications*. IBICA 2019. *Advances in Intelligent Systems and Computing*, vol 1180. Springer, Cham. https://doi.org/10.1007/978-3-030-49339-4_8
- [44] Bharati, S., Podder, P., and Paul, PK. (2019). Lung Cancer Recognition and Prediction According to Random Forest Ensemble and RUSBoost Algorithm Using LIDC Data. *International Journal of Hybrid Intelligent Systems*, 15(2): 91 – 100. DOI: 10.3233/HIS-190263
- [45] Bharati S, Podder P, Mondal MRH. (2020). Artificial Neural Network Based Breast Cancer Screening: A Comprehensive Review. *International Journal of Computer Information Systems and Industrial Management Applications*. 12:125-137.

- [46] Robel MRA, Bharati S, Podder P, Mondal MRH. (2020). IoT Driven Healthcare Monitoring System. Fog, Edge, and Pervasive Computing in Intelligent IoT Driven Applications. 30:161-176.
- [47] Madhavan MV, Thanh DNH, Khamparia A, Pande S, Malik R, Gupta D. (2021). Recognition and Classification of Pomegranate Leaves Diseases by Image Processing and Machine Learning Techniques. 66(3): 2939–2955.
- [48] Podder P, Bharati S, Mondal MRH. 10 Automated gastric cancer detection and classification using machine learning. In: Gupta D, Kose U, Le Nguyen B, Bhattacharyya S, Nguyen B (ed.) Artificial Intelligence for Data-Driven Medical Diagnosis. Berlin, Boston: De Gruyter; 2021. p.207-224. <https://doi.org/10.1515/9783110668322-010>
- [49] Bharati, S. & Podder, P. (2021). 1 Performance of CNN for predicting cancerous lung nodules using LightGBM. In D. Gupta, U. Kose, B. Le Nguyen, S. Bhattacharyya & B. Nguyen (Ed.), Artificial Intelligence for Data-Driven Medical Diagnosis (pp. 1-18). Berlin, Boston: De Gruyter. <https://doi.org/10.1515/9783110668322-001>
- [50] Bharati S., Podder P., Mondal M.R.H., Paul P.K. (2021). Applications and Challenges of Cloud Integrated IoMT. In: Hassanien A.E., Khamparia A., Gupta D., Shankar K., Slowik A. (eds) Cognitive Internet of Medical Things for Smart Healthcare. Studies in Systems, Decision and Control, vol 311. Springer, Cham. https://doi.org/10.1007/978-3-030-55833-8_4

8 Authors

Prajoy Podder is currently working as a masters researcher in the Institute of Information and Communication Technology (IICT), Bangladesh University of Engineering & Technology (BUET), Dhaka, Bangladesh. He published several Journal papers in reputed Journal and Conferences.

Aditya Khamparia PhD, works as an Associate Professor, Department of Computer Science, Lovely Professional University, Punjab, India. His research area is Machine Learning, Soft Computing, Educational Technologies, IoT, Semantic Web and Ontologies. He has published more than 50 scientific research publications in reputed International/National Journals and Conferences, which are indexed in various international databases.

M. Rubaiyat Hossain Mondal PhD, is currently working as a Professor in the Institute of Information and Communication Technology (IICT) at BUET, Dhaka, Bangladesh. His research interest includes artificial intelligence, bioinformatics, image processing, wireless communication and optical wireless communication.

Mohammad Atikur Rahman received BS in EEE from Ranada Prasad Shaha University, Narayanganj-1400, Bangladesh. He published several papers in reputed Conferences.

Subrato Bharati is currently working as a research assistant at IICT, BUET, Dhaka, Bangladesh. His research interest includes bio-informatics, wireless communication, image processing, machine learning and data science. He published several Journal papers in reputed Journal and Conferences.