

An ICU Admission Predictive Model for COVID-19 Patients in Saudi Arabia

Hamza Ghandorh¹, Muhammad Zubair Khan²
College of Computer Science and Engineering,
Taibah University
Medina 42353, Saudi Arabia

Raed Alsufyani³
Research and Knowledge Director
Aathar Alkhair Holding Company
Jeddah Governorate, Saudi Arabia

Mehshan Khan⁴
Department of EE
HITEC University
Museum Road, Taxila, Pakistan

Yousef M. Alsofayan⁵, Anas A. Khan⁶, Ahmed A. Alahmari⁷
Global Center for Mass Gatherings Medicine Ministry of Health,
P.O. Box 1146 Riyadh, 11176, Saudi Arabia

Abstract—Globally, COVID-19 already emerged in around 170 million confirmed cases of infected people and, as of May 31, 2021, affected more than 3.54 million deaths. This pandemic has given rise to numerous public health and socioeconomic issues, emphasizing the significance of unraveling the epidemic's history and forecasting the disease's potential dynamics. A variety of mathematical models have been proposed to obtain a deeper understanding of disease transmission mechanisms. Machine Learning (ML) models have been used in the last decade to identify patterns and enhance prediction efficiency in healthcare applications. This paper proposes a model to predict COVID-19 patients admission to the intensive care unit (ICU). The model is built upon robust known classification algorithms, including classic Machine Learning Classifiers (MLCs), an Artificial Neural Network (ANN) and ensemble learning. This model's strength in predicting COVID-19 infected patients is shown by performance analysis of various MLCs and error metrics. Among other used ML models, the ANN model resulted in the highest accuracy, 97.9% over other models. Mean Squared Error showed that the ANN method had the lowest error (0.0809). In conclusion, this paper could be beneficial to ICU staff to predict ICU admission based on COVID-19 patients' clinical characteristics.

Keywords—Covid-19; ANN; ensemble learning method; prediction; ICU admission; Saudi Arabia

I. INTRODUCTION

The 2019 coronavirus disease (COVID-19) emerged as a global public health emergency on December 12, 2020, affecting 220 countries worldwide. This disease is caused by an extreme acute respiratory syndrome called coronavirus 2 (SARS-CoV-2) [1, 2], first identified in Wuhan, China. This new virus began to transmit rapidly globally, and the WHO declared the epidemic to be International Public Health Emergency of International Concern (PHEIC), and soon after it was declared a pandemic [3, 4]. Around 170 million confirmed cases and more than 3.54 million deaths were registered worldwide, with a Case Fatality Rate (CFR) of 2 as of May 31, 2020, [5]. In the same time interval, the Kingdom of Saudi Arabia has registered a lower CFR 1.7 similar to the international clinical manifestation and risk factors of COVID-19 [6, 7, 8].

Various prevention measures have been promoted, namely encouraging the use of face masks, isolation, quarantine, lockdowns, and travel bans promoted locally and by the World

Health Organization (WHO) [9, 10]. As soon as COVID-19 infections appeared in a few developed countries, COVID-19 cases have been exponentially dispersed in developing countries. All over the world, scientists have made incredible efforts to cope with COVID-19; there are still many missing puzzle pieces they do not understand about the disease [11]. For instance, asymptomatic cases [12] could be between 5 % and 80 % of individuals who test positive for COVID-19, including children and young adults. The Centers for Disease Control and Prevention (CDC) stated that there are 11 common symptoms for COVID-19 [13], yet many COVID-19 infected patients did not develop vital sign abnormalities.

One of the most pressing questions about COVID-19 that was not answered is how deadly COVID-19 is. It is challenging to be confident about the death rate within a designated geographical area. Currently, between 0.02 % - 0.82 % of people infected with the virus are estimated to die [14], although the mortality rate could be lower if there are large numbers of asymptomatic patients. While in Saudi Arabia, it is estimated that there have been more than 7,362 death. Moreover, 1,438 critical cases have been recorded until March 2021, which is depicted in Fig. 1.

The massive volume and the increasing velocity of COVID-19 data pose an enormous challenge. Because they play a crucial role in disease transmission, asymptomatic carriers and healthcare workers should be given special consideration. The recent rapid and exponential rise in the number of infected patients has impelled accurate prediction models, for instance, Artificial Intelligence (AI), for potential outcomes.

Machine Learning (ML) and self-learning concepts have become tightly coupled terminology over the last decade. ML models have been used to detect patterns and improve prediction performance using statistical modeling. ML models operate upon input data and empirical information without direct programming. Probabilistic reasoning, trial and error, and other computational-intensive methods are key players in ML models [16].

ML has proved to be a major research area in the resolution of many highly complex and advanced real-world problems Rustom et al. [17]. These days, one of the main areas of ML application is healthcare. For example, Petropoulos and Makridakis [18] provided live projections with reported cases

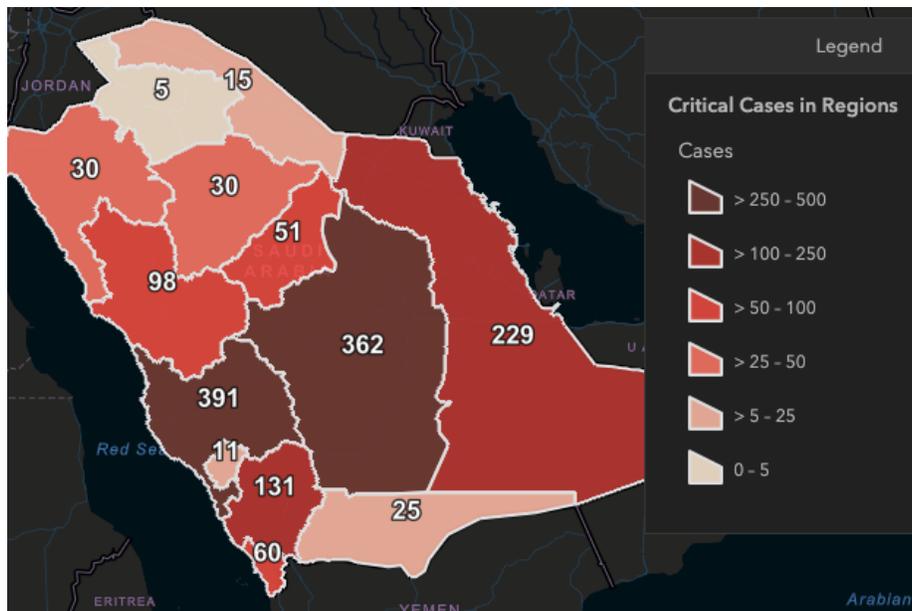


Fig. 1. Overview of Critical Cases Recorded in Saudi Arabia Until May 2021 [15].

of COVID-19. Grasselli et al. [19] also focused on predicting the epidemic and the early response to COVID-19. In the same way, these predictive systems can help manage current scenarios to direct early interventions to manage these diseases very efficiently.

There is a need for a novel model based on ML methods to investigate and predict the severity of the asymptomatic carriers and the possible death rate from the conditions mentioned earlier. This paper aims to propose a model to predict COVID-19 patients admission to the intensive care unit (ICU). The model is built upon robust known classification algorithms, including classic Machine Learning Classifiers (MLCs), Artificial Neural Network (ANN) and ensemble learning. The model may use clinical characteristics of COVID-19 on a national scale.

The contribution of this work is two-folded:

- This work presents a comparative analysis of multiple MLCs that intended to process COVID-19 patients' dataset in Saudi Arabia where an ANN and ensemble learning method were identified as good models for ICU admission prediction.
- This work serves as a starting point for the research community to explore different ML classifiers for better prediction of ICU admission.

The rest of the article is organized as follows: Section II demonstrates a few related works in which MLCs have been used for COVID-19 prediction. Section III describes the model details along with the dataset description and the data analysis. Section IV depicts the performance of the proposed model. Section V sheds a light upon a further discussion, and Section VI provides a conclusion and the future direction.

II. RELATED WORK

This section demonstrates the most popular MLCs and the architecture of the proposed model.

This pandemic has led to various public health and socio-economic concerns, highlighting the importance of unraveling the disease's evolution and predicting future dynamics. Various mathematical paradigms have often played a significant part in providing a deeper understanding of disease transmission mechanisms, adding significant insights to controlling the disease's spread. Calafiore et al. [20] and Nesteruk [21] have argued in favor of Mckendrick's suggested Susceptible Infectious Removed (SIR) disease model, and subsequently, one of the family models for human-to-human transmission is reasonably predictive. The Susceptible Exposed Infectious Removed (SEIR) model was created with many variants to forecast the possible dynamics of an outbreak. In a model that is based on population, integrating such real-world dynamics is still very difficult. Indeed, due to the lack of suitable actual historical data, research and prediction could go wrong. Instead, various models that focused on stochastic agents have been used as effective methods for monitoring the fine-grained effects of heterogeneous disease intervention policies on multiple disease outbreaks [22, 23, 24, 25]. However, due to the network structures' time-varying existence, this method's consistency can be a critical problem.

For reported and unreported infections, Li et al. [1] suggested an SEIR model after the virus's outbreak, integrating a meta-population framework, considering journeys between major cities in China. Their research discovered that before traveling was banned on January 23, 2020, about 86% of cases went unreported in Wuhan. As per their estimate, about 55% of asymptomatic spreaders were infectious on an individual basis, resulting in 79% of newly infected cases. In Verity et al. [26] later projected that about 63% of Italy cases were under-reported by reviewing an updated SIR model. Applying

the Susceptible-Infectious-Recovered-Deceased (SIRD) model to a Chinese official's statistics, Anastassopoulou et al. [27] estimated the COVID-19 pandemic's propagation dynamics in Hubei, using linear regression to assess the parameters of the model.

A more comprehensive SEIDIUQHRD¹ model has been proposed considering the above limitations of different proven mathematical models, taking into account all potential interactions, which can provide a more precise and stronger short-term and long-term estimation of the future COVID-19 dynamic model. As of May 11, the early stage's mathematical model parameters have changed dramatically. However, due to the strict containment measures and large-scale testing strategy, the outbreak has improved in many countries [28]. Considering the features of coronavirus infection reported by health organizations or quantitatively assessed in the literature, the model parameters' nominal values were considered [1, 26, 29, 30].

A few efforts aimed to provide a deeper understanding of the disease's spread and healthcare management aspects. ML models' capacity to predict the number of potential patients affected by COVID-19 is shown by Rustam et al. [17]. In this study, four typical forecasting models have been used to predict the threatening factors of COVID-19, such as linear regression (LR), minimum absolute shrinkage and selection operator (LASSO), support vector machine (SVM), and exponential smoothing (ES). The findings show that followed by LR and LASSO; the ES performed best out of all the models used, which performs well in predicting newly recorded incidents, death rate, and recovery rate, while SVM performs poorly in all projection scenarios given the available dataset.

The author in Iwendi et al. [31] provided a fine-tuned method for the Random Forests (RF) algorithm boosted by AdaBoost. This model uses the geographical, travel, health, and demographic data of COVID-19 patients and calculates the potential outcome of recovery or death. This study has a 94% accuracy and an F1 score of 0.86 for the used dataset.

Kutia et al. [32] attempted to break down consumer perceptions on China's eHealth applications and Ukraine's eHealth platform, which subsequently generated bits of knowledge and recommendations for refining an eHealth application (eZdorovya) specifically for health information benefits.

The proposal by Ardabili et al. [33] presented a relative analysis of ML and soft computing methods to forecast the COVID-19 pandemic. These two models present essential outcomes (i.e. multi-layered perceptron, adaptive network-based fuzzy inference system) based on the complexity of work that showed ML models were effective compared to their peers.

The study was given by Yan et al. [34], examining 404 infected patients' blood samples in the Wuhan region of China to classify critical predictive biomarkers of the seriousness of the disease. Three biomarkers that predict individual patients' survival with more than 90 % accuracy have been selected through ML tools for this function: lactic dehydrogenase (LDH), lymphocyte, and high-sensitivity C-reactive protein (hs-CRP). In particular, relatively high levels of LDH alone seem to play a crucial role in distinguishing the overwhelming

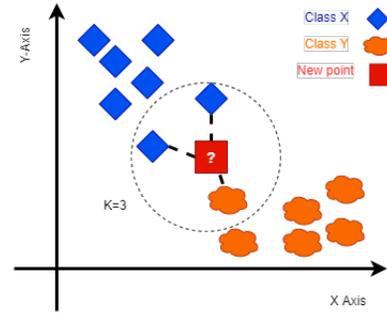


Fig. 2. Calculating Distance between a New Data point and Finding Closest Neighbors Through k-NN Classifier.

number of incidents that required immediate medication. This finding was consistent with a current medical understanding that various diseases, including pulmonary disorders (e.g. pneumonia) are correlated with elevated levels of LDH, causing tissue breakdown. Overall, the authors proposed an exact and operable formula for quickly predicting patients at the highest risk, enabling them to prioritize and potentially lower the mortality rate.

The research proposed by Pinter et al. [35] aimed to construct a hybrid approach to ML to predict COVID-19 and uses Hungary-based data to exemplify its potential. To forecast the time series mortality rate for infected individuals, the adaptive network-based fuzzy inference system (ANFIS) and multi-layered perceptron-imperialist competitive algorithm (MLP-ICA) hybrid ML methods are recommended. The online COVID-19 tracker suggested by Hamzah et al. [36] was used to interpret world sentiment trends on the dissemination of relevant health information and determine the political and economic effect of the virus's spread.

III. MATERIAL AND METHODS

This section demonstrates the most popular MLCs and the architecture of the proposed model.

A. Background

1) *k-Nearest Neighbors (k-NN)*: k-Nearest Neighbors (k-NN) is a simple supervised [16] ML model. It stores all input data and computes the distance between each new data point and all existing data points (i.e. neighbors). k-NN expects to label a new data point in a class based on the closest data points. Whenever many of the neighboring points become the majority, they will be designated as a class [37]. Fig. 2 depicts an overview of the k-NN model. The distance between a current data point and its neighbors is calculated as Eq. 1.

$$D(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}, \quad (1)$$

where D is the distance between x and y points.

Although k-NN is known for its accuracy and user-friendliness, it cannot tolerate multi-dimensional data with multiple features [16].

¹SEIDIUQHRD stands for Susceptible-Exposed-Symptomatic Infectious-Asymptomatic Infectious-Quarantined-Hospitalized-Recovered-Dead model.

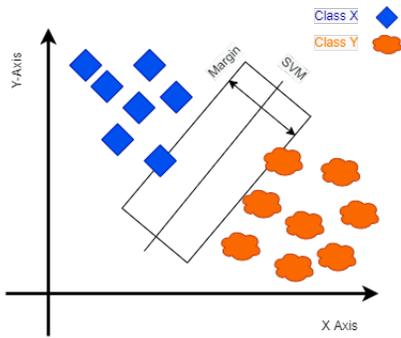


Fig. 3. Predicting the Class of a New Data Point using SVM Classifier.

2) *Support Vector Machine (SVM)*: Support Vector Machine (SVM) is an MLC used to predict categorical-based dataset decisions. SVM uniquely classifies data points by filtering them into binary or multiclass variables. It maintains a plane (i.e. hyperplane) to cover several dataset features. The maximum margin will be calculated in each plane, and the sides of the plane will be assigned to different classes. The plane is also a reinforcement used for future data points to be classified on either side of the plane's sides. Fig. 3 shows an overview of the SVM model. The maximum margin and the data point classification are calculated as Eq. 2, where the hypothesis function is h , vector x , and vector w .

$$h(x_i) = \begin{cases} +1 & \text{if } w \cdot x + b \geq 0 \\ -1 & \text{if } w \cdot x + b < 0 \end{cases} \quad (2)$$

Since SVM is useful in detecting complex relationships between classes, it is less sensitive to outliers and anomalies [16].

3) *Decision Tree (DT)*: Decision tree (DT) classifier intends to maintain a decision based upon learning a hierarchy of if\else tests of each decision point within a hierarchy of connecting nodes, starting from the root node to leaf node. It navigates all possible tests within the hierarchy and finds the most informative test of a target variable. Fig. 4 depicts an overview of a DT model. Gini index and entropy criteria are important measures to determine the optimum split of the tree features and are calculated through Eq. 3 and Eq. 4, respectively.

$$Gini_{index} = 1 - \sum_j p_j^2, \quad (3)$$

where $0 \leq Gini_{index} \leq 1$ and p_j is the probability of class j .

$$Entropy = - \sum_j p_j \cdot \log_2 \cdot p_j, \quad (4)$$

where p_j is the probability of class j .

Although DT is known to be used for classification and regression tasks and tends to work well with completely different scale features, it is over-fitting and may provide poor generalization performance for models [37].

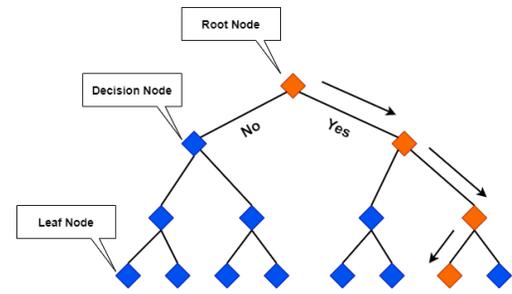


Fig. 4. Navigating Toward Optimum Split a the Tree Features through of DT Classifier.

4) *Random Forests (RF)*: Random Forests (RF) is a well-known regression and classification model. RF encompasses the collection of decision trees in which each tree could override part of the dataset. RF averages each tree's final results at the end of the prediction process and finalizes it. Fig. 5 provides an overview of the RF model.

Although RF surpasses the individual DT and its deficiencies, handles vast datasets, and its training stage can be rapidly executed in parallel, RF does not operate well on very high-dimensional or sparse data. Moreover, it requires massive computing resources for prediction tasks [37].

5) *Artificial Neural Network (ANN)*: An ANN is an advanced ML model used to construct prediction models. It mimics the slender projection's structural function through human brain nerve cells [16]. ANN has been seen as a generalization of linear models, executing different processing segments to reach a decision [37]. ANN consists of interconnected functions, known as nodes, that interact with each other through axon-like edges to maintain a network of nodes. Each input node then passes information through the network's edges to the next layer of nodes. The model's predicted output is compared with the actual output, which is understood to be right, using supervised learning. The expense's magnitude defines the difference between these two outcomes. Conventional neural networks can be subdivided into input, hidden, and output layers. Input (i.e. datasets' points) is initially received through the input layer, where the features are detected. The hidden layer(s) will then analyze and process the input features, and the final result will be shown as the output layer. The training goal is to reduce the value of the cost until the estimation of the model accurately reflects the correct output. One of the main advantages of neural networks is that they can detect the information contained in massive amounts of data given adequate time for computing, data and careful tuning. Fig. 6 depicts an overview of the ANN model. The cost of an ANN is calculated through the activation function as demonstrated in Eq. 5.

$$\hat{y} = w[0] * x[0] + w[1] * x[1] + \dots + w[p] * x[p] + b, \quad (5)$$

where \hat{y} is the weighted sum of the input features $x[0]$ to $x[p]$, and weighted by the learned coefficients $w[0]$ to $w[p]$.

It often takes a long time to train neural networks, especially large/strong ones. Besides, they require proper pre-processing of the information. Also, they work best with

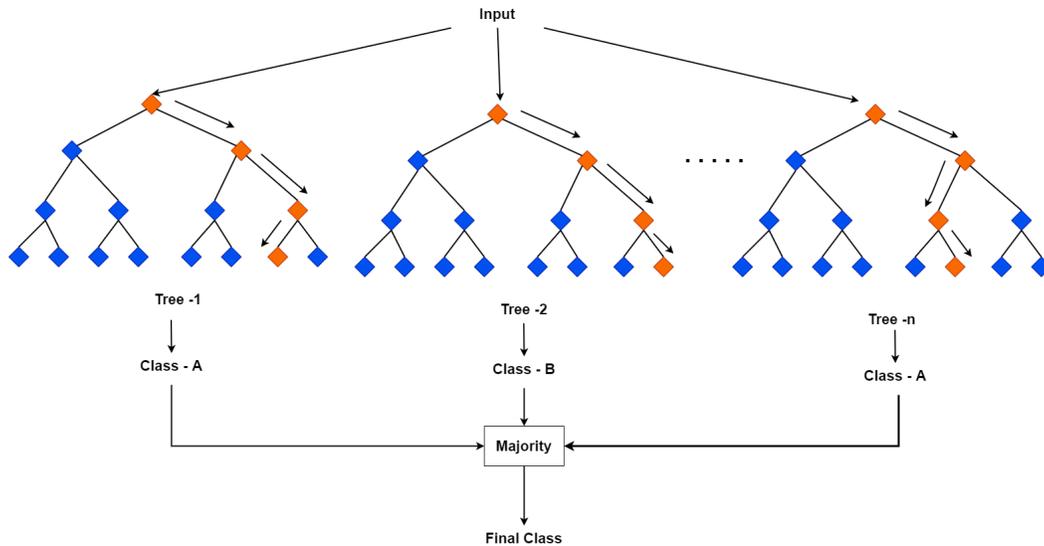


Fig. 5. Averaging each DT's Results to Arrive at the Final Class Through RF Classifier.

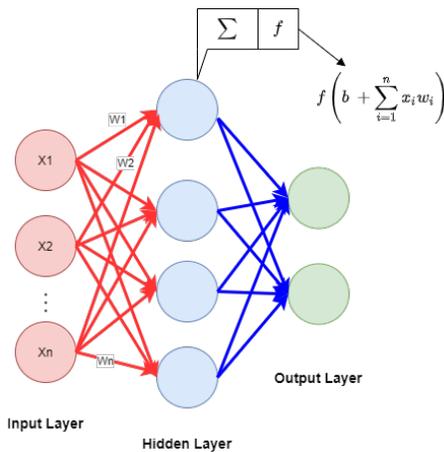


Fig. 6. Overview of ANN Classifier with Three General Layers, where x_i Represents an Input, w_i is Corresponding Weight, b is bias, and f is the Activation Function to be Applied to the Weighted Sum of the Inputs.

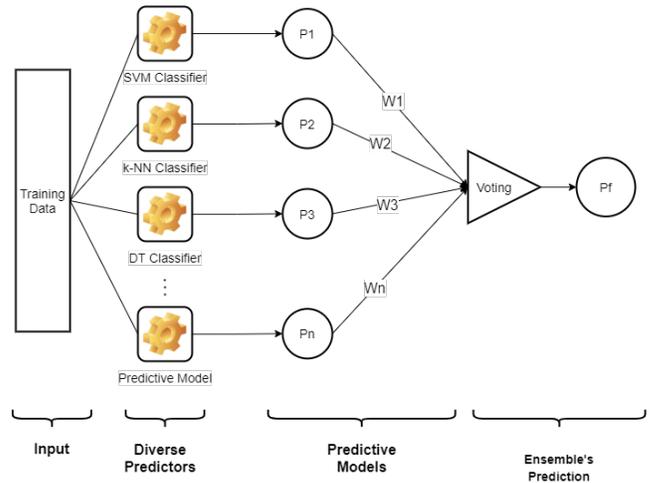


Fig. 7. Overview of an Aggregate of the Predictions of MLCs and the Prediction of Predict the Class that Gets the Most Votes, where P_n Represents Model Feature, W_n Represents Model Weight, and P_f Represents the Majority Vote.

“homogenous data”, where all features have similar meanings [16].

6) *Ensemble Learning Method*: Ensemble learning method - advanced ML classification model - considers and tests more than one ML model and finds the best model for input data to combine ML models. The ensemble learning method integrates various ML models into a single prediction using a complete analysis data voting system. These two methods, together, provide additional benefit over any homogeneous model. Fig. 7 depicts an overview of an ensemble learning method.

While the ensemble learning method performance exceeds a single algorithm in many other homogeneous models, the degree of complexity and sophistication of the method can be a potential drawback. The ensemble learning method produces the same advantages as the single ML model, where transparency and ease of interpretation are surrendered to the accuracy of a more complex algorithm, such as RF, bagging

or boosting [16].

B. Proposed Model

The current rapid and exponential increase in the number of infected patients has necessitated an accurate estimation of suitable ML models' potential outcomes. We propose a novel model based on MLCs to investigate and predict the severity of the asymptomatic carriers and the possible death rate. The model entails various MLCs, including k-NN (refer to Section III-A1), DT (refer to Section III-A3), RF (refer to Section III-A4), SVM (refer to Section III-A2) and the ensemble learning method (refer to Section III-A6). The model uses a private dataset's clinical characteristics in Saudi Arabia (refer to Section III-B1). In addition, model performance has been evaluated in terms of accuracy based on the confusion

matrix [38]. The proposed model can be built from scratch or by using the existing pre-trained models. Data acquired from the hospital(s) will be separated into training and testing clusters.

The proposed model architecture consists of five main phases, namely data acquisition, pre-processing, feature extraction, feature selection, and classification. An overview of the proposed model is presented as Fig. 8, and the primary taken stages are presented in Pseudocode 1.

Algorithm 1: Proposed algorithm to investigate and forecast the incidence and the potential death risk of the asymptomatic carriers.

Input: COVID-19 Positive Patients (CPP) Dataset

Output: Recovery from ICU, Recovery Without ICU, Death

Extract features from CPP Dataset

Pre-process features

Normalize input dataset

Select feature from the dataset

Initialize dataset_{train} & dataset_{test}

Train and validate (dataset_{train}, dataset_{test})

[X_{train}, X_{test}, Y_{train}, Y_{test}]

Initialize ML models

Input the readings to the model

Train the ML model

Test unseen input dataset

for each Fold in range (10) do

 Evaluate Loss, Validation Loss

 Evaluate Accuracy and Validation Accuracy

 Evaluate Precision, F-Score, Roc Curve and

 Confusion matrix

end

First, we acquired the private dataset in Saudi Arabia. Second, pre-processing was performed on the acquired dataset, and the entire health dataset was cleaned for prediction purposes. Third, the model used Principle Components Analysis (PCA) for feature selection and vital known classification methods from the dataset. Then, the selected features were utilized to train and test the MLCs. Finally, various classification methods were used to classify testing data.

1) *COVID-19 Positive Patients (CPP) Dataset:* Across all Saudi Arabian regions, a private COVID-19 Positive Patients (necessitated) dataset [6] was collected between March 1, 2020, and March 31, 2020, with no exclusion criteria. The diagnosis of COVID-19 was conducted based on the performance of real-time reverse transcription-polymerase chain reaction (RRT-PCR) tests in a quantitative study undertaken at the Saudi National Health Laboratory. The patients' data were stored at the Saudi Health Electronic Surveillance Network Database (HESN)² under the authority of the Global Center for Mass Gatherings Medicine. The HESN contains demographic characteristics, COVID-19-positive clinical, lab, and raw outcome data in all Saudi Arabian regions. All efforts have been made to preserve the data's confidentiality and autonomy. A study identification number was allocated to all subjects. In the dataset, 12.5 % of all patients served in healthcare facilities, and 17.3 % were asymptomatic.

²Site: <https://hesn.moh.gov.sa/webportal/>

Table I depicts an overview of the dataset, and Table II presents an overview of the dataset features' profile. The dataset contains 18 numerical variables, 15 Boolean variables, and 11 categorical variables. The dataset has 639 records with 44 features, and it contains demographic and clinical data for live patients.

2) *Data Pre-processing:* Since the dataset was unbalanced, there were many missing values, each variable was unique, and data were imputed to handle the missing entries as necessary. This dataset has three variable types: 1) Boolean, 2) numerical, and 3) categorical. The "Outcome" feature was designated as an independent variable, and the other features were designated as dependent variables during the models' training. "Outcome" feature has three classes, namely "Active ICU or Recovery with history of ICU", "Death", and "Recovery No ICU". In the dataset, we have a record of 639 patients. Out of 639 patients, 563 patients belong to the "Recovery No ICU" class, 65 patients belong to the "Active ICU or Recovery with history of ICU" class, and only 11 patients belong to the "Death" class. Due to shallow death cases in training data, sometimes the classifiers cannot correctly predict the "Death" classes. The Synthetic Minority Over-sampling Technique (SMOTE) [39] was applied to the whole dataset to deal with its unbalanced nature. SMOTE aims to balance the dataset by oversampling the minority class and duplicating the same entities without any new information. After applying SMOTE to the dataset, we have 563 patients in each class.

Boolean features ("Fever_PRESENT", "SoreThroat", "RunnyNose_PRESENT") entail "Yes" and "No" values where "Yes" means the symptom was present and "No" means the symptom was not present. In order to unify the code system within the entries, we replaced the "Yes" with "1" and "No" with "0" notations. In the "age_65" feature, there were two entries, "1" and "2" where "1" means a person was younger than 65, and "2" means a person was older than 65. In this variable, "1" was replaced with "0" and "2" was replaced with "1". Empty entries in these variables are replaced with "0" assuming that the symptom is not present. Other Boolean variables ("any_comorbidity", "DM1", "HTN1", "CRF1", "cardiac1", "asthma1", "cancer_immunodefecyncy1", "C_lungdisease") used the binary code system "0" and "1". Therefore, they are left untreated in the pre-processing part.

Numerical features ("comorbidities", "LOSdays", "Smoking", "dayofExposureifknown", "Incubation", "Temperature", "HEART_RATE", "RESPIRATOR", "SBP", "DBP", "SATURATION", "WHITE_CELLS", "CREATININE", "LYMPHOCYTES", "PLATELET", "NEUTROPHILS", "BLOOD") entailed missing values that were replaced with a median of their corresponding variable. Besides, categorical features ("ClassificationGroup", "Gender", "Nationality", "Outcome") had some missing values that were replaced with others. Other numerical features ("Myalgiaonset", "GISymptomsonset", "Headacheonset") entailed different scales that did not contribute equally to the models fitting. The "MinMax" scaler was applied to all these variables to scale the values on the same scale and prevent the problems caused by these values.

Categorical features ("Myalgiaonset", "GISymptomsonset", "Headacheonset") entailed intervals, for instance, "1_2" that implied an individual read of a symptom occurred every two

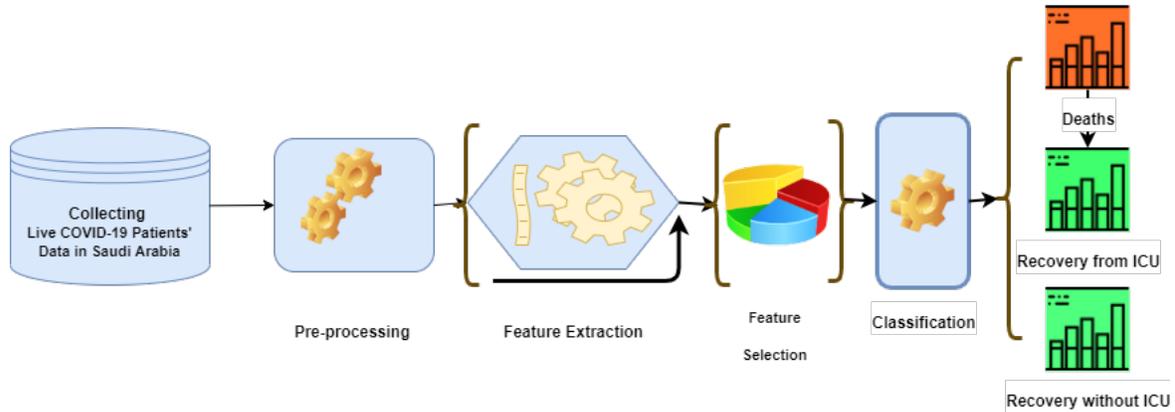


Fig. 8. Architecture of the Proposed Model to Investigate and Forecast the Incidence and the Potential Death Risk of the Asymptomatic Carriers.

TABLE I. DESCRIPTION OF THE 2019 CORONAVIRUS DISEASE (COVID-19) POSITIVE PATIENTS (CPP) DATASET.

Feature	Description	Initial Value	Null values(%)	Datatype
InvID	-	-	0	Categorical
ClassificationGroup	Epidemiological criteria	Case, Contact	0	Categorical
ClientName	-	-	0	Categorical
Outcome or Outcome_Modified	Admission status outcome	Recovery No ICU, Recovery with history of ICU, Death	0	Categorical
LOSdays	length of stay in days	-	1.3	Numerical
HCW_totalpop	Patient's occupation (Medical staff, Military, others)	"0", "1", "2"	0	Categorical
comorbidities	-	-	0.2	numerical
any_comorbidity	Any comorbidity?	(Yes, No)	0	Categorical
morethan2comorbidities	Two or more comorbidities?	(Yes, No)	0	Categorical
DM1	Diabetes?	(Yes, No)	0	Categorical
HTN1	Hypertension?	(Yes, No)	0	Categorical
CRF1	Chronic kidney disease?	(Yes, No)	0	Categorical
cardiac1	Heart diseases?	(Yes, No)	0	Categorical
asthma1	Asthma and chronic lung disease?	(Yes, No)	0	Categorical
cancer_immunodeficiency1	immunodeficiency?	(Yes, No)	0	Categorical
C_lungdisease	Lung disease?	(Yes, No)	0	Categorical
Smoking	Smoker?	(Yes, No)	0.2	Categorical
age_65	Age Above 65?	(Yes, No)	0	Categorical
Gender	Patient gender	Male, Female	0	Categorical
Nationality	nationality	Saudi Arabia, Egypt, Philippines, Sudan, India, and many others.	5.6	Categorical
SYMPTOMATIC	Symptomatic?	(Yes, No)	57.7	Categorical
Fever_PRESENT	Fever?	(Yes, No)	38	Categorical
Cough_PRESENT	Cough?	(Yes, No)	65.1	Categorical
SoreThroat_PRESENT	Sore Throat?	(Yes, No)	77.6	Categorical
RunnyNose_PRESENT	Runny nose?	(Yes, No)	85	Categorical
Headacheonset	Headache frequency	"0", "1_1", "1_2", "1_3"	21.6	Categorical
Myalgiaonset	Myalgia frequency	"0", "1_1", "1_2", "1_5"	21.6	Categorical
GI symptomsonset	GI symptoms frequency including abdominal pain, vomiting, or diarrhea	"0", "1_1", "1_2", "1_3", "1_4"	21.6	Categorical
dayofExposureifknown	-	-	64.9	Numerical
Incubation	-	-	65.1	Numerical
Temperature	oral temperature of 38° or higher	-	58.5	Numerical
HEART_RATE	Heart beats per minute	-	82.2	Numerical
RESPIRATORY	number of breaths for an entire minute	-	83.6	Numerical
SBP	Systolic Blood Pressure	-	82.3	Numerical
DBP	Diastolic Blood Pressure	-	82.2	Numerical
SATURATION	Oxygen level	-	59.9	Numerical
WHITE_CELLS	White cell count	-	94.7	Numerical
CREATININE	Creatine phosphate from muscle and protein metabolism count	-	96.9	Numerical
LYMPHOCYTES	lymphocyte count of less than 1,500 per 1 Mio. m ³	-	96.4	Numerical
PLATELET	Platelet counts in the blood	-	94.8	Numerical
NEUTROPHILS	White blood cell type level	-	95.6	Numerical
SEVERITY	Patients' conditions	"0", "3"	0	Categorical

TABLE II. PROFILE INFORMATION OF THE CPP DATASET

Feature	Min.	Max.	Mean	Std. deviation
LOSdays	0	60	7.71	9.59
comorbidities	0	5	0.47	0.87
age_computed	0	84	36.59	15.57
dayofExposureifknown	1	30	8.5	6.30
Incubation	1	30	7.21	6.12
Temperature	0	39.1	37.05	2.37
HEART_RATE	63	125	89.82	14.19
RESPIRATORY	14	30	19.87	2.48
SBP	60	188	125.12	17.22
DBP	57	116	75.03	10.46
SATURATION	69	100	96.68	3.63
WHITE_CELLS	0	17	6.53	3.96
CREATININE	0	145	64.55	43.63
LYMPHOCYTES	6.3	58.1	24.43	13.00
PLATELET	107	572	246.52	92.53
NEUTROPHILS	1.37	93.4	48.68	32.52

TABLE III. SAMPLE OF CONFUSION MATRIX

		Actual class	
		P	N
Predicted class	P	T_P	F_P
	N	F_N	T_N

days. There were some empty entries in these variables, and they were replaced with the median of each variable.

“Client name” and “InvID” are removed because these variables have no role in classification.

C. Experimental Setup

In this work, the ML methods were executed using Python programming language (version 3.7) in Google Colaboratory (Colab)³. Colab is a web-based platform intended to run python code and customized for AI and data analysis tasks. Colob provides on-demand computing services for ML-related tasks. The following specifications were used: CPU (Intel Xeon 2.20 GHz, family 6, Model 79), RAM (1 GB), and Storage (Google Drive with 69 GB), including all the necessary libraries: Pandas, sklearn, NumPy, Seaborn, SciPy, Keras, ELI5, and TensorFlow.

D. Performance Evaluation Metrics

To accurately predict ICU admission depending on multiple COVID-19 symptoms and patients’ clinical characteristics, accuracy is vital. A comparative analysis has been conducted between MLCs and the ensemble learning method for COVID-19 infected patient classification in this study. Based on the confusion matrix parameters, “accuracy”, “precision”, and “f-score” will be calculated to investigate the accuracy of the proposed approach. The confusion matrix is represented in Table III.

A few notations are used within the evaluation metrics. True Positive (T_P) is the number of true positives classified by the prediction model. True Negative (T_N) is the number of true negatives classified by the prediction model. False Positive (F_P) is the number of false positives classified by

the prediction model. False Negative (F_N) is the number of false negatives classified by the prediction model.

1) *Accuracy*: The accuracy predictor is the ratio between the total accurate predictions ($T_P + T_N + F_P + F_N$) and the total data points of the classifier. Accuracy defines the full variety of correct predictions laid out in percentage. Eq. 6 calculates Accuracy as follows:

$$Accuracy = \frac{T_P + T_N}{T_P + T_N + F_P + F_N}, \quad (6)$$

where $0.0 < Accuracy < 1.0$

2) *Precision*: The ratio of the True Positive (T_P) samples to the sum of the True Positive (T_P) and False Positive (F_P) samples is equal to the precision predictor. Precision is outlined because of the entire range of correct positive predictions portrayed in percentage. Eq. 7 calculates Precision as follows:

$$Precision = \frac{T_P}{T_P + F_P} \quad (7)$$

3) *F Score*: The F score is defined as the harmonic mean of precision and recall of the model. Eq. 8 calculates F Score as follows:

$$F \text{ score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (8)$$

E. Error Metrics

We calculated four error metrics: Root Mean Squared Error (RMSE), Relative Squared Error (RSE), Mean Absolute Error (MAE), and Relative Absolute Error (RAE).

1) *Root Mean Squared Error (RMSE)*: RMSE determines the mean of the magnitude of the error. Eq. 9 calculates RMSE as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (p_i - a_i)^2}{n}}, \quad (9)$$

where a is the actual target, and p is the predicted target.

2) *Relative Squared Error (RSE)*: RSE compares the sum of errors of the model compared to a simple predictor (using the average). Eq. 10 calculates RSE as follows:

$$RSE = \frac{\sum_{i=1}^n (p_i - a_i)^2}{\sum_{i=1}^n (\bar{a} - a_i)^2} \quad (10)$$

3) *Mean Absolute Error (MAE)*: MAE measures the average of all the absolute errors. Eq. 11 calculates MAE as follows:

$$MAE = \frac{\sum_{i=1}^n |p_i - a_i|}{n} \quad (11)$$

4) *Relative Absolute Error (RAE)*: RAE calculates the square root of the relative squared error. Eq. 12 calculates RAE as follows:

$$RAE = \frac{\sum_{i=1}^n |p_i - a_i|}{\sum_{i=1}^n |\bar{a} - a_i|} \quad (12)$$

³Site: <https://colab.research.google.com/>

TABLE IV. COMPARISON OF THE PROPOSED MODEL WITH THE RELATED MODELS

ML Architecture	Accuracy (%)
Pal et al. [40]	77.6
Tang et al. [41]	87.5
Yan et al. [34]	90
Iwendi et al. [31]	94
Proposed	97.93

TABLE V. OVERVIEW OF PERFORMANCE EVALUATION METRICS OF VARIOUS MLCs AND THE ENSEMBLE LEARNING METHOD

Classifiers	Accuracy	Precision	AUC	F-Score
k-NN	0.87968	0.889014	0.7997	0.86935
SVM	0.94872	0.949557	0.9115	0.94746
DT	0.88166	0.882669	0.8138	0.87966
RF	0.96647	0.96691	0.9358	0.96585
ANN	0.9783	0.97937	0.9508	0.97778
Ensemble learning method	0.94675	0.949887	0.9048	0.9445

IV. RESULTS

This section demonstrates our results from the constructed ML models and the associated evaluation metrics.

A. Comparative Analysis of Various Classifiers

We presented an evaluation of the proposed model for COVID-19 predictive models with the related methods presented in [40], [41], [34], and [31] in Table IV. The proposed model provided 97.93% accuracy, while the comparative models attained an average accuracy of 87.28%, so that the proposed model showed a 10.7% performance gain.

We measured various performance metrics of the used classifiers: Accuracy, Precision, Area Under the ROC Curve (AUC), and F-score. We calculated ANN's results by comparing them with the outcomes obtained from MLCs and the ensemble learning method. We divided the COVID-19 dataset using the beforementioned classifiers into 30% of the dataset for the testing mode and 70% of the dataset for the training mode. All the classifiers' execution time took 0.03 seconds, except ANN, which took 45 seconds.

The performance of the ML methods was calculated and is shown in Fig. 9 and portrayed in tabular form in Table V. Accuracy, Precision, AUC, and F1-score scores of all the classifiers are represented in Fig. 10, Fig. 11, Fig. 12, Fig. 13, respectively.

Using MAE, RMSE, RAE, and RRSE, the error rate of each predictor was calculated and shown in Fig. 14 or Table VI. In Fig. 15, Fig. 16, Fig. 17, and Fig. 18, the different error rates obtained for different classifiers are shown, respectively.

MAE is used to calculate the average absolute difference between the expected and the observed set of values, given that the weight of each difference is the same. The following figure shows that ANN gives the lowest error at 8%, whereas the decision tree is 42%, SVM is 17%, and RF is 13%. Therefore, the ANN model provides a 92% match to actual values with 8% error, while RF provides an 87% match to actual values with only 13% error. The MAE graph is shown in Fig. 15.

TABLE VI. DIFFERENT ERROR METRICS GIVEN BY VARIOUS MLCs AND THE ENSEMBLE LEARNING METHOD

Classifiers	MAE	RMSE	RAE	RRSE
k-NN	0.3452	0.5875	0.2912	0.7247
SVM	0.1755	0.419	0.1382	0.5168
DT	0.426	0.6527	0.3294	0.8051
RF	0.1341	0.3662	0.1	0.4518
ANN	0.0809	0.2844	0.0618	0.3508
Ensemble learning method	0.1598	0.3997	0.1323	0.493

RMSE also calculates the median magnitude of the variations as MAE (i.e. errors). The square root of the mean of the square deviations is RMSE. As RMSE is more suitable than MAE, the ANN model has 28% RMSE, indicating a 78% accurate classification of COVID-19 ICU recoveries, death, and recoveries with 28% error, while RMSE falls between 39% and 60% of the remaining ML models. The RMSE graph is shown in Fig. 16.

RAE is the same as RRSP, which is determined by dividing MAE by the simple classifier error received. The smaller the value of RAE, therefore, the better the prediction. Fig. 17 shows that the proposed ANN method attained 6% of RAE, a superior RAE value.

RRSP provides the forecasts' squared error relative to every data value's mean. It gives more accurate results than simple classifiers by normalizing the values obtained from the simple classifiers (e.g. k-NN or ANN). RRSP divides the total squared error by splitting it with ASE (absolute squared error) obtained from the simple classifiers. Furthermore, the error is minimized by producing a normalized value square root. As shown in Fig. 18, the ANN model exceeds 35%, while other classifiers are greater than RRSE.

The different proposed methods (k-NN, SVM, DT, ANN, RF, and the ensemble learning method) for classification on the CPP dataset yields 87.96%, 94.87%, 88.16%, 97.83%, 96.64%, and 94.67%, respectively. The accuracy of various ML and the ensemble learning methods was ensured compared to the evaluation of classifier error rates by Table VI, as it considers the relative significance of each factor found in the study. Thus, the ANN model provides greater precision than other variants on the CPP dataset by alleviating data inconsistencies.

Table VII shows the p-values obtained using the student test while examining possible significant differences by various pairs of classifiers. We calculated p-values between MLCs against each other, and then calculated p-values between MLCs and the ensemble learning method. By comparing the ANN and k-NN classifiers, the p-values are smaller than 0.05, by which there was no significant difference (i.e. rejecting the null hypothesis). Comparing ANN, RF and the ensemble learning method, the p-values are smaller than 0.05, by which there was no significant difference (i.e. rejecting the null hypothesis).

However, when we compare ANN and SVM against RF, the p-values are larger than 0.05, by which there was a significant difference (i.e. accepting the null hypothesis). When we compare k-NN and SVM against RF, the p-value is smaller than 0.05, by which there was no significant difference (i.e. rejecting the null hypothesis). When we compare k-NN and

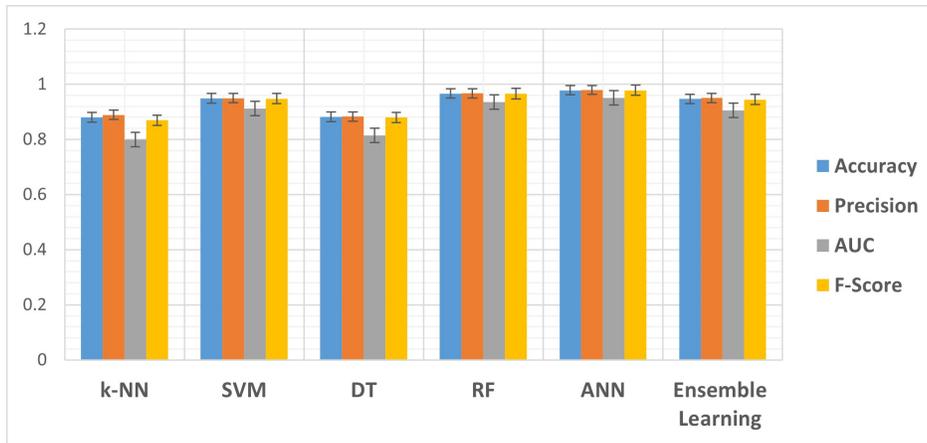


Fig. 9. Comparison of Performance Analysis of Various MLCs and the Ensemble Learning Method.

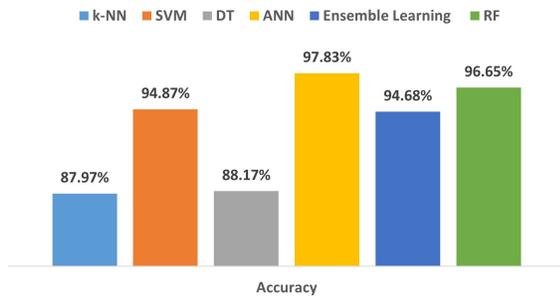


Fig. 10. Accuracy Scores of Various MLCs and the Ensemble Learning Method.

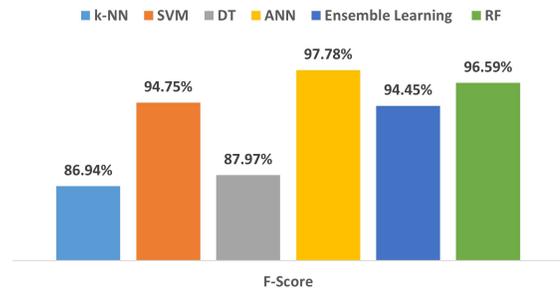


Fig. 13. F-Score Scores of Various MLCs and the Ensemble Learning Method.

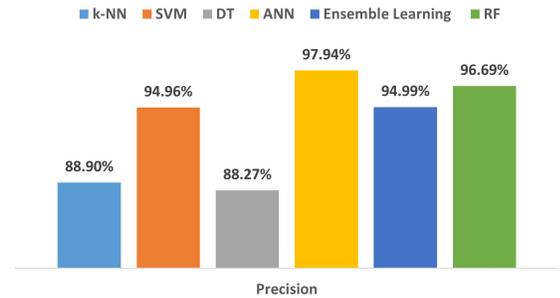


Fig. 11. Precision Score of Various MLCs and the Ensemble Learning Method.

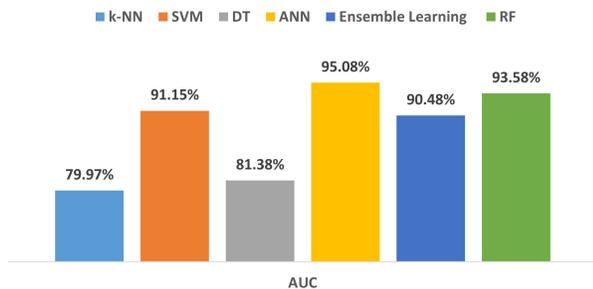


Fig. 12. AUC Scores of Various MLCs and the Ensemble Learning Method.

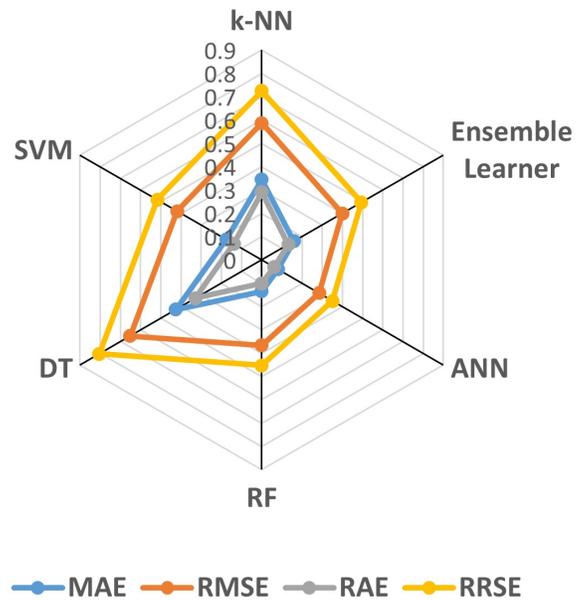


Fig. 14. Different Error Metrics Given by Various MLCs and the Ensemble Learning Method.

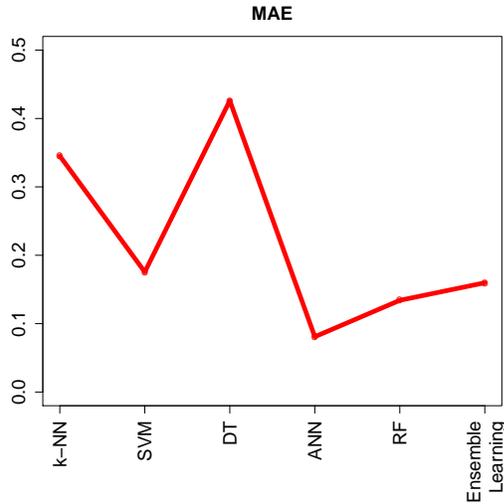


Fig. 15. Mean Absolute Error (MAE) Analysis of Individual MLCs and the Ensemble Learning Method.

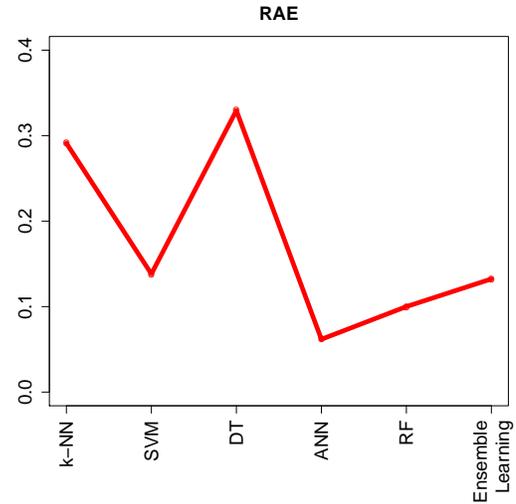


Fig. 17. Relative Absolute Error (RAE) from Various MLCs and the Ensemble Learning Method.

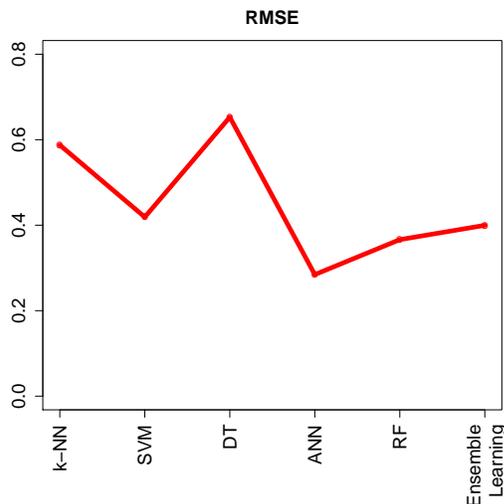


Fig. 16. Root Mean Square Error (RMSE) Analysis of Individual MLCs and the Ensemble Learning Method.

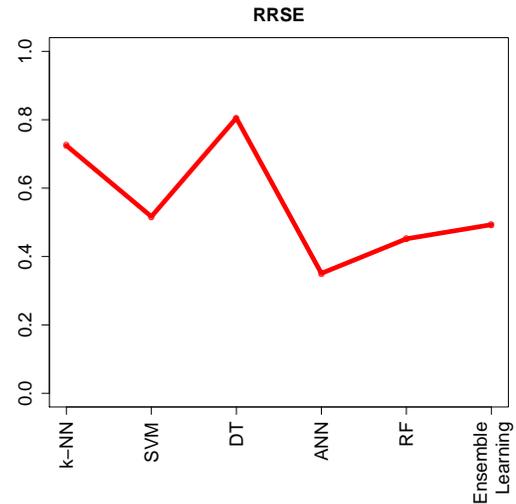


Fig. 18. Root Relative Squared Error (RRSE) from Various MLCs and the Ensemble Learning Method.

TABLE VII. P-VALUES OF VARIOUS CLASSIFIER SCORES

Algorithms	t-test	p value
ANN - k-NN	3.191458	0.012775
ANN - SVM	0.910476	0.38918
ANN - DT	2.106662	0.06822
ANN - RF	-0.81257	0.439975
ANN - Ensemble learning method	3.191458	0.012775
k-NN - SVM	-2.80261	0.023105
k-NN - DT	-1.06956	0.316024
k-NN - RF	-4.12081	0.00334
k-NN - Ensemble learning method	0	1
SVM - DT	1.596006	0.149154
SVM - RF	-2.15321	0.063458
SVM - Ensemble learning method	2.802612	0.023105
DT - RF	-2.99951	0.017085
DT - Ensemble learning method	1.069559	0.316024
RF - Ensemble learning method	4.120807	0.00334

DT against the ensemble learning method, the p-value is larger than 0.05, by which there was a significant difference (i.e. accepting the null hypothesis). Similarly, the null hypothesis was accepted, and there was a significant difference when comparing SVM against RF. When comparing SVM with the ensemble learning method, the p-value was less than 0.05, by which there was no significant difference (i.e. rejecting the null hypothesis). Finally, when we compare DT's p-value with RF and RF with the ensemble learning method, the null hypothesis is rejected, so there is no significant difference among them. However, when comparing DT with the ensemble learning method, the null hypothesis was accepted.

V. DISCUSSION

Early COVID-19 prediction could help reduce healthcare systems' enormous burden by diagnosing patients with it. In

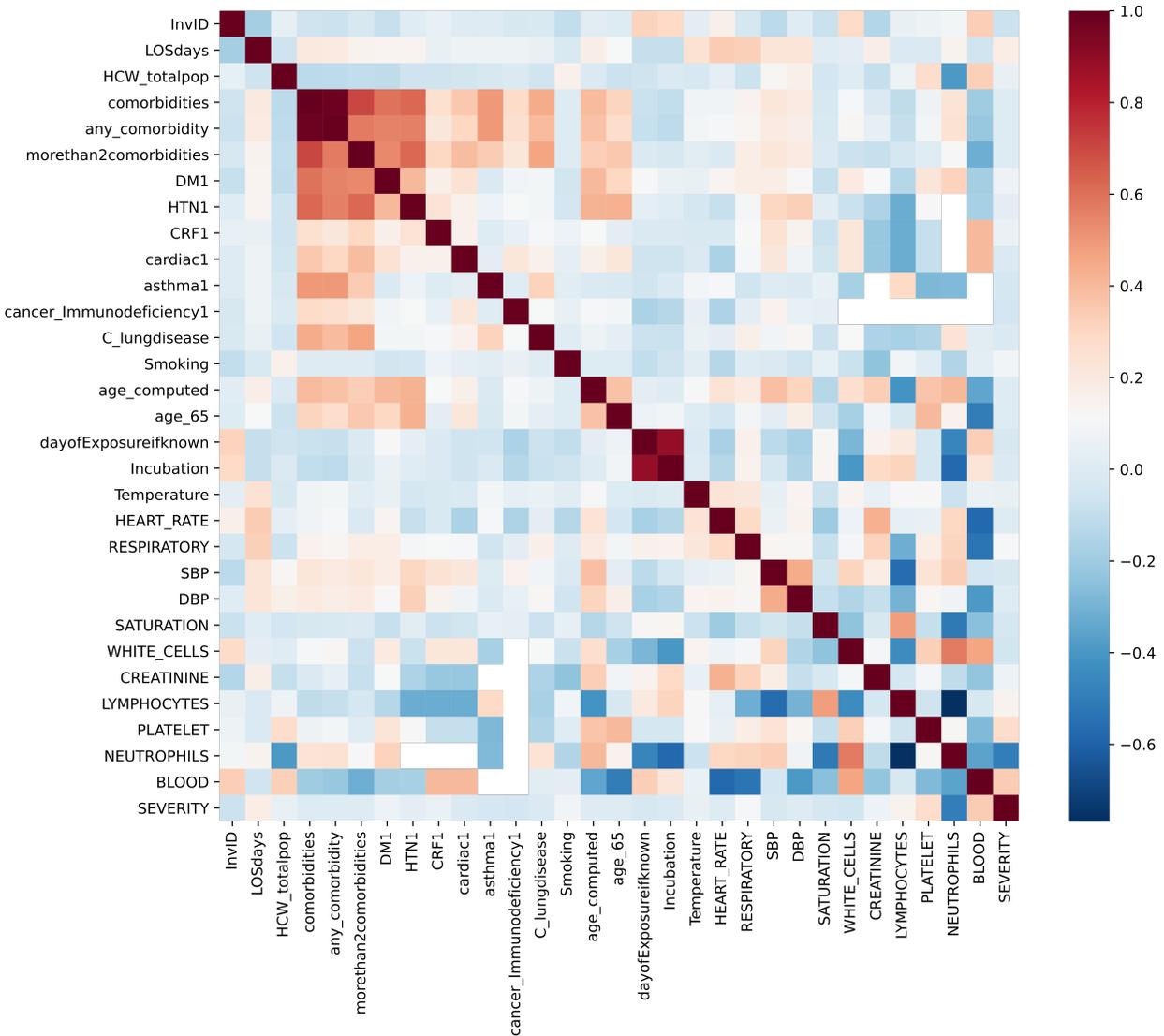


Fig. 19. Heatmap of Correlation Coefficient of CPP Dataset Features.

this study, k-NN, SVM, DT, RF, ANN, and the ensemble learning methods were constructed to predict patients' ICU admission with a CPP dataset for Saudi Arabian residents. Accuracy, Precision, AUC, and F-score were used to measure all models' performance.

Fig. 19 demonstrates various characteristics, including clinical conditions (fever, cough, sore throat, runny nose, headache, myalgia), GI symptoms, and blood work indicators (oxygen saturation, immune cells, creatinine, lymphocytes, platelet, neutrophils) and others, against the "Outcome" feature, which is an independent variable. Besides, commodities increase "LoSdays" and "Severity". Also, "age_65" correlates with "commodities". It was not expected that there would be no correlation between smoking and "LoSdays" or "Severity".

The feature importance for the dataset is shown for ANN and the ensemble learning method in Fig. 20 and Fig. 21, respectively.

From Fig. 20, the ANN model indicated that "LoSdays" is

the most important feature among all the dependent features, while some clinical characteristics, such as "Myalgiaonset" or "Headacheonset", are the least important features of the dataset. From Fig. 21, the ensemble learning method indicated that a patient's occupation and "LoSdays" are the most important features among all the dependent features of the dataset, while some clinical characteristics, such as cough or hypertension, are the least important features.

VI. CONCLUSION AND FUTURE WORK

Non-clinical methods, for instance, ML models, have been utilized as an alternative means for diagnosis and prognosis. In this study, a model was proposed to predict COVID-19 patients admission to the ICU was established with the CPP dataset using MLCs and the ensemble learning method in Saudi Arabia. The models were trained with 70% of the training data and tested with 30% of the data remaining. In terms of precision, the model developed with ANN was the best model among all models with 97.83%. This study

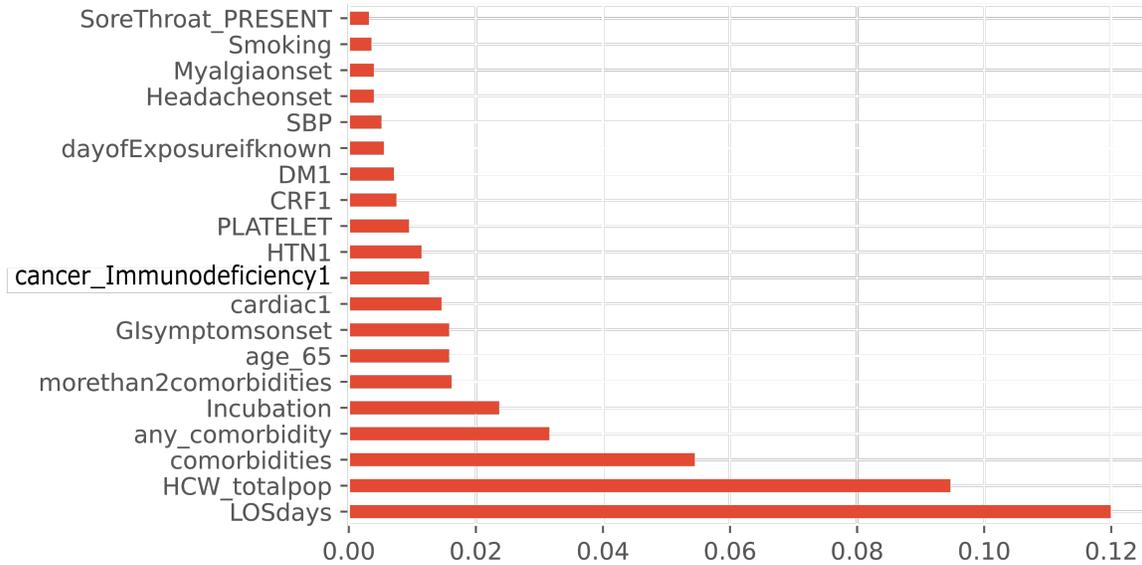


Fig. 20. Feature Importance Computed from ANN that was Fitted to CPP Dataset.

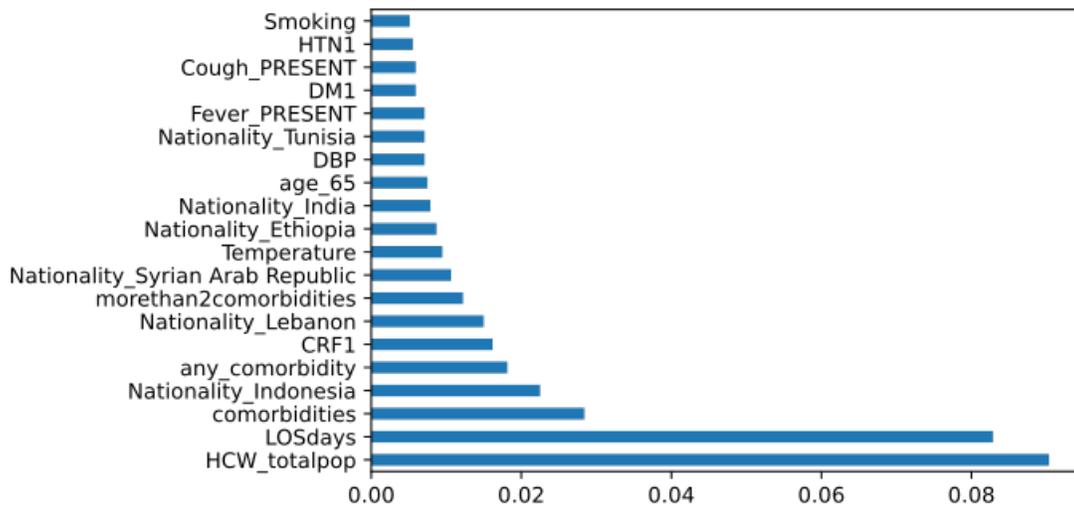


Fig. 21. Feature Importance Computed from the Ensemble Learning Method that was Fitted to CPP Dataset.

shows how it is possible to develop, validate, and use ML Predictive COVID-19 infection models as tools for ICU admission prediction. The models developed would potentially be vital in the battle against COVID-19 in healthcare. Future work will compare the current ML models against another prediction model, for instance, Ada boost. The developed models could be incorporated into decision-support systems for semi-autonomous diagnostic devices with rapid screening and diagnosis of possible outbreaks.

A. Data Availability Statement

The dataset presented in this study is private property for the Global Center for Mass Gatherings Medicine.

B. Authors Contributions

MZK, YA, RA, AA, AAK: idea conceptualization and project administration. HG, MZK, RA: methodology, investigation, data curation, and writing—original draft preparation. HG, MK: software. HG, MZK: validation and visualization. HG, MZK: formal analysis. YA, RA, AA: resources. HG, MZK, YA, RA, AA: writing—review and editing, supervision. The authors declared that they have read and agreed to the manuscript’s published version.

C. Funding

There is no funding for the project.

D. Conflict of Interests

All authors declared that there are no conflicts of interest.

ACKNOWLEDGMENT

This project has resulted from the collaboration between the Global Center for Mass Gatherings Medicine and the College of Computer Science and Engineering in Taibah University, starting from October 1, 2020 until February 20, 2021.

REFERENCES

- [1] R. Li, S. Pei, B. Chen, Y. Song, T. Zhang, W. Yang, and J. Shaman, "Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2)," *Science*, vol. 368, no. 6490, pp. 489–493, May 2020. [Online]. Available: <https://www.sciencemag.org/lookup/doi/10.1126/science.abb3221>
- [2] F. Zhou, T. Yu, R. Du, G. Fan, Y. Liu, Z. Liu, J. Xiang, Y. Wang, B. Song, X. Gu, L. Guan, Y. Wei, H. Li, X. Wu, J. Xu, S. Tu, Y. Zhang, H. Chen, and B. Cao, "Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study," *The Lancet*, vol. 395, no. 10229, pp. 1054–1062, Mar. 2020. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0140673620305663>
- [3] World Health Organization, "Statement on the second meeting of the International Health Regulations (2005) Emergency Committee regarding the outbreak of novel coronavirus (2019-nCoV)," Jan. 2020, publisher: World Health Organization. [Online]. Available: <https://cutt.ly/tkrihlz>
- [4] —, "WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020," Mar. 2020, publisher: World Health Organization. [Online]. Available: <https://cutt.ly/KkrikaE>
- [5] "WHO COVID-19 Dashboard," 2020. [Online]. Available: <https://covid19.who.int/>
- [6] Y. M. Alsofayan, S. M. Althunayyan, A. A. Khan, A. M. Hakawi, and A. M. Assiri, "Clinical characteristics of covid-19 in Saudi Arabia: A national retrospective study," *Journal of Infection and Public Health*, vol. 13, no. 7, pp. 920–925, 2020. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1876034120304925>
- [7] World Health Organization, "Saudi Arabia: WHO Coronavirus Disease (COVID-19) Dashboard," publisher: World Health Organization. [Online]. Available: <https://covid19.who.int>
- [8] A. Khan, S. Althunayyan, Y. Alsofayan, R. Alotaibi, A. Mubarak, M. Arafat, A. Assiri, and H. Jokhdar, "Risk factors associated with worse outcomes in COVID-19: a retrospective study in Saudi Arabia," *Eastern Mediterranean Health Journal*, vol. 26, no. 11, pp. 1371–1380, Nov. 2020. [Online]. Available: <https://applications.emro.who.int/emhvj/v26/11/1020-3397-2020-2611-1371-1380-eng.pdf>
- [9] Ministry of Health, KSA, "MOH Publications - Saudi Arabia's Experience in Health Preparedness and Response to COVID-19 Pandemic," The Saudi Ministry of Health, Riyadh, Saudi Arabia, Tech. Rep., Aug. 2020. [Online]. Available: <https://www.moh.gov.sa/en/Ministry/MediaCenter/Publications/Pages/Publications-2020-10-27-001.aspx>
- [10] World Health Organization, "Operational planning guidance to support country preparedness and response," World Health Organization, Geneva, Switzerland, Tech. Rep., May 2020. [Online]. Available: <https://cutt.ly/pkriWU0>
- [11] J. Gallagher, "Coronavirus: What we still don't know about Covid-19," *BBC News*, Mar. 2020. [Online]. Available: <https://www.bbc.com/news/health-52006988>
- [12] C. Heneghan, J. Brassey, and T. Jefferson, "COVID-19: What proportion are asymptomatic?" Apr. 2020. [Online]. Available: <https://www.cebm.net/covid-19/covid-19-what-proportion-are-asymptomatic/>
- [13] Centers for Disease Control and Prevention, "Coronavirus Disease 2019 (COVID-19) – Symptoms," Dec. 2020. [Online]. Available: <https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html>
- [14] G. Meyerowitz-Katz and L. Merone, "A systematic review and meta-analysis of published research data on COVID-19 infection-fatality rates," *Epidemiology*, preprint, May 2020. [Online]. Available: <http://medrxiv.org/lookup/doi/10.1101/2020.05.03.20089854>
- [15] Ministry of Health, KSA, "COVID-19 Dashboard: Saudi Arabia," 2021. [Online]. Available: <https://covid19.moh.gov.sa/>
- [16] O. Theobald, *Machine learning for absolute beginners: a plain English introduction*. Scatterplot Press, 2018.
- [17] F. Rustam, A. A. Reshi, A. Mehmood, S. Ullah, B.-W. On, W. Aslam, and G. S. Choi, "COVID-19 Future Forecasting Using Supervised Machine Learning Models," *IEEE Access*, vol. 8, pp. 101 489–101 499, 2020. [Online]. Available: <https://ieeexplore.ieee.org/document/9099302/>
- [18] F. Petropoulos and S. Makridakis, "Forecasting the novel coronavirus COVID-19," *PLOS ONE*, vol. 15, no. 3, p. e0231236, Mar. 2020. [Online]. Available: <https://dx.plos.org/10.1371/journal.pone.0231236>
- [19] G. Grasselli, A. Pesenti, and M. Cecconi, "Critical Care Utilization for the COVID-19 Outbreak in Lombardy, Italy: Early Experience and Forecast During an Emergency Response," *JAMA*, vol. 323, no. 16, p. 1545, Apr. 2020. [Online]. Available: <https://jamanetwork.com/journals/jama/fullarticle/2763188>
- [20] G. C. Calafiore, C. Novara, and C. Possieri, "A Modified SIR Model for the COVID-19 Contagion in Italy," *arXiv:2003.14391 [physics]*, Mar. 2020, arXiv: 2003.14391. [Online]. Available: <http://arxiv.org/abs/2003.14391>
- [21] I. Nesteruk, "Statistics-based predictions of coronavirus 2019-nCoV spreading in mainland China," *Epidemiology*, preprint, Feb. 2020. [Online]. Available: <http://medrxiv.org/lookup/doi/10.1101/2020.02.12.20021931>
- [22] N. Ferguson, D. Laydon, G. Nedjati Gilani, N. Imai, K. Ainslie, M. Baguelin, S. Bhatia, A. Boonyasiri, Z. Cucunuba Perez, G. Cuomo-Dannenburg, A. Dighe, I. Dorigatti, H. Fu, K. Gaythorpe, W. Green, A. Hamlet, W. Hinsley, L. Okell, S. Van Elsland, H. Thompson, R. Verity, E. Volz, H. Wang, Y. Wang, P. Walker, P. Winskill, C. Whittaker, C. Donnelly, S. Riley, and A. Ghani, "Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand," Mar. 2020. [Online]. Available: <http://spiral.imperial.ac.uk/handle/10044/1/77482>

- [23] S. L. Chang, N. Harding, C. Zachreson, O. M. Cliff, and M. Prokopenko, "Modelling transmission and control of the COVID-19 pandemic in Australia," *Nature Communications*, vol. 11, no. 1, p. 5710, Nov. 2020. [Online]. Available: <https://www.nature.com/articles/s41467-020-19393-6>
- [24] B. Wilder, M. Charpignon, J. Killian, H.-C. Ou, A. Mate, S. Jabbari, A. Perrault, A. Desai, M. Tambe, and M. S. Majumder, "the role of age distribution and family structure on Covid-19 dynamics: A preliminary modeling assessment for Hubei and Lombardy," 2020.
- [25] M. A. Ruiz Estrada and E. Koutrouas, "The Networks Infection Contagious Diseases Positioning System (NICDP-System): The Case of Wuhan-COVID-19," *SSRN Electronic Journal*, 2020. [Online]. Available: <https://www.ssrn.com/abstract=3548413>
- [26] R. Verity, L. C. Okell, I. Dorigatti, P. Winskill, C. Whittaker, N. Imai, G. Cuomo-Dannenburg, H. Thompson, P. G. T. Walker, H. Fu, A. Dighe, J. T. Griffin, M. Baguelin, S. Bhatia, A. Boonyasiri, A. Cori, Z. Cucunubá, R. FitzJohn, K. Gaythorpe, W. Green, A. Hamlet, W. Hinsley, D. Laydon, G. Nedjati-Gilani, S. Riley, S. van Elsland, E. Volz, H. Wang, Y. Wang, X. Xi, C. A. Donnelly, A. C. Ghani, and N. M. Ferguson, "Estimates of the severity of coronavirus disease 2019: a model-based analysis," *The Lancet Infectious Diseases*, vol. 20, no. 6, pp. 669–677, Jun. 2020. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S1473309920302437>
- [27] C. Anastassopoulou, L. Russo, A. Tsakris, and C. Siettos, "Data-based analysis, modelling and forecasting of the COVID-19 outbreak," *PLOS ONE*, vol. 15, no. 3, p. e0230405, Mar. 2020. [Online]. Available: <https://dx.plos.org/10.1371/journal.pone.0230405>
- [28] Q. Lin, S. Zhao, D. Gao, Y. Lou, S. Yang, S. S. Musa, M. H. Wang, Y. Cai, W. Wang, L. Yang, and D. He, "A conceptual model for the coronavirus disease 2019 (COVID-19) outbreak in Wuhan, China with individual reaction and governmental action," *International Journal of Infectious Diseases*, vol. 93, pp. 211–216, Apr. 2020. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S120197122030117X>
- [29] J. M. Read, J. R. Bridgen, D. A. Cummings, A. Ho, and C. P. Jewell, "Novel coronavirus 2019-nCoV: early estimation of epidemiological parameters and epidemic predictions," *Infectious Diseases (except HIV/AIDS)*, preprint, Jan. 2020. [Online]. Available: <http://medrxiv.org/lookup/doi/10.1101/2020.01.23.20018549>
- [30] J. T. Wu, K. Leung, and G. M. Leung, "Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study," *The Lancet*, vol. 395, no. 10225, pp. 689–697, Feb. 2020. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0140673620302609>
- [31] C. Iwendi, A. K. Bashir, A. Peshkar, R. Sujatha, J. M. Chatterjee, S. Pasupuleti, R. Mishra, S. Pillai, and O. Jo, "COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm," *Frontiers in Public Health*, vol. 8, p. 357, Jul. 2020. [Online]. Available: <https://www.frontiersin.org/article/10.3389/fpubh.2020.00357/full>
- [32] S. Kutia, S. H. Chauhdary, C. Iwendi, L. Liu, W. Yong, and A. K. Bashir, "Socio-technological factors affecting user's adoption of ehealth functionalities: A case study of China and Ukraine ehealth systems," *IEEE Access*, vol. 7, pp. 90777–90788, 2019. [Online]. Available: <https://ieeexplore.ieee.org/document/8744204/>
- [33] S. F. Ardabili, A. Mosavi, P. Ghamisi, F. Ferdinand, A. R. Varkonyi-Koczy, U. Reuter, T. Rabczuk, and P. M. Atkinson, "COVID-19 Outbreak Prediction with Machine Learning," *Health Informatics*, preprint, Apr. 2020. [Online]. Available: <http://medrxiv.org/lookup/doi/10.1101/2020.04.17.20070094>
- [34] L. Yan, H.-T. Zhang, J. Goncalves, Y. Xiao, and M. Wang, "A machine learning-based model for survival prediction in patients with severe COVID-19 infection," *Epidemiology*, preprint, Mar. 2020. [Online]. Available: <http://medrxiv.org/lookup/doi/10.1101/2020.02.27.20028027>
- [35] G. Pinter, I. Felde, A. Mosavi, P. Ghamisi, and R. Gloaguen, "COVID-19 Pandemic Prediction for Hungary; A Hybrid Machine Learning Approach," *Mathematics*, vol. 8, no. 6, p. 890, Jun. 2020. [Online]. Available: <https://www.mdpi.com/2227-7390/8/6/890>
- [36] F. A. B. Hamzah, C. H. Lau, H. Nazri, D. V. Ligo, G. Lee, C. L. Tan, M. K. B. M. Shaib, U. H. B. Zaidon, A. B. Abdullah, M. H. Chung, C. H. Ong, and P. Y. Chew, "CoronaTracker: World-wide COVID-19 Outbreak Data Analysis and Prediction," *nCoV*, preprint, Mar. 2020. [Online]. Available: http://www.who.int/bulletin/online_first/20-255695.pdf
- [37] A. C. Müller and S. Guido, *Introduction to machine learning with Python: a guide for data scientists*, 1st ed. Sebastopol, CA: O'Reilly Media, Inc, 2016.
- [38] S. V. Stehman, "Selecting and interpreting measures of thematic classification accuracy," *Remote Sensing of Environment*, vol. 62, no. 1, pp. 77 – 89, 1997. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0034425797000837>
- [39] K. W. Bowyer, N. V. Chawla, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *CoRR*, vol. abs/1106.1813, 2011. [Online]. Available: <http://arxiv.org/abs/1106.1813>
- [40] R. Pal, A. A. Sekh, S. Kar, and D. K. Prasad, "Neural network-based country wise risk prediction of COVID-19," *Applied Sciences*, vol. 10, no. 18, p. 6448, Sep. 2020, arXiv: 2004.00959. [Online]. Available: <http://arxiv.org/abs/2004.00959>
- [41] Z. Tang, W. Zhao, X. Xie, Z. Zhong, F. Shi, J. Liu, and D. Shen, "Severity Assessment of Coronavirus Disease 2019 (COVID-19) Using Quantitative Features from Chest CT Images," *arXiv:2003.11988 [cs, eess]*, Mar. 2020, arXiv: 2003.11988. [Online]. Available: <http://arxiv.org/abs/2003.11988>