Journal of Multidisciplinary Healthcare

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

# Early Prediction of COVID-19 Ventilation Requirement and Mortality from Routinely Collected Baseline Chest Radiographs, Laboratory, and Clinical Data with Machine Learning

Abdulrhman Fahad Aljouie,[1,2] Ahmed Almazroa,[2,3] Yahya Bokhari,[1,2] Mohammed Alawad,[1,2] Ebrahim Mahmoud,[4] Eman Alawad,[4] Ali Alsehawi,[5] Mamoon Rashid,[1,2] Lamya Alomair,[1,2] Shahad Almozaai,[6] Bedoor Albesher,[6] Hassan Alomaish,[5] Rayyan Daghistani,[5] Naif Khalaf Alharbi,[2,7] Manal Alaamery,[2,8–10] Mohammad Bosaeed,[2–4] Hesham Alshaalan[5]

[1]Bioinformatics Section, King Abdullah International Medical Research Center, Riyadh, Saudi Arabia; [2]King Saud bin Abdulaziz University for Health Sciences, Riyadh, Saudi Arabia; [3]Department of Imaging Research, King Abdullah International Medical Research Center, Riyadh, Saudi Arabia; [4]Department of Medicine, Ministry of National Guard - Health Affairs, Riyadh, Saudi Arabia; [5]Department of Medical Imaging, Ministry of National Guard - Health Affairs, Riyadh, Saudi Arabia; [6]College of Pharmacy, King Saud bin Abdulaziz University for Health Sciences, Riyadh, Saudi Arabia; [7]Department of Infectious Disease Research, King Abdullah International Medical Research Center, Riyadh, Saudi Arabia; [8]Department of Developmental Medicine, King Abdullah International Medical Research Center, Riyadh, Saudi Arabia; [9]KACST-BWH Center of Excellence for Biomedicine, Joint Centers of Excellence Program, King Abdulaziz City for Science and Technology (KACST), Riyadh, Saudi Arabia; [10]King Abdulaziz City for Science and Technology (KACST)-Saudi Human Genome Satellite Lab at Abdulaziz Medical City, Ministry of National Guard Health Affairs (MNGHA), Riyadh, Saudi Arabia

Correspondence: Mohammed Alawad
Bioinformatics Section, King Abdullah International Medical Research Center, King Saud bin Abdulaziz University for Health Sciences, Riyadh, Saudi Arabia
Email malawad@sdaia.gov.sa

Ahmed Almazroa
Department of Imaging Research, King Abdullah International Medical Research Center King Saud bin Abdulaziz University for Health Sciences, Riyadh, Saudi Arabia
Email almazroaah@ngha.med.sa

**Background:** Coronavirus disease 2019 (COVID-19), caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), emerged in Wuhan, China, in late 2019 and created a global pandemic that overwhelmed healthcare systems. COVID-19, as of July 3, 2021, yielded 182 million confirmed cases and 3.9 million deaths globally according to the World Health Organization. Several patients who were initially diagnosed with mild or moderate COVID-19 later deteriorated and were reclassified to severe disease type.

**Objective:** The aim is to create a predictive model for COVID-19 ventilatory support and mortality early on from baseline (at the time of diagnosis) and routinely collected data of each patient (CXR, CBC, demographics, and patient history).

**Methods:** Four common machine learning algorithms, three data balancing techniques, and feature selection are used to build and validate predictive models for COVID-19 mechanical requirement and mortality. Baseline CXR, CBC, demographic, and clinical data were retrospectively collected from April 2, 2020, till June 18, 2020, for 5739 patients with confirmed PCR COVID-19 at King Abdulaziz Medical City in Riyadh. However, of those patients, only 1508 and 1513 have met the inclusion criteria for ventilatory support and mortalilty endpoints, respectively.

**Results:** In an independent test set, ventilation requirement predictive model with top 20 features selected with reliefF algorithm from baseline radiological, laboratory, and clinical data using support vector machines and random undersampling technique attained an AUC of 0.87 and a balanced accuracy of 0.81. For mortality endpoint, the top model yielded an AUC of 0.83 and a balanced accuracy of 0.80 using all features with balanced random forest. This indicates that with only routinely collected data our models can predict the outcome with good performance. The predictive ability of combined data consistently outperformed each data set individually for intubation and mortality. For the ventilator support, chest X-ray severity annotations alone performed better than comorbidity, complete blood count, age, or gender with an AUC of 0.85 and balanced accuracy of 0.79. For mortality, comorbidity alone achieved an AUC of 0.80 and a balanced accuracy of 0.72, which is higher than models that use either chest radiograph, laboratory, or demographic features only.

**Conclusion:** The experimental results demonstrate the practicality of the proposed COVID-19 predictive tool for hospital resource planning and patients' prioritization in the current COVID-19 pandemic crisis.

**Keywords:** COVID-19, mortality, NIV, X-rays, CBC, random forest, SMOTE; machine learning

Received: 31 May 2021
Accepted: 15 July 2021
Published: 30 July 2021

## Introduction

Globally, the cumulative number of Coronavirus Disease 2019 (COVID-19) cases is increasing daily. Consequently, it has caused a rapid surge of critically ill patients with ventilatory support and mortality rates.[1] Although the respiratory system is the primary target for the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), other organ systems complications can also participate in the cause of death from COVID-19.[2,3] Based on the clinical experience, SARS-CoV-2 infection has shown substantial heterogeneity; it spans from asymptomatic patients to mild, moderate, and severe disease forms with low survival rates.[4] Many studies[5–7] have reported that some of the COVID-19 patients are assigned moderate disease grade at initial diagnosis and later, during the course of the disease, are reclassified into severe type due to new or worsening symptoms. Hence, early estimation of COVID-19′ severity helps clinicians prioritize patients and monitor cases that are more likely to deteriorate for timely intervention. It also helps them to plan hospital resources better. There are two main streams of studies to assess the severity of COVID-19: 1) studies to investigate biomarkers to test their association with severe cases, and 2) studies aiming to build severity predictive models from clinical symptoms, tests results, or chest radiographic scans collected at the time of diagnosis, or at a predefined fixed time before the occurrence of the outcome endpoint. Many studies[5,8,9] have reported that in severe cases, the significant factors are age,[8,9] comorbidity,[8,9] and lymphopenia.[5,8,9] Other studies have identified features in X-ray of COVID-19 patients that are associated with severe disease types, such as bilateral peripheral ground-glass opacity (GGO), consolidation, and other radiological manifestations.-[10,11] Toussie et al have investigated chest radiography as an independent prognostic factor of the disease outcome. The authors tested the association among hospital admissions, intubation, prolonged stay (>10 days), and the severity of baseline X-ray images in COVID-19 young patients, after adjusting for demographics and comorbidity.[12] Each X-ray image was divided into six regions (three zones per lung), where each region was assigned a score. The overall score ranged between 0 and 6. Instead, here we investigated the use of chest radiographs in adults with laboratory findings, clinical, and demographic data.

For the related work, Shi et al[13] used five features: age, lactate dehydrogenase (LDH), C-reactive protein (CRP), CD4+ T cell counts, and mass of infection (MOI) in the whole lung. The MOI is a quantitative parameter, obtained from a computer tomography (CT) scan, to predict patient's infection severity from baseline indicators. Furthermore, Gong et al[14] utilized age and six serological variables (ie serum lactate dehydrogenase, C-reactive protein, red blood cell distribution width variation coefficient, blood urea nitrogen, albumin, and direct bilirubin) as an input to least absolute shrinkage and selection operator (LASSO) algorithm and logistic regression to predict severe versus nonsevere cases. Various studies have attempted to utilize diagnostic tests and clinical data to predict COVID-19 outcome.[15–18] Cheng et al trained a random forest model with time-series data (vital signs, nursing assessment, laboratory data, and electrocardiograms) to predict within 24 hours ICU transfer. They achieved an accuracy of 76% on the test set (30% of the original data). Wu et al[19] developed four logistic regression models to predict severe vs nonsevere COVID-19 types. For model 1, they used hospital employment and age, which achieved an AUC of 0.83 on the validation set. Model 2 used hospital employment, age, body temperature, and time of onset and achieved 0.74 AUC on the validation set. Model 3 was based on CT semantic features and age and achieved an AUC of 83 on the validation set. Model 4 used seven features (age, lymphocyte (proportion), CRP, LDH, creatine kinase, urea and calcium) and yielded an AUC of 0.90 on the validation set.[19] Ryan et al[16] compared machine learning to Sepsis Related Organ Failure Assessment (qSOFA), Modified Early Warning Score (MEWS), and CURB-65 severity scores to predict patients outcome in Medical Information Mart for Intensive Care (MIMIC) and COVID-19 data from a community hospital. They have built an XGBoost model to predict in-hospital mortality of COVID-19 patients at 12-, 24-, 48-, and 72-hours. In the community hospital data, their model yielded an F1 score of 0.57 when predicting mortality before 72-hours using the last three collected observations of laboratory and clinical variables, and the model outperformed qSOFA, MEWS, and CURB-65 risk scores. Note that the data is imbalanced. However, the trained XGBoost model does not account for the effect of skewed class distribution, which may explain its low F1 score on the test set.

Organizations such as Fleischner Society have issued a consensus statement exploring the application of imaging in patient's diagnosis and risk stratification.[20] The American College of Radiology and the Society of Thoracic Radiology have also recommended against the use of CT scan and two-view chest radiography for large-scale screening and diagnosis.[21] However, various investigations[22–26] have examined the utility of imaging for screening and prognosis of COVID-19 and have demonstrated high classification accuracy rates.

The purpose of the study is to develop a holistic mortality and ventilation requirement machine learning-based classification models using a heterogeneous combination of a patient's chest X-ray (CXR), laboratory, underlying health condition, age, and gender data. This data was collected at the time of diagnosis of patients infected with COVID-19, at King Abdulaziz Medical City, Riyadh, Saudi Arabia, to support an early decision to predict the severity of COVID-19 disease.

## Materials and Methods

### Data Collection and Study Design

We retrospectively collected data of admitted confirmed COVID-19 patients (positive RT-PCR test) to King Abdulaziz Medical City in Riyadh. Starting from the first case, which was on April 2, 2020, till June 18, 2020; the total number of patients is 5739.

The aim of this study is to assess combining the baseline (within 3 days of COVID-19 diagnosis) chest X-ray (CXR) image severity annotations, complete blood counts (CBC), age, gender, and comorbidity data of each patient to predict at the diagnosis two endpoints: ventilation requirement and mortality. We further investigate the ability to differentiate between mechanical ventilation (MV) and non-invasive ventilation (NIV).

For ventilation requirement, Figure 1 shows the inclusion criteria flowchart, ie, a patient that lacks one of the following baseline features is excluded: 1) confirmed RT-PCR for COVID-19, 2) age >18, 3) CXR, 4) CBC test results, 5) availability of medical history, and 6) Full code status at the time of COVID-19 diagnosis. The total patients that met these criteria are 1508 with a mean age of 54.8 ±16.9 and 43% females and 57% males. The intubation or invasive ventilation status is assigned if the patient underwent NIV or MV within 30 days of the admission date.

For the mortality endpoint, the inclusion criteria are identical to ventilation requirements except that we included COVID-19 patients assigned no-code before the SARS-CoV-2 test as shown in Figure 2. The total patients that met the inclusion criteria for mortality endpoint are 1513. A comparison between patients groups age, gender, CBC, and comorbidity for the mortality and ventilator support is described in Tables 1 and 2.

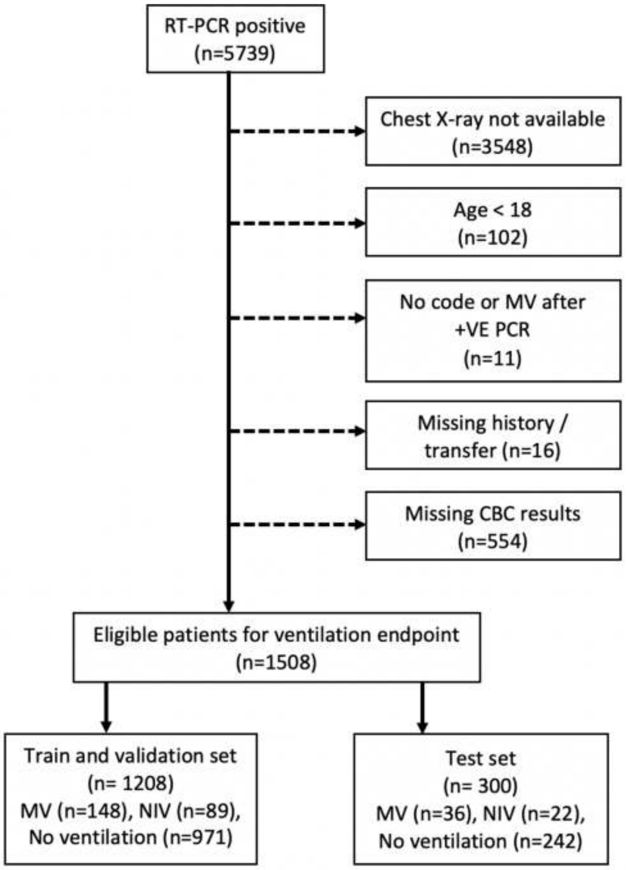

**Figure 1** Flowchart of selection criteria for ventilation requirement prediction endpoint. Patients that have any missing data are excluded from this study.

The local hospital ventilation criteria in COVID-19 were obtained from the Saudi Ministry of Health Mechanical Ventilation Protocol for COVID-19 (May 2020); this protocol is aligned with other international recommendations and previous guidelines to provide objective parameters for intubation and ventilation of COVID-19 patients.[1,27,28] The indications for mechanical ventilation were: increase work of breathing and sign of organ failure (eg, altered mental status, low BP, increased lactate, signs of cardiac ischemia), acute hypoxic respiratory failure not responding to HFNC nor NIV for a maximum of 2 hours. Hypoxia with acute decreased level of consciousness and cannot protect his airway, hypoxia with large copious secretions, hypercapnic respiratory failure not responding to HFNC nor NIV, hemodynamically unstable, and to consider for a patient on HFNC or NIV therapy and transfer by ambulance.

The CBC data contain ten features: hematocrit, hemoglobin, mean corpuscular hemoglobin concentration (MCHC), mean corpuscular hemoglobin (MCH), mean corpuscular

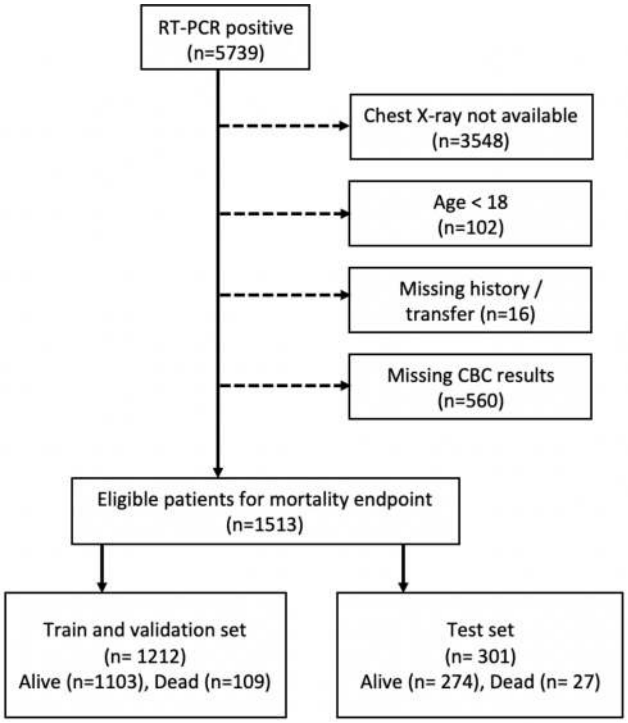
RT-PCR positive (n=5739)
Chest X-ray not available (n=3548)
Age < 18 (n=102)
Missing history / transfer (n=16)
Missing CBC results (n=560)
Eligible patients for mortality endpoint (n=1513)
Train and validation set (n= 1212) Alive (n=1103), Dead (n=109)
Test set (n= 301) Alive (n= 274), Dead (n= 27)

**Figure 2** Patients selection criteria flowchart for mortality prediction.

volume (MCV), mean platelet volume (MPV), red blood cell count (RBC), platelet count, red cell distribution width (RDW), and white blood cells (WBC). The comorbidity data contain ten binary features: cancer, coronary artery disease, hypertension, asthma, chronic obstructive pulmonary disease (COPD), type 2 diabetes, liver cirrhosis, chronic hepatitis B, chronic hepatitis C, and chronic kidney disease (CKD) all stages. The CXR data contain 12 features annotated by radiologists and is explained in detail in the subsequent section. The CBC, age, gender, and COVID-19 test results were automatically obtained from electronic health records, whereas the comorbidity, non-invasive and mechanical ventilation, and mortality data were manually curated from admission and discharge charts.

## Radiologists Scoring of Chest Radiographs

Baseline chest X-rays were divided into 12 regions, six regions per lung. Images were distributed randomly among four radiologists with experience of 3–15 years. Each radiologist then scored each region according to the presentation of ground-glass opacity and consolidation severity. Zero is assigned if no manifestation is found, one for mild/moderate, and two for severely affected zones. Figure 3 demonstrates the segmentation of chest posteroanterior X-ray. The resulting matrix was then used alone to train a discriminative classifier to predict both endpoints. Additionally, it was combined with other laboratory and clinical features.

## Chest X-Rays Inter-Rater Variability Assessment

To gauge the inter-rater variability, we randomly selected 7% (n=110) of patients' images. All four radiologists scored each zone of these selected images. Accordingly, we had a total of 12 matrices corresponding to the number of lung zones annotations (see Figure 3). In each matrix, the rows are patients and the columns are radiologists' ratings. We assessed the inter-rater agreement with Fleiss-kappa statistic, and we used kappam.fleiss from irr package in R software (version 3.5.1).

## Data Preprocessing and Feature Selection

Here, we describe the method used for handling missing values, feature normalization, and feature selection to remove irrelevant ones. Patients with any missing data were excluded from the analysis to limit bias, and thus we eliminated the need for imputation. As a preprocessing step, we standardized the data using z-score before fitting a model with a support vector machine (SVM) since the features present in the data differ in magnitude. When using SVM, features with larger values will dominate. Besides, it is challenging to calculate kernel, therefore, standardization can speed up the training phase. In the training set, the z-score transforms separate distributions of features in the training set into a standardized distribution. Each feature vector has a value between [−1, +1], where each data point has a 0 mean and 1 standard deviation from the distribution mean. Data points that are above the mean get a positive score, and a negative score if the data points are below the mean. We apply the same transformation for the test set. We used scikit-learn[29] StandardScaler and the implementation is described in equation 1.

$$X_j' = \frac{X_j - u_j}{s_j} \tag{1}$$

where $X_j$' is the jth transformed feature. $X_j$ is the original *jth* feature, $u_j$ and $s_j$ are the samples mean and standard deviation of the jth feature in the training set.

Feature selection technique is a common preprocessing step in machine learning to improve classifiers performance and reduce training and inference time. In our analysis, we used ReliefF.[30] Most heuristic filter approaches assume no interactions between features; however, relief-based

**Table 1** COVID-19 Cohort Ventilation Requirement and Clinical Features Descriptive Statistics

| Feature | No Ventilation (n= 1213) (Mean ±SD, IQR) | Non-Invasive Ventilation (n=111) (Mean ±SD, IQR) | Mechanical Ventilation (n=184) (Mean ±SD, IQR) | P-value |
|---|---|---|---|---|
| Age | (52.52 ± 16.94, 25) | (60.23 ± 15.76, 9.5) | (61.69 ± 14, 20.25) | 1.5e–15* |
| Male/female | 635/578 | 82/29 | 140/44 | 8.9e–12* |
| Haematocrit | (0.42 ± 0.06, 0.075) | (0.43 ± 0.05, 0.026) | (0.42 ± 0.07, 0.074) | 0.19 |
| Hgb, g/L | (134.12 ± 19.7, 25) | (137.5 ± 17.4, 9) | (135.37 ± 23.26, 25) | 0.14 |
| MCHC, g/L | (319.75 ± 10.5, 13) | (320.71 ± 10.1, 6) | (319.72 ± 10.78, 13) | 0.62 |
| MCH, pg, | (27.79 ± 2.66, 2.7) | (28.32 ± 1.87, 1) | (28.11 ± 2.23, 2.32) | 0.011 |
| MCV, fL | (86.92 ± 7.86, 8.3) | (88.3 ± 5.43, 2.59) | (87.96 ± 6.87, 7.57) | 0.017 |
| MPV, fL | (8.21± 1.04, 1.3) | (8.08 ± 1.17, 0.60) | (7.95 ± 1, 1.32) | 0.004* |
| RBC, $10^{12}$/L | (4.84 ± 0.67, 0.77) | (4.87 ± 0.64, 0.35) | (4.82 ± 0.78, 0.89) | 0.85 |
| Platelet count | (250.1 ± 90.4, 106) | (250.4 ± 115.3, 54) | (255.22 ± 101.6, 132.7) | 0.81 |
| RDW, % | (13.55 ± 1.81, 1.7) | (13.5 ± 1.78, 0.6) | (13.77 ± 1.44, 1.62) | 0.17 |
| WBC, $10^9$/L | (6.89 ± 4.26, 3.4) | (9.01 ± 3.85, 2.27) | (9.47 ± 5.24, 6.19) | 4.8e-13* |
| Cancer, count | 47 (3.87%) | 3 (2.70%) | 6 (3.26%) | 0.90 |
| CAD[a], count | 107 (8.82%) | 10 (9.01%) | 24 (13.04%) | 0.18 |
| Hypertension, count | 511 (42.13%) | 59 (53.15%) | 108 (58.70%) | 2.7e-05* |
| Asthma, count | 135 (11.13%) | 10 (9.01%) | 15 (8.15%) | 0.40 |
| COPD[b], count | 11 (0.91%) | 3 (2.70%) | 2 (1.09%) | 0.13 |
| T2D[c], count | 511 (42.13%) | 57 (51.35%) | 109 (59.24%) | 2.8e-05* |
| Liver cirrhosis, count | 10 (0.82%) | 1 (0.90%) | 3 (1.63%) | 0.50 |
| CHB[d], count | 8 (0.66%) | 0 (0.00%) | 1 (0.54%) | 1 |
| Chronic HCV[e], count | 3 (0.25%) | 0 (0.00%) | 1 (0.54%) | 0.58 |
| CKD[f], count | 97 (8.00%) | 11 (9.91%) | 27 (14.67%) | 0.01 |

**Notes**: [a]Coronary artery disease, [b]Chronic obstructive pulmonary disease, [c]Chronic hepatitis B, [d]Type 2 diabetes, [e]Chronic hepatitis C virus, [f]Chronic kidney disease (all stages), *p<0.01.

algorithms do not make this assumption and thus are able to detect feature dependencies.[31] ReliefF first sets a 0 value to all features in the quality weight W[F]. Then, the algorithm loops through $m$ random (without replacement) observations $R$. For a given sample $R_i$, RelifF finds its $k$ nearest neighbors from the same class $H_j$ and $k$ nearest neighbor from each of the other classes $M_j$ (C), where C is the other classes in a multiclass dataset. The W[F] vector is then updated for all the features based on how these features separate the observation $R_i$ from $H_j$ and $M_j$. The weight vector is normalized to a value between [−1, 1], where a higher positive weight is assigned if the feature can differentiate $R_i$ observation from $M_j$ and a negative weight otherwise. We have set $k$=3, and m to the default value of RelifF Python package. The ReliefF features weights in the train set are shown in Table 3.

## Class Balancing in Train Set

The patients cohort's distribution of outcome classes is extremely skewed. Mechanical ventilation and mortality classes represent 7% and 9% of the data, respectively. When the training data are severely imbalanced, the classifier tends to focus more on learning the majority class and generally yields a lower predictive ability.[32] Hence, several resampling and cost-sensitive learning techniques have been proposed to tackle the class imbalance problem.[33] Here, we have used the Synthetic Minority Over-sampling TEchnique (SMOTE).[34] This increases the number of samples in the minority class by creating a random synthetic data point along the line of the feature space of the $k$ nearest neighbors to each minority class record. For $k$, we used the default value of 5 in imbalanced-learn implementation.[35] Thus, we increased the number of instances of the minority class to match the number of majority class instances. Another technique that we used to balance the training set is Adaptive Synthetic (ADASYN) sampling approach.[36] Unlike SMOTE algorithm, ADASYN upsamples the minority class by creating more instances from the harder to learn minority class examples. We also experimented with random undersampling (RUS), which randomly downsamples the majority class such that the training set becomes balanced.

**Table 2** COVID-19 Cohort Vital Status at Discharge and Clinical Features Descriptive Statistics

| **Feature** | **Alive (n=1377)**<br>**(Mean ±SD, IQR)** | **Deceased (n=136)**<br>**(Mean ±SD, IQR)** | **P-value** |
|---|---|---|---|
| Age | (53.42 ± 16.63, 24.29) | (69.17 ± 13.74, 19.02) | < 2.2e-16* |
| Male/female | 760/617 | 100/36 | 5.62e-05* |
| Haematocrit | (0.42 ± 0.06, 0.07) | (0.41 ± 0.07, 0.1) | 0.0344 |
| Hemoglobin (g/L) | (134.9 ± 19.8, 24) | (129.83 ± 22.97, 31) | 0.0139 |
| MCHC (g/L) | (319.92 ± 10.54, 14) | (318.56 ± 10.35, 13) | 0.1474 |
| MCH (pg) | (27.86 ± 2.59, 2.6) | (28.08 ± 2.26, 2.53) | 0.2804 |
| MCV (fL) | (87.07 ± 7.66, 7.9) | (88.18 ± 7.07, 8.55) | 0.0864 |
| MPV, fL | (8.18 ± 1.05, 1.3) | (8.08 ± 0.99, 1.2) | 0.2598 |
| RBC, $10^{12}$/L | (4.86 ± 0.67, 0.77) | (4.63 ± 0.79, 1.13) | 0.0013* |
| Platelet count | (250.34 ± 91.98, 110) | (252.24 ± 111.85, 109.25) | 0.8483 |
| RDW (%) | (13.52 ± 1.73, 1.7) | (14.21 ± 2.04, 2) | 0.0001* |
| WBC, $10^9$/L | (7.24 ± 4.48, 3.68) | (8.56 ± 4.08, 4) | 0.0004* |
| Cancer, count | 47 (3.41%) | 14 (10.29%) | 7.67e-05* |
| CAD[a], count | 107 (7.77%) | 28 (20.59%) | 5.57e-06* |
| Hypertension, count | 511 (37.11%) | 94 (69.12%) | 6.02e-09* |
| Asthma, count | 135 (9.80%) | 8 (5.88%) | 0.0817 |
| COPD[b], count | 11 (0.80%) | 3 (2.21%) | 0.168 |
| Type 2 diabetes, count | 511 (37.11%) | 97 (71.32%) | 1.82e-10* |
| Liver cirrhosis | 10 (0.73%) | 4 (2.94%) | 0.0306 |
| CHB[c], count | 8 (0.58%) | 1 (0.74%) | 0.5726 |
| Chronic HCV[d], count | 3 (0.22%) | 2 (1.47%) | 0.0426 |
| CKD[e] (all stages), count | 97 (7.04%) | 29 (21.32%) | 5.03e-07* |

**Notes**: [a]Coronary artery disease, [b]Chronic obstructive pulmonary disease, [c]Chronic hepatitis B, [d]Chronic hepatitis C virus, [e]Chronic kidney disease, *p<0.01.

## Classifiers

Here, we briefly describe the four classifiers used in this study: linear Support Vector Machine (SVM), Random Forest (RF), Linear Regression (LR), and eXtreme Gradient Boosting (XGB) to model COVID-19 disease outcome.

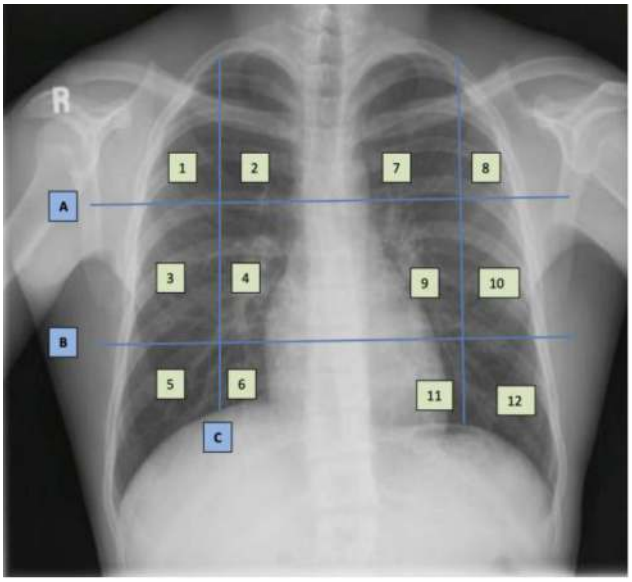

**Figure 3** Frontal chest X-ray lung zone segmentation. The horizontal lines A and B represent the upper and lower poles of the hilum. The vertical line C is from the junction of the middle/inner third of the clavicle to the diaphragm. The light green squares are the regions in which radiologists assign a severity score.

In a linearly separable two-class classification problem, SVM[37] finds the optimal hyperplane that maximizes the distance between the two classes' closest points, also called support vectors. In this study, SVM with a linear kernel is used. For non-linearly separable data, the cost term *C* is introduced to balance misclassification and maximum margin.The optimal *C* hyperparameter was selected using grid search in the validation set.

RF[38] is a type of ensemble method that constructs a strong classifier from many weak classifiers. For a classification problem, RF builds a user predefined number of decision trees to form a forest, and then it takes the majority vote of these trees to output a single class prediction. For each tree, RF selects random samples with replacement and randomly generates subset of features to decide each candidate node split; typically, the one with the highest Gini impurity or entropy. To measure the split quality, we have used Gini impurity, which is the default measure in Scikit-learn.[29]

BRF is a random forest classifier that takes into account the class distribution. BRF creates a balanced

**Table 3** Features Ranking with ReliefF in the Train Dataset

| Rank | MV[a] | | NIV[b] | | MV, NIV, None[c] | | Mortality | |
|---|---|---|---|---|---|---|---|---|
| | **Feature** | **Weight** | **Feature** | **Weight** | **Feature** | **Weight** | **Feature** | **Weight** |
| 1 | CXR (zone 11) | 0.23 | CXR (zone 11) | 0.21 | CXR (zone 11) | 0.20 | CXR (zone 11) | 0.24 |
| 2 | CXR (zone 12) | 0.19 | CXR (zone 12) | 0.14 | CXR (zone 12) | 0.17 | CXR (zone 12) | 0.24 |
| 3 | Age | 0.10 | Gender | 0.13 | CXR (zone 10) | 0.13 | Age | 0.19 |
| 4 | CXR (zone 5) | 0.10 | CXR (zone 9) | 0.10 | CXR (zone 5) | 0.12 | Type 2 diabetes | 0.15 |
| 5 | CXR (zone 10) | 0.10 | WBC | 0.10 | CXR (zone 3) | 0.10 | Gender | 0.15 |
| 6 | Gender | 0.10 | CXR (zone 6) | 0.08 | CXR (zone 4) | 0.08 | MCHC | 0.13 |
| 7 | CXR (zone 3) | 0.10 | CXR (zone 5) | 0.08 | CXR (zone 9) | 0.08 | CXR (zone 9) | 0.13 |
| 8 | CXR (zone 9) | 0.09 | CXR (zone 10) | 0.07 | Gender | 0.07 | CXR (zone 5) | 0.12 |
| 9 | MCHC | 0.08 | MPV | 0.07 | CXR (zone 6) | 0.07 | CXR (zone 6) | 0.11 |
| 10 | CXR (zone 6) | 0.08 | CXR (zone 4) | 0.07 | WBC | 0.05 | CXR (zone 10) | 0.10 |
| 11 | Type 2 diabetes | 0.08 | Age | 0.06 | CXR (zone 1) | 0.04 | CXR (zone 3) | 0.09 |
| 12 | MCV | 0.08 | Platelet_count | 0.06 | Haematocrit | 0.03 | CXR (zone 2) | 0.09 |
| 13 | MPV | 0.08 | CXR (zone 3) | 0.06 | CAD | 0.03 | Platelet_count | 0.09 |
| 14 | MCH | 0.07 | Type 2 diabetes | 0.05 | CKD | 0.03 | MCV | 0.08 |
| 15 | CXR (zone 4) | 0.07 | MCH | 0.05 | Hypertension | 0.02 | MPV | 0.08 |
| 16 | WBC | 0.06 | MCV | 0.04 | MCV | 0.02 | CXR (zone 1) | 0.08 |
| 17 | Hypertension | 0.06 | MCHC | 0.04 | Hemoglobin | 0.02 | CKD | 0.08 |
| 18 | CXR (zone 1) | 0.05 | RBC | 0.04 | COPD | 0.02 | MCH | 0.07 |
| 19 | Platelet_count | 0.05 | Hemoglobin | 0.04 | MCH | 0.02 | CAD | 0.07 |
| 20 | RBC | 0.05 | RDW | 0.03 | Platelet_count | 0.02 | CXR (zone 4) | 0.07 |
| 21 | Haematocrit | 0.05 | Haematocrit | 0.03 | RDW | 0.01 | RBC | 0.06 |
| 22 | Hemoglobin | 0.05 | Hypertension | 0.03 | CXR (zone 8) | 0.01 | Hypertension | 0.06 |
| 23 | RDW | 0.04 | CKD | 0.03 | CXR (zone 2) | 0.01 | Hemoglobin | 0.05 |
| 24 | Asthma | 0.04 | CXR (zone 8) | 0.02 | Chronic HCV | 0.00 | Haematocrit | 0.05 |
| 25 | CKD[d] | 0.04 | CXR (zone 2) | 0.02 | CXR (zone 7) | 0.00 | RDW | 0.05 |
| 26 | CXR (zone 2) | 0.03 | CXR (zone 1) | 0.02 | Liver cirrhosis | 0.00 | Cancer | 0.04 |
| 27 | CXR (zone 8) | 0.02 | CAD | 0.01 | RBC | 0.00 | CXR (zone 7) | 0.03 |
| 28 | CAD[e] | 0.01 | CXR (zone 7) | 0.01 | CHB | 0.00 | CXR (zone 8) | 0.03 |
| 29 | Chronic HCV[f] | 0.00 | Asthma | 0.00 | MCHC | −0.01 | WBC | 0.03 |
| 30 | CXR (zone 7) | 0.00 | COPD | 0.00 | Cancer | −0.01 | Asthma | 0.02 |
| 31 | Liver cirrhosis | 0.00 | Chronic HCV | 0.00 | Asthma | −0.01 | Chronic HCV | 0.01 |
| 32 | CHB[g] | 0.00 | CHB | 0.00 | MPV | −0.01 | Chronic HCV | 0.00 |
| 33 | COPD[h] | 0.00 | Liver cirrhosis | 0.00 | Type 2 diabetes | −0.02 | Liver cirrhosis | 0.00 |
| 34 | Cancer | −0.01 | Cancer | −0.01 | Age | −0.03 | COPD | 0.00 |

**Notes**: [a]Mechanical ventilation versus non-invasive ventilation or no ventilation, [b]Non-invasive ventilation or mechanical ventilation versus no ventilation, [c]Three-class classification (mechanical ventilation versus non-invasive ventilation versus no ventilation). [d]Chronic kidney disease, [e]Coronary artery disease, [f]Chronic hepatitis C virus, [g]Chronic hepatitis B, [h]Chronic obstructive pulmonary disease.

https://doi.org/10.2147/JMDH.S322431

train set via randomly selecting records at every iteration to build each tree by under-sampling the majority class.

Gradient boosting is an ensemble method that iteratively combines weak learners to build strong learner. It uses decision tree as base learner and adaptively adds trees that minimizes a differentiable loss function. We used XGBoost implementation of Gradient boosting in python.

## Model Selection and Performance Estimation

To assess the predictive ability and generalizability of the constructed models, we split the data into train set and independent test set with 80:20 ratio by conducting a stratified random sampling to make the class distribution equal in both the train and test sets. We further split the train set, using stratified sampling, to create a validation set (20% of the train set). Grid search was performed on the validation set to tune the four classifiers hyperparameters and to choose the best operating point (classification threshold) that maximizes Youden J index (see equation 5). We then refit a model on the entire train set with the selected optimal hyperparameters and operating point and predicted the test set.

For SVM with a linear kernel and logistic regression, using the train and validation sets, we searched for the best C value from the set: [0.1, 1, 10]. For random forest and XGBoost, we selected the best number of trees to grow from the set: [10, 100, 1000]. We used the same technique for the four prediction tasks performed in this study: MV, NIV and MV vs no ventilation, MV vs NIV vs no ventilation, and mortality.

## Performance Metrics

For imbalanced data, the overall accuracy is not an appropriate classifier performance measure. To assess the models' predictive ability in the test set, the balanced accuracy, and area under the ROC Curve (AUC) were used. Balanced accuracy takes class distribution into account by averaging the recall from each class; thus, it is a better measure of performance than accuracy for imbalanced data sets. For the best performing models, the confusion matrix is reported, which shows the number of true positive (TP), false positive (FP), true negative (TN) and false negative (FN) counts. Equations 2–4 show the metrics used in this study.

$$Sensitivity = \frac{TP}{TP + FN} \tag{2}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3}$$

$$Balanced\ accuracy = \frac{Sensitivity + Specificity}{2} \tag{4}$$

$$Youden\ index = Sensitivity + Specificity - 1 \tag{5}$$

AUC was used to select the best model with grid search of hyperparameters in the validation set and to select the optimal classification decision threshold (operating point) that maximizes the Youden index. Since the data is imbalanced, this performance metric takes class distribution into account by averaging the recall from each class; thus, it is a better measure of performance than accuracy for imbalanced data sets.

## Results

Among the patients included in our study, 7% underwent non-invasive ventilation (either high-flow nasal cannula (HFNC) or bi-level/continuous positive airway pressure (BIPAP/CPAP)) and 12% received mechanical ventilation. Table 1 shows the clinical characteristics and descriptive statistics for each ventilation requirement cohort subgroup separately, ie, patients who were not ventilated, patients who received only non-invasive ventilation, and patients who required mechanical ventilation. Data were analyzed using R software to compare groups using one-way ANOVA for continuous variables and Chi-square or two-tailed Fisher's exact test for categorical variables as appropriate.

Table 2 shows the clinical characteristics and descriptive statistics for patients discharge status, 9% of COVID-19 patients in this study died during hospitalization. The analysis revealed that younger patients and female gender are more likely to survive the disease.

To choose the best hyperparameter for classifiers, we performed grid search on 80% of the train set and validated on the validation set (20% of train set). Figure 4 shows the receiver operating characteristic curves (ROC curve) of the top constructed models and the point that maximizes the Youden index. In the validation set, the top MV model was built with logistic regression, random undersampling, regularization parameter C set to 10, and top 20 features selected using ReliefF. The model achieved an AUC of 0.80. For ventilation requirement (MV+NIV),

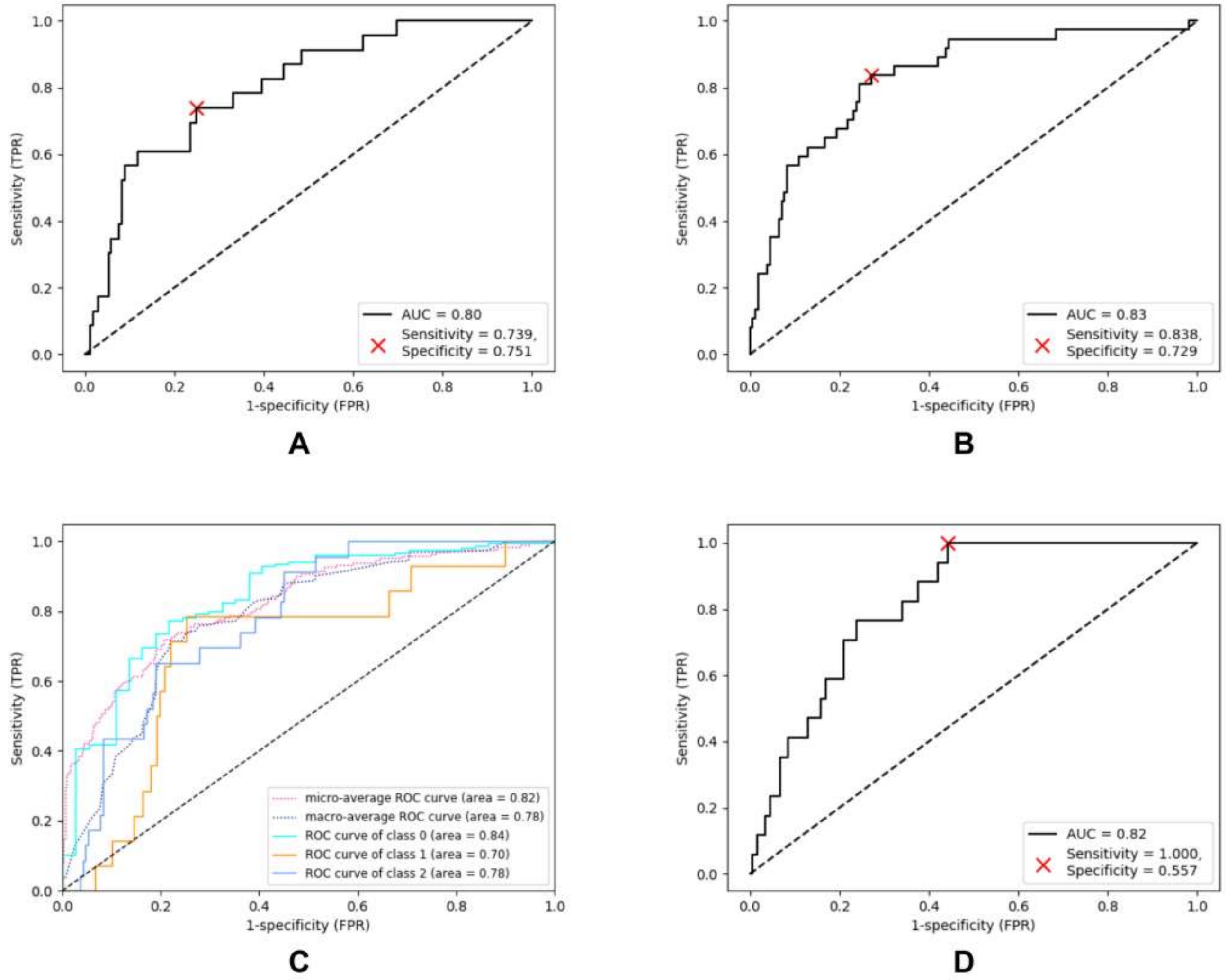

**Figure 4** Receiver operator characteristic curves for ventilation support and mortality end points prediction on the validation sets. The optimum cutoff point that maximizes Youden index, identified with a red X mark. (**A**) MV with LR and top 20 selected features, and random undersampling. (**B**) Ventilation requirement (MV+NIV) prediction with the best performing model, linear SVM with 20 top ranked features and random downsampling. (**C**) three-class classification (MV, NIV, no ventilation) with linear SVM and SMOTE on top ranked 20 features. (**D**) Mortality prediction with BRF and all features. The C hyperparameter for SVM represents the tradeoff between maximizing the margin minimizing the error, and the C for logistic regression represents the value assigned to control the regularization strength.

constructing a model with linear SVM and top 20 features with the misclassification penalty set to 1 yielded an AUC of 0.83. Also, we found that the best performing model for three classes classification (MV, NIV, and no ventilation) was built using SVM with linear kernel, SMOTE, top 20 features, and a C set to 0.1 with an AUC of 0.78. For mortality, the best performing model was created with balanced random forest and using all the combined features, with the number of estimators set to 1000. This model scored an AUC of 0.82 in the validation set.

The top-performing models in the test set for classifying 1) mechanical ventilation (MV) and non-invasive ventilation (NIV) + no ventilaton, 2) MV+NIV and no ventilation, and 3) three classes (MV, NIV, and no ventilation), and 4) alive, dead patients within 30 days follow-up are reported in Figure 5. The results show that, in the test set, the best model for MV prediction was obtained with logistic regression and random undersampling with an AUC of 82 and a balanced accuracy of 0.79. For MV+NIV prediction, the best model achieved an AUC of 0.86 and a balanced accuracy of 0.81. For the 3 class classification task (MV, NIV, and no ventilation), the best model yielded an AUC 0.78 of and a balanced accuracy of 0.56 with SVM and top 20 features selected with ReliefF. For mortality prediction, the top model attained 0.83 AUC and 0.80 balanced accuracy with BRF using the combined data. Table 4 shows the performance of the four classifiers SVM, LR, RF, and XGBoost on the test set.

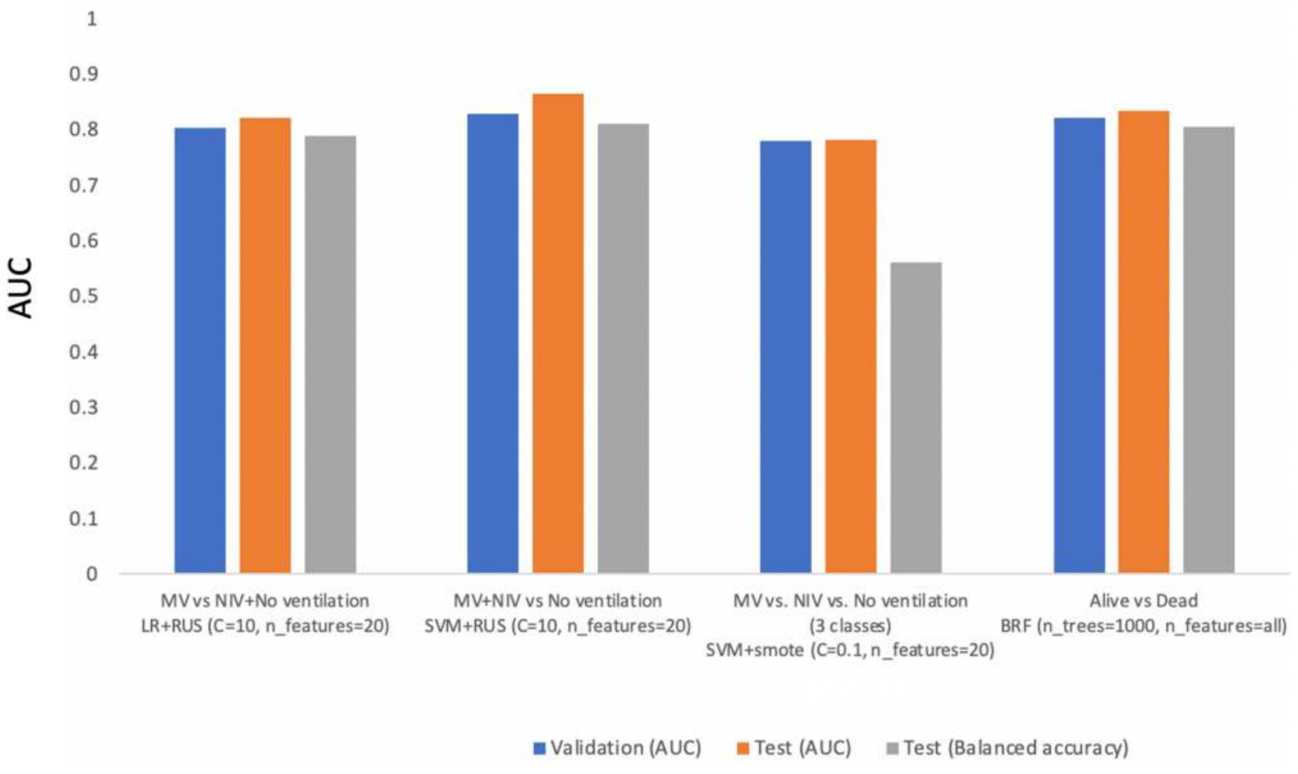

**Figure 5** Best performing models on test and validation sets for mortality and ventilation support need. MV represents mechanical ventilation prediction. MV+NIV represents a model that predicts the need for either mechanical ventilation or non-invasive ventilation versus no ventilation. The C hyperparameter for SVM represents the tradeoff between maximizing the margin minimizing the error, and the C for logistic regression represents the value assigned to control the regularization strength.

Figure 6 shows the confusion matrix for the best performing models in the test set for the four prediction tasks tackled in this study. The top-performing model for MV yielded a specificity of 0.80 and a sensitivity of 0.78 with logistic regression after applying random undersampling to the majority in the train dataset. The best model for ventilation requirement achieved a specificity of 0.71 and a sensitivity of 0.91 using a linear SVM classifier and random undersampling. The balanced random forest model for mortality prediction has a specificity of 0.61 and a sensitivity of 1.

Table 3 shows the top features ReliefF weights in the train set, where CXR zone 11 and 12 radiological semantic features are ranked top-1 and top-2 in with the weight of 0.23 and 0.19 for MV, 0.21, and 0.14 for MV+NIV, 0.20 and 0.17 for MV, NIV, and no ventilation, and a weight of 0.24 and 0.24 in the mortality train data set.

Interobserver variability of severity scores among radiologists for the randomly selected CXR image annotations as Kappa values using Fleiss Kappa test is reported in Table 5. Landis and Koch have suggested the interpretations of the strength of rater agreement for Kappa statistic to be divided into six categories (poor: <0.00, slight: 0.00–0.20, fair: 0.21–0.40, moderate: 0.41–0.60, substantial: 0.61–0.80, and almost perfect: 0.81–1.00).[39] Lung zone 3 (see Figure 3 for lung segmentation) has the best inter-rater agreement (kappa = 0.71), which is equivalent to substantial agreement. The lowest inter-rater agreement attained was in region 7 (kappa = 0.21), which corresponds to a fair agreement.

Figure 7 shows the severity scores distribution for the 12 lung zones from baseline CXR as a heatmap of patients that underwent MV or those who did not need any ventilation, as well as patients who were discharged alive or died during hospitalization.

## Discussions

In this work, we assessed the predictive ability of combined CBC, age, gender, comorbidity, and CXRs severity annotated data. This data was collected at the time of COVID-19 diagnosis in our hospital. The proposed models can be used to predict mechanical ventilation (MV), mechanical and non-invasive ventilation, and mortality early on to prioritize patients and manage the allocation of hospital resources. We have found that balancing the train set, classifiers consistently yielded better performance than using the original skewed data (results not

**Table 4** Performance at Optimal Operating Point of Four ML Algorithms and Different Balancing Techniques for Prediction of Ventilation Requirement and Mortality

| Features Set | Model | MV[a] | MV+NIV[b] | MV, NIV, None[c] | Mortality |
|---|---|---|---|---|---|
| **(Balanced Accuracy, AUC)** | | | | | |
| | SVM+ADASYN | (0.70, 0.71) | (0.63, 0.71) | (0.42, 0.65) | (0.67, 0.78) |
| | SVM+SMOTE | (0.64, 0.72) | (0.62, 0.71) | (0.44, 0.65) | (0.69, 0.78) |
| | SVM+RUS | (0.66, 0.71) | (0.62, 0.69) | (0.41, 0.66) | (0.66, 0.78) |
| | LR+ADASYN | (0.66, 0.72) | (0.66, 0.71) | (0.40, 0.66) | (0.69, 0.78) |
| | LR+SMOTE | (0.64, 0.72) | (0.66, 0.71) | (0.39, 0.66) | (0.68, 0.78) |
| **Age + gender** | LR+RUS | (0.65, 0.71) | (0.65, 0.69) | (0.41, 0.67) | (0.64, 0.78) |
| | XGB+ADASYN | (0.65, 0.70) | (0.61, 0.67) | (0.41, 0.63) | (0.52, 0.64) |
| | XGB+SMOTE | (0.65, 0.69) | (0.60, 0.68) | (0.46, 0.67) | (0.66, 0.70) |
| | XGB+RUS | (0.61, 0.68) | (0.62, 0.68) | (0.49, 0.67) | (0.63, 0.66) |
| | RF+ADASYN | (0.61, 0.68) | (0.57, 0.61) | (0.42, 0.60) | (0.58, 0.63) |
| | RF+SMOTE | (0.65, 0.68) | (0.58, 0.62) | (0.47, 0.64) | (0.60, 0.65) |
| | BRF | (0.63, 0.70) | (0.60, 0.65) | (0.47, 0.65) | (0.69, 0.73) |
| | SVM+ADASYN | (0.67, 0.71) | (0.75, 0.82) | (0.50, 0.74) | (0.49, 0.72) |
| | SVM+SMOTE | (0.70, 0.74) | (0.75, 0.81) | (0.56, 0.75) | (0.51, 0.75) |
| | SVM+RUS | (0.64, 0.70) | (0.76, 0.82) | (0.47, 0.74) | (0.70, 0.73) |
| | LR+ADASYN | (0.70, 0.74) | (0.79, 0.85) | (0.48, 0.76) | (0.53, 0.73) |
| | LR+SMOTE | (0.71, 0.74) | (0.79, 0.85) | (0.56, 0.77) | (0.52, 0.73) |
| | LR+RUS | (0.70, 0.72) | (0.79, 0.84) | (0.53, 0.76) | (0.67, 0.74) |
| **X-ray** | XGB+ADASYN | (0.57, 0.69) | (0.77, 0.82) | (0.52, 0.76) | (0.58, 0.63) |
| | XGB+SMOTE | (0.68, 0.73) | (0.77, 0.82) | (0.51, 0.76) | (0.58, 0.65) |
| | XGB+RUS | (0.71, 0.72) | (0.79, 0.82) | (0.50, 0.73) | (0.54, 0.75) |
| | RF+ADASYN | (0.63, 0.56) | (0.74, 0.70) | (0.47, 0.62) | (0.53, 0.51) |
| | RF+SMOTE | (0.52, 0.63) | (0.71, 0.75) | (0.47, 0.62) | (0.43, 0.49) |
| | BRF | (0.70, 0.74) | (0.79, 0.82) | (0.51, 0.75) | (0.71, 0.72) |
| | SVM+ADASYN | (0.64, 0.72) | (0.61, 0.68) | (0.43, 0.62) | (0.62, 0.66) |
| | SVM+SMOTE | (0.66, 0.72) | (0.62, 0.69) | (0.42, 0.62) | (0.61, 0.65) |
| | SVM+RUS | (0.64, 0.73) | (0.64, 0.68) | (0.40, 0.64) | (0.60, 0.61) |
| | LR+ADASYN | (0.65, 0.71) | (0.61, 0.68) | (0.42, 0.62) | (0.59, 0.66) |
| | LR+SMOTE | (0.63, 0.71) | (0.62, 0.68) | (0.42, 0.61) | (0.53, 0.66) |
| | LR+RUS | (0.66, 0.72) | (0.66, 0.68) | (0.43, 0.61) | (0.59, 0.62) |
| **CBC** | XGB+ADASYN | (0.61, 0.67) | (0.57, 0.65) | (0.42, 0.56) | (0.54, 0.60) |
| | XGB+SMOTE | (0.64, 0.70) | (0.61, 0.65) | (0.36, 0.58) | (0.62, 0.62) |
| | XGB+RUS | (0.65, 0.73) | (0.63, 0.69) | (0.46, 0.59) | (0.62, 0.63) |
| | RF+ADASYN | (0.59, 0.63) | (0.59, 0.65) | (0.37, 0.57) | (0.58, 0.62) |
| | RF+SMOTE | (0.60, 0.66) | (0.60, 0.64) | (0.37, 0.57) | (0.57, 0.63) |
| | BRF | (0.64, 0.71) | (0.59, 0.67) | (0.42, 0.61) | (0.60, 0.66) |
| | SVM+ADASYN | (0.50, 0.59) | (0.49, 0.64) | (0.40, 0.51) | (0.70, 0.77) |
| | SVM+SMOTE | (0.50, 0.59) | (0.56, 0.60) | (0.37, 0.57) | (0.70, 0.66) |
| | SVM+RUS | (0.50, 0.56) | (0.56, 0.49) | (0.37, 0.52) | (0.68, 0.77) |
| | LR+ADASYN | (0.53, 0.57) | (0.52, 0.63) | (0.39, 0.55) | (0.72, 0.80) |
| | LR+SMOTE | (0.52, 0.59) | (0.55, 0.61) | (0.42, 0.58) | (0.70, 0.78) |

(*Continued*)

**Table 4** (Continued).

| Features Set | Model | MV[a] | MV+NIV[b] | MV, NIV, None[c] | Mortality |
|---|---|---|---|---|---|
| **(Balanced Accuracy, AUC)** | | | | | |
| **Comorbidity** | LR+RUS | (0.51, 0.57) | (0.55, 0.58) | (0.37, 0.59) | (0.70, 0.77) |
| | XGB+ADASYN | (0.54, 0.61) | (0.61, 0.60) | (0.41, 0.55) | (0.71, 0.76) |
| | XGB+SMOTE | (0.53, 0.60) | (0.55, 0.58) | (0.39, 0.58) | (0.71, 0.75) |
| | XGB+RUS | (0.54, 0.63) | (0.56, 0.57) | (0.40, 0.58) | (0.67, 0.75) |
| | RF+ADASYN | (0.54, 0.61) | (0.56, 0.61) | (0.41, 0.55) | (0.68, 0.74) |
| | RF+SMOTE | (0.55, 0.60) | (0.55, 0.57) | (0.38, 0.56) | (0.71, 0.74) |
| | BRF | (0.58, 0.63) | (0.50, 0.57) | (0.39, 0.56) | (0.70, 0.79) |
| | SVM+ADASYN | (0.73, 0.81) | (0.78, 0.86) | (0.53, 0.79) | (0.77, 0.83) |
| | SVM+SMOTE | (0.73, 0.81) | (0.80, 0.86) | (0.56, 0.78) | (0.78, 0.83) |
| | SVM+RUS | (0.74, 0.81) | (0.81, 0.87) | (0.53, 0.77) | (0.74, 0.82) |
| | LR+ADASYN | (0.73, 0.82) | (0.80, 0.86) | (0.53, 0.79) | (0.78, 0.85) |
| | LR+SMOTE | (0.74, 0.81) | (0.81, 0.87) | (0.53, 0.79) | (0.77, 0.84) |
| **Combined** | LR+RUS | (0.79, 0.82) | (0.80, 0.86) | (0.56, 0.76) | (0.75, 0.82) |
| | XGB+ADASYN | (0.74, 0.82) | (0.78, 0.84) | (0.43, 0.78) | (0.72, 0.81) |
| | XGB+SMOTE | (0.76, 0.83) | (0.75, 0.82) | (0.46, 0.67) | (0.73, 0.82) |
| | XGB+RUS | (0.72, 0.72) | (0.78, 0.83) | (0.53, 0.74) | (0.73, 0.78) |
| | RF+ADASYN | (0.76, 0.80) | (0.76, 0.82) | (0.45, 0.78) | (0.68, 0.77) |
| | RF+SMOTE | (0.75, 0.80) | (0.78, 0.84) | (0.45, 0.78) | (0.71, 0.80) |
| | BRF | (0.77, 0.80) | (0.78, 0.86) | (0.53, 0.78) | (0.80, 0.83) |

**Notes**: [a]Mechanical ventilation versus non-invasive ventilation or no ventilation, [b]Non-invasive ventilation or mechanical ventilation versus no ventilation, [c]Three class classification: mechanical ventilation versus non-invasive ventilation versus no ventilation.

shown). Random downsampling for creating a balanced train set with logistic regression and random forest produced a better model compared to SMOTE or ADASYN for mechanical ventilation and MV+NIV and mortality prediction. However, SMOTE with linear SVM yielded the better model for the three-class classification problem (MV, NIV, no ventilation). We have compared the performance of random forest and XGBoost, which are capable of capturing nonlinear relationships between features and the target, with logistic regression and linear SVM since these are widely used classifiers with good generalization. In all of the prediction tasks attempted in the current study, combined clinical, laboratory, and radiological data with ReliefF feature selection consistently performed better than each data set individually. The model LR with RUS and top 20 selected features for the combined data sources achieved 0.82 AUC in predicting mechanical ventilation requirement in the test set. When we merged both MV and NIV patients into a positive class and assigned the no ventilation to the negative class, we observed that linear SVM with RUS outperformed other models with an AUC of 0.86 in top 20 ranked features. The top achieving model in the test set for mortality prediction yielded an AUC score of 0.83 and a balanced accuracy of 0.80, using all features from combined data and balanced random forest. The dramatic difference in performance between fitting a model with and without data balancing is perhaps due to severe data skewness (see Tables 1 and 2). Comorbidity features alone, which are a binary feature for diseases listed in Table 1, yielded a predictive ability of 0.78 AUC for mortality prediction. However, as shown in Table 4, building a model with CXRs severity scores to predict intubation and invasive ventilation outperforms models that use age, gender, medical history, or laboratory data alone, which is consistent with what other studies have reported.[40] In accordance with earlier reports, males (57% in our cohort) were infected more than females,[7,41] and the number of males is higher in both mortality and severe disease type (MV: 76%, and mortality: 73% in our cohort) than females.[42] Other studies investigate the use of complete blood count parameters to stratify patients based on disease severity.[43,44] ACE2 protein, which is a receptor of COVID-19, is expressed in lymphocytes. It is worth mentioning that numerous studies have concluded that lymphopenia is associated with the severity of COVID-19.[5,45] Similarly, we have observed that WBC is

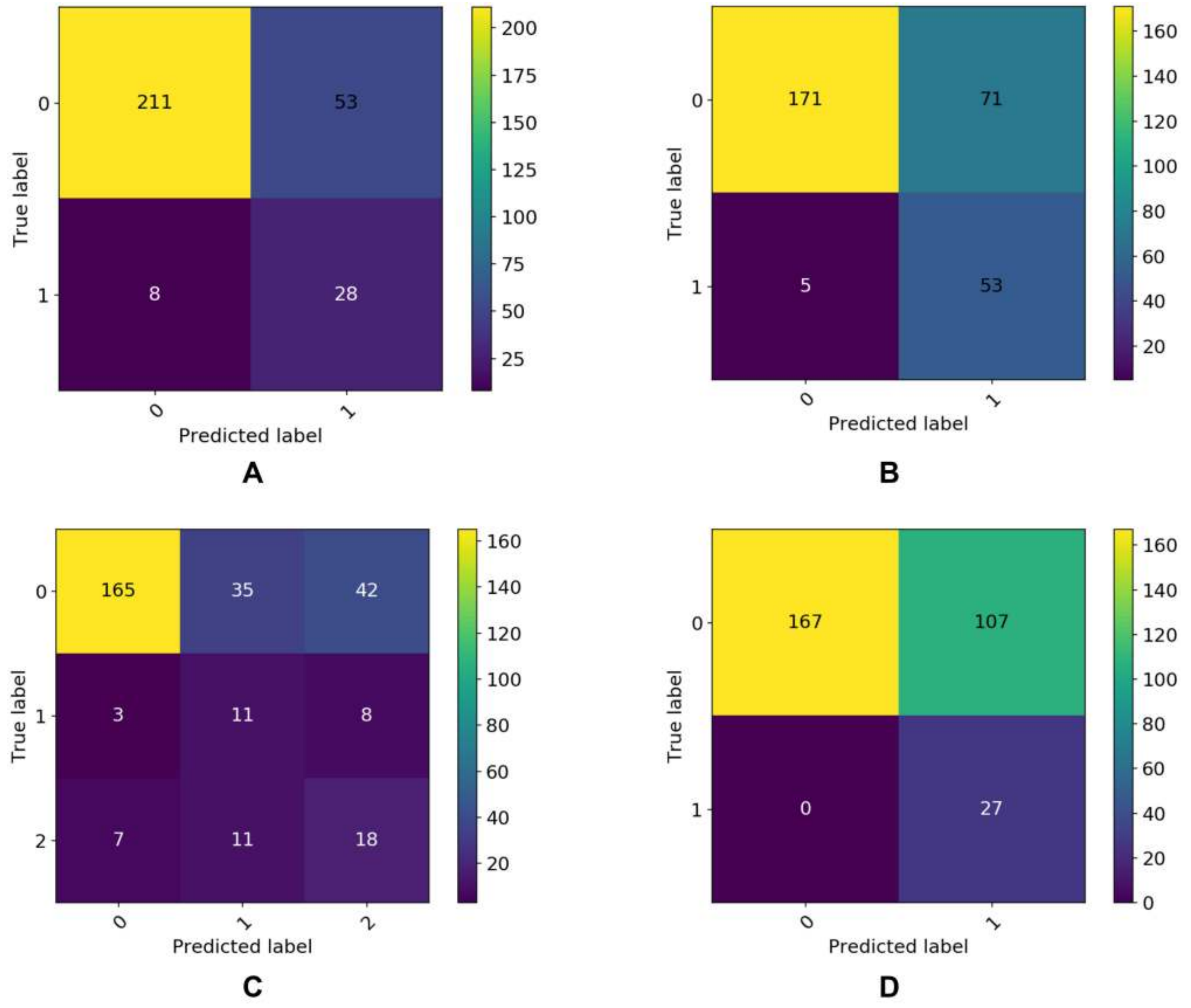

**Figure 6** (**A**) mechanical ventilation prediction confusion matrix in the test set with linear SVM and SMOTE; (**B**) ventilation requirement (MV or NIV) prediction in the test set with BRF; (**C**) MV vs NIV vs no ventilation in the test set prediction with BRF; (**D**) mortality prediction in the test set with BRF.

significantly different (P = 4.8e-13) between patients who required ventilation and those who did not require ventilation during hospitalization. There is growing evidence that older COVID-19 patients tend to be more critically ill than younger patients.[46,47] This is also consistent with what we have found in this study, where increased age is significantly associated with disease severity and mortality (see Tables 1 and 2). We have found that White Blood Cell (WBC) count was higher in patients with mechanical ventilation and non-invasive ventilation compared to patients that required no ventilation. For mortality, we have found that WBC count was significantly higher (*p*-value = 0.0004) in deceased patients compared to patients that were discharge alive (see Table 2). Studies have shown that patients who died from COVID-19 had significantly higher levels of neutrophils (%) compared to patients who survived, and the patients who died had significantly lower levels of lymphocyte (%) than those who survived.[48–51]

The most affected zones with GGO and consolidation in mechanical ventilation and mortality groups are the lower lobes. In addition, semantic radiological features obtained from zone 11, and 12 (see Figure 3) ranked top-1 and top-2 based on ReliefF weight as reported in Table 3 for MV and MV+NIV and mortality prediction task. This indicates that the radiological semantic features (GGO) of lower lobe is an important biomarker of severe COVID-19 disease type. Figure 7

**Table 5** Fleiss Kappa Scores of Four Raters' Agreement on 110 Randomly Selected Patients for the 12 CXR Lung Regions Semantic Features

| CXR Lung Zone | Kappa | Z-Score | P-value |
|---|---|---|---|
| 1 | 0.552 | 14.6 | 0 |
| 2 | 0.451 | 11.9 | 0 |
| 3 | 0.713 | 20.3 | 0 |
| 4 | 0.698 | 20.2 | 0 |
| 5 | 0.526 | 16.3 | 0 |
| 6 | 0.54 | 16.6 | 0 |
| 7 | 0.216 | 6.16 | 7.06e–10 |
| 8 | 0.608 | 16.5 | 0 |
| 9 | 0.493 | 13.6 | 0 |
| 10 | 0.66 | 18.5 | 0 |
| 11 | 0.416 | 13.3 | 0 |
| 12 | 0.536 | 17 | 0 |

shows a clear difference of severity distributions between MV and no ventilation groups, where severe opacities are more pronounced in patients with poor outcome. Our finding confirms the conclusions of many studies that COVID-19 pneumonia in CXR images predominantly appears as airspace opacity and often in bilateral, peripheral, and lower zone.[10,52,53]

We focused on baseline variables that are readily available and routinely collected at the disease diagnosis time. For example, several studies have used CT scans to assess the severity of COVID-19 disease.[54,55] Although CT scan shows the COVID-19 pneumonia presentations of ground-glass opacity and pulmonary consolidation and has a better sensitivity than CXR;[56,57] however, it is not the standard of care for COVID-19 diagnosis in many countries.[57,58] Among other reasons, it is preferable to avoid CT scan as a first-line imaging method to limit unnecessary extra radiation exposure, prevent CT device shortages, and minimize cross-infection by using portable X-ray machines.[58]

The current study has some limitations. First of all, this is a single-center study; so, the predictive model generalizability to other hospitals has not been tested. We plan to extend this work to include patients from other healthcare providers. Another limitation is that the class distribution is severely imbalanced, and therefore, having roughly balanced data could enhance our model's predictive ability. In the future, we plan to use stratified random sampling with a larger sample size. Furthermore, we have not tested the predictive ability of features such as vitals such as temperature, respiratory rate and SPO2, which have been found to be effective in

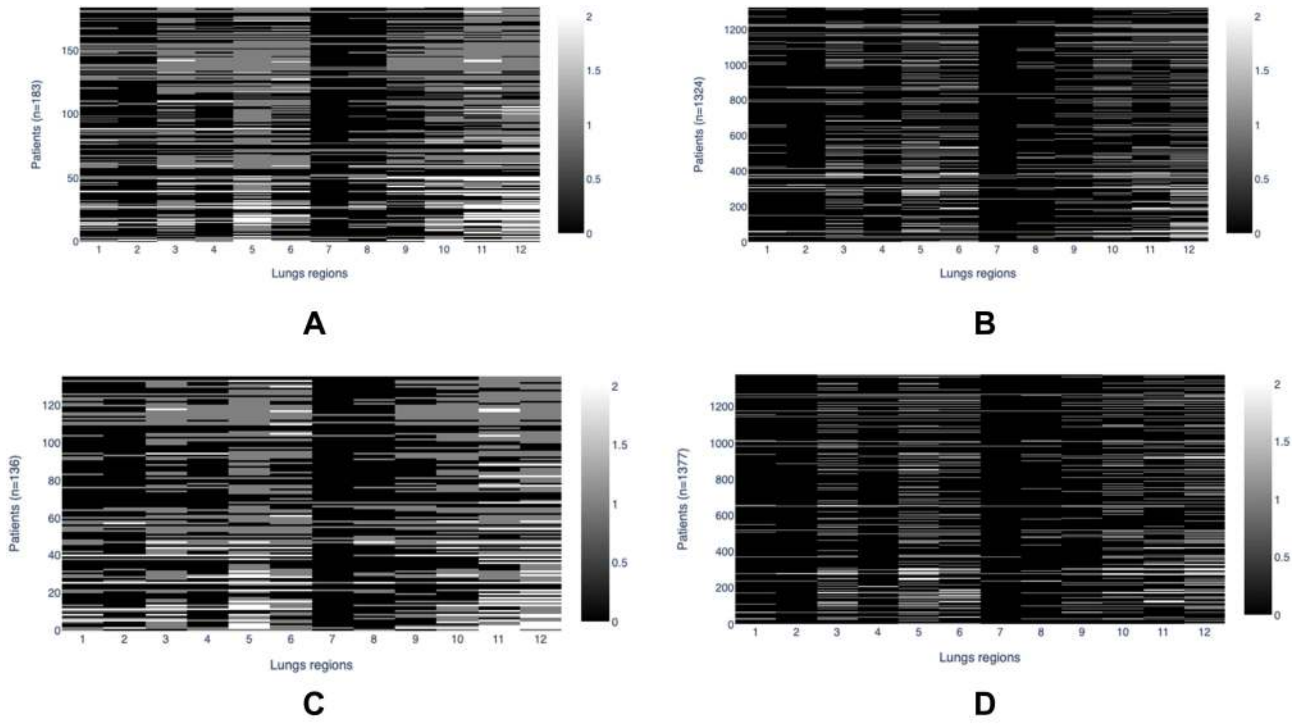

**Figure 7** Heatmap of severity scores distribution from the baseline CXR images in the 12 lung region: (**A**) mechanical ventilation; (**B**) no ventilation; (**C**) dead; (**D**) discharged alive. White color represents severe GGO/consolidation, gray color represents mild/moderate manifestations, and black color represents no presentations of abnormality.

predicting COVID-19 prognosis.[59–63] Lastly, the CXR severity scores depend on radiologists' observations, which are prone to interrater variability. Thus, extracting relevant features directly from CXR images using deep learning techniques may lead to a more robust model with an automated end-to-end prediction system.

## Conclusions

In summary, we have developed models to predict COVID-19 patients' invasive mechanical ventilation requirement and mortality. In conclusion, these models scored an AUC of 0.82 and 0.83, respectively, in an independent test set from top 20 selected features with ReliefF from combined baseline radiological, laboratory, and clinical features. For this independent test set, all the data was collected retrospectively at King Abdulaziz Medical City in Riyadh. We selected best models using grid search on the validation set for ventilation requirement and mortality prediction from a combination of four machine learning classifiers (linear support vector machines, logistic regression, random forest, and XGBoost) and three data balancing techniques (random undersampling, SMOTE, and ADASYN) and top ranked features based on ReliefF algorithm weights. The proposed tool provides insights into the efficient planning of hospital resources and patients' prioritization in the current COVID-19 pandemic crisis.

## Data Sharing Statement

We made the datasets generated during and/or analyzed in the current study publicly available in the Mendeley Data repository, "KAMC_COVID-19", Mendeley Data, V2, doi: 10.17632/r6t9tmzzmz.2.

## Ethical Consideration

This retrospective study was approved on July 6, 2020, by the ethical committee (IRB approval number: RC/357/R) at King Abdulaziz Medical City in Riyadh, Saudi Ministry of National Guard Health Affairs. The data used was anonymized, and all the personal information was de-identified to preserve the privacy of the human subjects. Additionally, due to the minimal risk to the patients, the written informed consent was waived. This study was conducted in accordance with the ethical principles outlined in the Declaration of Helsinki.

## Acknowledgments

We thank the KAIMRC Research Data Management group for their technical help.

## Author Contributions

All authors made substantial contributions to conception and design, acquisition of data, or analysis and interpretation of data; took part in drafting the article or revising it critically for important intellectual content; agreed to submit to the current journal; gave final approval of the version to be published; and agreed to be accountable for all aspects of the work.

## Funding

This research received no external funding.

## Disclosure

The authors declare no conflicts of interest for this work.

## References

1. Alhazzani W, Møller MH, Arabi YM, et al. Surviving sepsis campaign: guidelines on the management of critically ill adults with Coronavirus disease 2019 (COVID-19). *Intensive Care Med*. 2020;46(5):854–887.
2. Guan W-J, Ni Z-Y, Hu Y, et al. Clinical characteristics of Coronavirus disease 2019 in China. *N Engl J Med*. 2020;382(18):1708–1720. doi:10.1056/NEJMoa2002032
3. Richardson S, Hirsch JS, Narasimhan M, et al. Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York City area. *JAMA*. 2020;323 (20):2052–2059. doi:10.1001/jama.2020.6775
4. Yadaw AS, Li Y-C, Bose S, Iyengar R, Bunyavanich S, Pandey G. Clinical features of COVID-19 mortality: development and validation of a clinical prediction model. *Lancet Digit Health*. 2020;2(10):e516–e525. doi:10.1016/S2589-7500(20)30217-X
5. Tan L, Wang Q, Zhang D, et al. Lymphopenia predicts disease severity of COVID-19: a descriptive and predictive study. *Signal Transduct Target Ther*. 2020;5(1):33. doi:10.1038/s41392-020-0148-4
6. Cao W, Liu X, Bai T, et al. High-dose intravenous immunoglobulin as a therapeutic option for deteriorating patients with Coronavirus disease 2019. *Open Forum Infect Dis*. 2020;7(3):ofaa102. doi:10.1093/ofid/ofaa102
7. Zhou F, Yu T, Du R, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet*. 2020;395(10229):1054–1062. doi:10.1016/S0140-6736(20)30566-3
8. Qin C, Zhou L, Hu Z, et al. Dysregulation of immune response in patients with Coronavirus 2019 (COVID-19) in Wuhan, China. *Clin Infect Dis*. 2020;71(15):762–768. doi:10.1093/cid/ciaa248
9. Li K, Wu J, Wu F, et al. The clinical and chest CT features associated with severe and critical COVID-19 pneumonia. *Invest Radiol*. 2020;55(6):327–331. doi:10.1097/RLI.0000000000000672
10. Wong HYF, Lam HYS, Fong AH, et al. Frequency and distribution of chest radiographic findings in patients positive for COVID-19. *Radiology*. 2020;296(2):E72–E78. doi:10.1148/radiol.2020201160
11. Xu Z, Shi L, Wang Y, et al. Pathological findings of COVID-19 associated with acute respiratory distress syndrome. *Lancet Respir Med*. 2020;8(4):420–422. doi:10.1016/S2213-2600(20)30076-X
12. Toussie D, Voutsinas N, Finkelstein M, et al. Clinical and chest radiography features determine patient outcomes in young and middle-aged adults with COVID-19. *Radiology*. 2020;297(1):E197–E206. doi:10.1148/radiol.2020201754

13. Shi W, Peng X, Liu T, et al. A deep learning-based quantitative computed tomography model for predicting the severity of COVID-19: a retrospective study of 196 patients. *Ann Transl Med*. 2021;9(3):216. doi:10.21037/atm-20-2464
14. Gong J, Ou J, Qiu X, et al. *Multicenter Development and Validation of a Novel Risk Nomogram for Early Prediction of Severe 2019-Novel Coronavirus Pneumonia*. Rochester, NY: Social Science Research Network; 2020.
15. Cheng FY, Joshi H, Tandon P, et al. Using machine learning to predict ICU transfer in hospitalized COVID-19 patients. *J Clin Med*. 2020;9(6):1668. doi:10.3390/jcm9061668
16. Ryan L, Lam C, Mataraso S, et al. Mortality prediction model for the triage of COVID-19, pneumonia, and mechanically ventilated ICU patients: a retrospective study. *Ann Med Surg*. 2020;59:207–216. doi:10.1016/j.amsu.2020.09.044
17. Parchure P, Joshi H, Dharmarajan K, et al. Development and validation of a machine learning-based prediction model for near-term in-hospital mortality among patients with COVID-19. *BMJ Support Palliat Care*. 2020: Epub. doi:10.1136/bmjspcare-2020-002602
18. Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ*. 2020;369:m1328. doi:10.1136/bmj.m1328
19. Wu G, Yang P, Xie Y, et al. Development of a clinical decision support system for severity risk prediction and triage of COVID-19 patients at hospital admission: an international multicentre study. *Eur Respir J*. 2020;56(2):2001104. doi:10.1183/13993003.01104-2020
20. Rubin GD, Ryerson CJ, Haramati LB, et al. The role of chest imaging in patient management during the COVID-19 pandemic: a multinational consensus statement from the Fleischner society. *Chest*. 2020;158(1):106–116. doi:10.1016/j.chest.2020.04.003
21. ACR recommendations for the use of chest radiography and Computed Tomography (CT) for suspected COVID-19 infection; 2020. Available from: https://www.acr.org/Advocacy-and-Economics/ACR-Position-Statements/Recommendations-for-Chest-Radiography-and-CT-for-Suspected-COVID19-Infection. Accessed July 19, 2021.
22. Apostolopoulos ID, Mpesiana TA. Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks. *Phys Eng Sci Med*. 2020;43(2):635–640. doi:10.1007/s13246-020-00865-4
23. Sethy PK, Behera SK, Ratha PK, Biswas P. Detection of coronavirus disease (COVID-19) based on deep features and support vector machine. *Int J Math Eng Manag Sci*. 2020;5(4):643–651.
24. Yoo SH, Geng H, Chiu TL, et al. Deep learning-based decision-tree classifier for COVID-19 diagnosis from chest X-ray imaging. *Front Med (Lausanne)*. 2020;7:427. doi:10.3389/fmed.2020.00427
25. Singh D, Kumar V, Vaishali KM, Kaur M. Classification of COVID-19 patients from chest CT images using multi-objective differential evolution-based convolutional neural networks. *Eur J Clin Microbiol Infect Dis*. 2020;39(7):1379–1389. doi:10.1007/s10096-020-03901-z
26. Ahuja S, Panigrahi BK, Dey N, Rajinikanth V, Gandhi TK. Deep transfer learning-based automated detection of COVID-19 from lung CT scan slices. *Appl Intell*. 2020;51:1–15.
27. Mechanical ventilation protocol for COVID-19; 2020. Available from: https://www.moh.gov.sa/Ministry/MediaCenter/Publications/Documents/Mechanical-Ventilition.pdf. Accessed July 04, 2021.
28. COVID-19 Treatment Guidelines Panel. Coronavirus disease 2019 (COVID-19) treatment guidelines; 2021. National Institutes of Health. Available from: https://www.covid19treatmentguidelines.nih.gov/management/critical-care/oxygenation-and-ventilation/. Accessed July 04, 2021.
29. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *J Mach Learn Res*. 2011;12:2825–2830.
30. Kononenko I. Estimating attributes: analysis and extensions of RELIEF. In: Bergadano F, De Raedt L, editors. *Machine Learning: ECML-94*. Berlin, Heidelberg: Springer Berlin Heidelberg; 1994:171–182.
31. Robnik-šikonja M, Kononenko I. Theoretical and empirical analysis of ReliefF and RReliefF. *Mach Learn*. 2003;53(1):23–69. doi:10.1023/A:1025667309714
32. Liu N, Li X, Qi E, Xu M, Li L, Gao B. A novel ensemble learning paradigm for medical diagnosis with imbalanced data. *IEEE Access*. 2020;8:171263–171280. doi:10.1109/ACCESS.2020.3014362
33. López V, Fernández A, García S, Palade V, Herrera F. An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics. *Inf Sci (NY)*. 2013;250:113–141. doi:10.1016/j.ins.2013.07.007
34. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Int Res*. 2002;16(1):321–357.
35. Lemaître G, Nogueira F, Aridas CK. Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. *J Mach Learn Res*. 2017;18(1):559–563.
36. Haibo H, Yang B, Garcia EA, Shutao L. ADASYN: adaptive synthetic sampling approach for imbalanced learning. 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence); 2008; 1322–1328.
37. Cortes C, Vapnik V. Support-vector networks. *Mach Learn*. 1995;20(3):273–297. doi:10.1007/BF00994018
38. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5–32. doi:10.1023/A:1010933404324
39. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159–174. doi:10.2307/2529310
40. Hui TCH, Khoo HW, Young BE, et al. Clinical utility of chest radiography for severe COVID-19. *Quant Imaging Med Surg*. 2020;10(7):1540–1550. doi:10.21037/qims-20-642
41. Alsofayan YM, Althunayyan SM, Khan AA, Hakawi AM, Assiri AM. Clinical characteristics of COVID-19 in Saudi Arabia: a national retrospective study. *J Infect Public Health*. 2020;13(7):920–925. doi:10.1016/j.jiph.2020.05.026
42. Pradhan A, Olsson P-E. Sex differences in severity and mortality from COVID-19: are males more vulnerable? *Biol Sex Differ*. 2020;11(1):53. doi:10.1186/s13293-020-00330-7
43. Zeng F, Li L, Zeng J, et al. Can we predict the severity of coronavirus disease 2019 with a routine blood test? *Pol Arch Intern Med*. 2020;130(5):400–406.
44. Foy BH, Carlson JCT, Reinertsen E, et al. Association of red blood cell distribution width with mortality risk in hospitalized adults with SARS-CoV-2 infection. *JAMA Netw Open*. 2020;3(9):e2022058. doi:10.1001/jamanetworkopen.2020.22058
45. Henry B, Cheruiyot I, Vikse J, et al. Lymphopenia and neutrophilia at admission predicts severity and mortality in patients with COVID-19: a meta-analysis. *Acta Biomed*. 2020;91:e2020008.
46. Liu Y, Mao B, Liang S, et al. Association between ages and clinical characteristics and outcomes of Coronavirus disease 2019. *Eur Respir J*. 2020;55:2001112. doi:10.1183/13993003.01112-2020
47. Romero Starke K, Petereit-Haack G, Schubert M, et al. The age-related risk of severe outcomes due to COVID-19 infection: a rapid review, meta-analysis, and meta-regression. *Int J Environ Res Public Health*. 2020;17(16):5974. doi:10.3390/ijerph17165974
48. Zuo Y, Zuo M, Yalavarthi S, et al. Neutrophil extracellular traps and thrombosis in COVID-19. *J Thromb Thrombolysis*. 2021;51(2):446–453. doi:10.1007/s11239-020-02324-z
49. Karthikeyan A, Garg A, Vinod PK, Priyakumar UD. Machine learning based clinical decision support system for early COVID-19 mortality prediction. *Front Public Health*. 2021;9:475. doi:10.3389/fpubh.2021.626697

50. Lagunas-Rangel FA. Neutrophil-to-lymphocyte ratio and lymphocyte-to-C-reactive protein ratio in patients with severe coronavirus disease 2019 (COVID-19): a meta-analysis. *J Med Virol*. 2020;92(10):1733–1734. doi:10.1002/jmv.25819
51. Chowdhury MEH, Rahman T, Khandakar A, et al. An early warning tool for predicting mortality risk of COVID-19 patients using machine learning. *Cognit Comput*. 2021; Epub. doi:10.1007/s12559-020-09812-7
52. Rousan LA, Elobeid E, Karrar M, Khader Y. Chest x-ray findings and temporal lung changes in patients with COVID-19 pneumonia. *BMC Pulm Med*. 2020;20(1):245. doi:10.1186/s12890-020-01286-5
53. Yoon SH, Lee KH, Kim JY, et al. Chest radiographic and CT findings of the 2019 Novel Coronavirus disease (COVID-19): analysis of nine patients treated in Korea. *Korean J Radiol*. 2020;21(4):494–500. doi:10.3348/kjr.2020.0132
54. Yu Z, Li X, Sun H, et al. Rapid identification of COVID-19 severity in CT scans through classification of deep features. *Biomed Eng Online*. 2020;19(1):63. doi:10.1186/s12938-020-00807-x
55. Homayounieh F, Ebrahimian S, Babaei R, et al. CT radiomics, radiologists and clinical information in predicting outcome of patients with COVID-19 pneumonia. *Radiology*. 2020;2(4):e200322.
56. Stephanie S, Shum T, Cleveland H, et al. Determinants of chest X-ray sensitivity for COVID- 19: a multi-institutional study in the United States. *Radiology*. 2020;2(5):e200337.
57. Jacobi A, Chung M, Bernheim A, Eber C. Portable chest X-ray in coronavirus disease-19 (COVID-19): a pictorial review. *Clin Imaging*. 2020;64:35–42. doi:10.1016/j.clinimag.2020.04.001
58. Cozzi D, Albanesi M, Cavigli E, et al. Chest X-ray in new Coronavirus disease 2019 (COVID-19) infection: findings and correlation with clinical outcome. *Radiol Med*. 2020;125(8):730–737. doi:10.1007/s11547-020-01232-9
59. Xie J, Hungerford D, Chen H, et al. Development and external validation of a prognostic multivariable model on admission for hospitalized patients with COVID-19. *medRxiv*. 2020; preprint. doi:10.1101/2020.03.28.20045997
60. Lu X, Jiang L, Chen T, et al. Continuously available ratio of SpO2/FiO2 serves as a noninvasive prognostic marker for intensive care patients with COVID-19. *Respir Res*. 2020;21(1):194. doi:10.1186/s12931-020-01455-4
61. Wang K, Zuo P, Liu Y, et al. Clinical and laboratory predictors of in-hospital mortality in patients with Coronavirus disease-2019: a cohort study in Wuhan, China. *Clin Infect Dis*. 2020;71(16):2079–2088. doi:10.1093/cid/ciaa538
62. Marcos M, Belhassen-García M, Sánchez-Puente A, et al. Development of a severity of disease score and classification model by machine learning for hospitalized COVID-19 patients. *PLoS One*. 2021;16(4):e0240200. doi:10.1371/journal.pone.0240200
63. Bolourani S, Brenner M, Wang P, et al. A machine learning prediction model of respiratory failure within 48 hours of patient admission for COVID-19: model development and validation. *J Med Internet Res*. 2021;23(2):e24246.