Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 12, No. 1, December : 2023

EARLY STAGE IDENTIFICATION AND PREDICTION OF COVID-19 PATIENTS USING MACHINE LEARNING

Shanaaya, B.Tech CSE-AI, Dept. of Data Science and Artificial Intelligence, IGDTUW, New Delhi, India, shanaaya018btcsai21@igdtuw.ac.in.
Navita, Research Scholar, Department of Computer Science & Applications, Maharshi Dayanand University, Rohtak, Haryana, India, Navitamehra55@gmail.com.
Bal Kishan, Assistant Professor, Department of Computer Science & Applications, Maharshi Dayanand University, Rohtak, Haryana, India, Balkishan246@gmail.com.

Abstract
The COVID-19 pandemic has significantly altered human lifestyles, by affecting various aspects of daily life such as culture, education, business, social interactions, and marketing activities within limited boundaries. This global health crisis has placed immense strain on healthcare systems worldwide, prompting a need for innovative solutions to navigate the challenges it presents. In response to this pandemic, various technologies including Machine Learning (ML), Artificial Intelligence (AI) and the Internet of Things (IoT) have emerged as potential solutions. The current work explores the efficiency of ML techniques in the early identification and prediction of COVID-19. The current work proposed a model to effectively analyze the dataset collected from different sources in order to identify and predict whether the patient is infected by COVID-19 or not. The proposed model works with the four most important machine learning techniques named logistic regression, k-nearest neighbor, XGBoost, random forest tree, and Bayesian as an optimization technique. The results reveal that among all machine learning techniques, RFT achieved the best ROC_AUC score of 97.43 with the lowest MSE of 2.203.

Keywords: Machine Learning (ML), Internet of Things (IoT), COVID-19.

I. Introduction

The global epidemic caused by the new Coronavirus of 2019, commonly known as COVID-19, has triggered widespread panic worldwide. Compared to the 1918 H1N1 influenza pandemic, COVID-19 stands out as one of the most destructive and hazardous outbreaks. Identified as the primary source of COVID-19, the "Severe Acute Respiratory Syndrome Coronavirus Two (SARS-CoV-2)" is implicated [1]. The first patient of COVID-19 was tracked in December 2019 in Wuhan, China[2]. World Health Organization reported that there have been 100,075 deaths and 15,225,252 reported cases reported up to April 10, 2020. This underscores the swift and widespread progression of COVID-2019 since the onset of December 2020. The impact of COVID-2019 has now reached 172 countries.

The indicators of COVID-19 exhibit significant variability, ranging from no discernible signs to severe and potentially life-threatening symptoms. Individuals can be either symptomatic or asymptomatic. Cough, fever, headaches, breathing challenges, fatigue, no taste and no smell are typical symptoms that appear among COVID-19 patients [3]**.** Unfortunately, currently, no proven treatment or vaccination still exists. The sole method the world can employ to control this situation is to minimize its spread by utilizing approaches such as social isolation, hand hygiene, and the use of face masks. Nevertheless, technology could play a crucial role in mitigating its spread through early prediction or detection of new cases and assist in monitoring its progression[4]**.** Through the implementation of lockdowns and effective strategies, the infection rate saw a decline in September 2021. However, this did not mark the conclusion of the COVID-19 challenges. In March 2021, the second wave proved to be significantly more devastating than the initial wave. Various regions in the country grappled with shortages of essential resources such as vaccines, doctors, hospital beds, oxygen cylinders, and other

healthcare services [5]**.** At the April end, India recorded the highest number of active cases globally. On April 30, 2021, India became the first country to document over 400,000 new cases within 24 hours[6]. As of March 2022, this figure decreased to 21,530. From the start of this pandemic, the "World Health Organization (WHO) "has been actively engaged in identifying effective tools and has globally launched the "COVID-19 vaccine" on January 16, 2021. India initiated its "vaccination program" on the same date, administering both the "AstraZeneca vaccine (Covishield)" and the native "Covaxin" The primary objective of the vaccine is to safeguard against "severe acute respiratory syndrome coronavirus2 (SARS-CoV-2)," the virus responsible for causing COVID-19. However, till now, there is still no effective treatment or safety against it.

In the future, the virus could potentially mutate into different forms, creating critical situations that would significantly impact the healthcare sector. To mitigate the strain on medical facilities, governments, and healthcare institutions may collaborate with smart and intelligent technologies for the early detection and treatment of cases with a higher likelihood of survival. This strategic approach would enable highly efficient utilization of inadequate medical resources for effective treatments[7]**.**

The key objective of this research is to identify the patient suffering from COVID-19 as well as to make predictions about the patient's health status based on their health symptoms. Moreover, this study also explores the impact of COVID-19, preliminary clinical observation, vital signs, and demographic features to predict patient health status.

The paper is organized as follows. Section 2 details the work done by different researchers in this domain. The materials and the methodology followed in this work are described in Section 3. Section 4 details about experimental setups and results. Finally, the conclusion of this work is described in Section 5.

II. Related Work

ML techniques play a crucial role in preventing the spread of the COVID-19 pandemic by offering early identification and prediction of this disease. This involves the analysis of various types of data including X-rays, clinical data, CT scans blood samples etc.

Yan et al., 2020 [8] predict the survival rate of patients suffering from severe COVID-19 through the analysis of various risk factors and available data. The author works with the XGBoost (XGB) classifier and identifies three main risk factors. The proposed model makes predictions about the risk of death with 0.90 accuracy and 0.95 precision. Such a model acts as an effective tool for clinicians in identifying critical and helping them to reduce the mortality rate. The proposed model was implemented only on a small dataset of 29 records only which may affect the accuracy of results.

Yao et al., 2020 [9]employed the SVM model to classify patients based on the severity of their symptoms. The SVM was applied for binary classification using a dataset of 137 records, encompassing both highly severe patients and those with mild symptoms. This dataset included information from urine and blood test results. The results highlighted approximately 32 factors with significant correlations to severe patients, achieving an accuracy of 0.815. Notably, among these factors, age, and gender played pivotal roles in the classification of patients between mild and severe patients. Patients near the age of 65 exhibited a higher likelihood of experiencing severe cases than their counterparts, and male patients faced an elevated risk of developing severe COVID-19 symptoms. Regarding urine and blood test samples, the study emphasized that features derived from blood test results exhibited more notable differences among severe and mild cases compared to features from urine test results.

Hu et al., 2021 [10] utilized the LR (logistic regression) model to detect the severity among patients infected by COVID-19. Their dataset comprised clinical as well as demographic data from 115 individuals with nonsevere conditions and 68 individuals with severe conditions. High sensitivity, age, lymphocyte count, C-reactive protein level, and d-dimer level—were identified as the most important

features in distinguishing between mild and severe cases. The model's evaluation demonstrated effective prediction with an AUC_ROC score of 0.881, 0.839 of sensitivity, and 0.794 of specificity.
In a study conducted by [11], a sample of 3927 COVID-19 patients was used for predicting the risk of mortality using the XGBoost (XGB) model. The proposed model attained an accuracy score of 0.85 and 0.90 of the AUC_ROC score. Additionally, [12]developed a mortality prediction model using the LR approach and implemented it on data collected from 1969 individuals affected by COVID-19. Their study identified O2 and age levels as significant attributes, achieving an accuracy of 0.89, sensitivity of 0.82, and specificity of 0.81.
From the literature study, it is found that machine learning plays a significant role in the effective analysis of data and for better prediction. While several investigations have been conducted to predict and forecast outcomes, there remains a necessity for further exploration and expansion of findings related to COVID-19, leveraging real datasets from clinical records.
The key objective of this study is to develop a model for early identification and prediction of COVID-19 patients based on their major health symptoms.

III. Materials and Methodology

This section will cover the dataset description and the methodology used.

3.1 Data Set Description

This study was conducted on a "symptom-based dataset" recorded from different individuals and is accessible from Kaggle. The collected symptoms adhere to the standard outlined by the "World Health Organization (WHO)" to ascertain the presence of COVID-19 in individuals. Biosensors-based medical devices like smart temperature recorder is used to collect the dataset encompassing a total of 21 attributes. Table 1 provides detailed information about each attribute.

Table 1.COVID-19 Symptom Dataset Description

Attribute No.	Attribute Name
1	Breathing Problem
2	Dry Cough
3	Fever
4	Running Nose
5	Sore throat
6	Asthma
7	Headache
8	Chronic Lung Disease
9	Heart Disease
10	Diabetes
11	Hyper Tension
12	Gastrointestinal
13	Fatigue
14	Abroad travel
15	Attended Large Gathering
16	Contact with COVID Patient
17	Family working in Public Exposed Places
18	Visited Public Exposed Places

The dataset comprises 20 variables that potentially influence the prediction of COVID-19, and one class variable indicating the presence and absence of COVID-19.

3.2 Data Preprocessing

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 12, No. 1, December : 2023

The process of transforming data into a comprehensible format is called data preprocessing. Real-world data generally contains noise, and missing values or may contain incomplete data that prevent its direct use in machine learning models. Data preprocessing is a crucial step in which we clean and adapt the data to make it suitable for machine learning models, thereby enhancing accuracy and efficiency. The key steps followed during data preprocessing are:

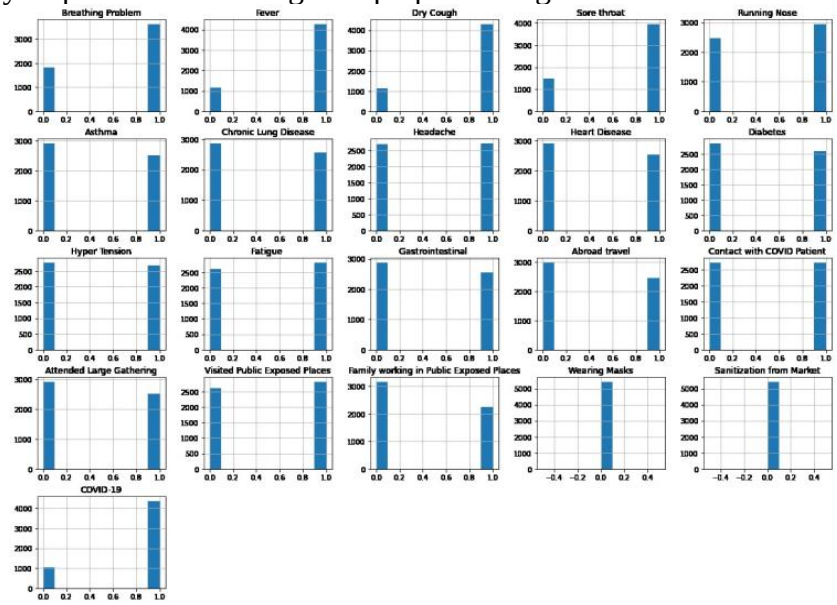


Figure 1. A HistMap Representation of All Associated Features of COVID-19

1.Feature Removal: After analyzing Figure 1, we reveal that "wearing masks" and "sanitization from the market" have a single constant value, 'no.' Since these features do not affect prediction, they can be dropped from the dataset.

2.Data Encoding: Label encoding is widely used for managing categories by assigning each label a numerical value depending on alphabetical order. The given dataset contains all the attributes in the form of 'yes' or 'no'. Therefore, label encoding was applied to translate them to 0 and 1, facilitating better understanding by the machine learning model.

3.Splitting the dataset: The next step is to split the dataset into train and test datasets. The given dataset is partitioned into an 80:20 ratio which means 80% data is used for training the model and 20% of the data is used for testing the model. All independent attributes (20 attributes) into X and independent attribute COVID-19 into Y to predict whether the patient has COVID or not.

3.3Prediction Models

In the current study, four machine learning techniques namely random forest, logistic regression (LR), k-nearest neighbor, and extreme gradient boosting (XGB) were used. Each technique is briefly described as follows:

1. Logistic Regression: Logistic Regression (LR) is frequently employed to establish associations between dependent and independent categorical variables. A dependent variable with only two values (like 0 and 1, true and false, or yes and no) is referred to as binary logistic regression. If a dependent variable contains more than two values, it is termed multinomial logistic regression. A mathematical expression used by LR to predict the changeability of the dependent variable, allowing for the estimation of probabilities associated with different outcomes based on the values of independent variables [13].

2. K-Nearest Neighbor: K-NN (K-Nearest Neighbors) is a straightforward and user-friendly model utilized for both classification and regression tasks. It relies on the k Neighbor's approach, where data classification is determined by measuring similarity. The algorithm assesses the imbalance among

existing and new data points in the sample space, assigning the new data to a class with similar data [14].

3. Random Forest: Random Forest (RF) is predominantly employed to address complex problems and is constructed by integrating multiple decision trees. It finds application in both classification and regression tasks. RF utilizes the bagging method to generate results, with the final decision being determined through majority voting across individual trees. This methodology is particularly effective in mitigating the overfitting problem associated with decision tree.

4. XGBoost: The Extreme Gradient Boosting algorithm (XGBoost) is a highly efficient ensemble learning method capable of managing missing values and aggregating weak predictors to create a more effective model [15]. It applies to both classification and prediction tasks. XGBoost achieves loss function reduction through the implementation of the gradient descent method for optimizing the objective function. Notably, XGBoost mitigates overfitting concerns by leveraging a set of learners to construct a robust model, contributing to reduced runtime. Its flexibility and efficiency have led to its widespread adoption, making it a popular choice in numerous successful data mining competitions.

5. Bayesian optimization: Various ML methods, including Logistic Regression, K-Nearest Neighbor, Random Forest, and XGBoost include many hyperparameters that must be chosen. The selection of hyperparameter values significantly influences the performance of these machine-learning models. Frequently used optimization methods include random search, grid search, and Bayesian optimization in the literature but the current work utilizes Bayesian optimization (BO) for finding optimal hyperparameters. Notably, the BO algorithm stands out as an efficient and effective global optimization approach that is designed upon the principles of Gaussian processes and Bayesian inference [16]**.** A key advantage associated with BO is to reduce the time required to reach the optimal parameter set by leveraging information from past evaluations when selecting the next set of hyperparameters.

IV. Experimental Results and Analysis

A detailed description of experimental results is given in this section. The implementation of the proposed model is carried out on Jupyter Notebook using Python as a programming language. The general procedure followed for the prediction of patients infected by COVID-19 is shown in Figure 2. The performance of the proposed model was evaluated using accuracy, precision, specificity, sensitivity, and F1-score, MSE (Mean Squared Error) as the standard evaluation measures. Additionally, the “area under the curve (AUC) and the receiver operating characteristic (ROC)” were employed to compare the classifiers. The ROC analysis is a widely utilized method for examining the trade-off between true-positive (sensitivity) and false-positive rate (specificity) in diagnostic tests. MSE measures the difference between the anticipated class value and the actual class. It provides insight into how accurately the constructed model predicts the labels of the samples in contrast to the original values in the dataset. A model with a low MSE is found to be more effective compared to one with a higher MSE. The experimentation results achieved through the implementation are shown in Table 3 in the form of different performance evaluation measures.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 12, No. 1, December : 2023

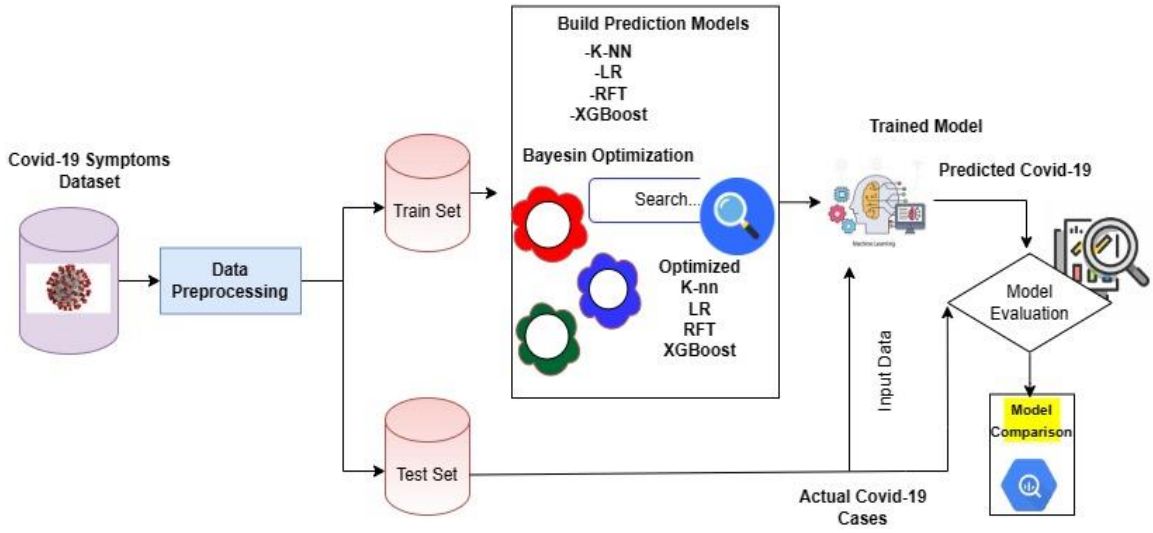


Figure 2. Model Used for Identification and Prediction of COVID-19 Patient

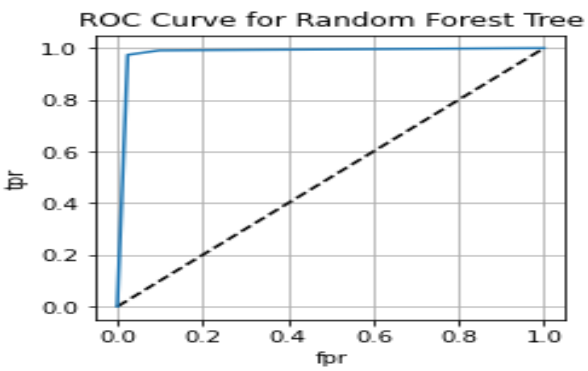
The optimized set of hyperparameters selected through Bayesian Optimization methods is depicted in Table 2.

Table 2. Optimal Parameter by Using Bayesian Optimization Method

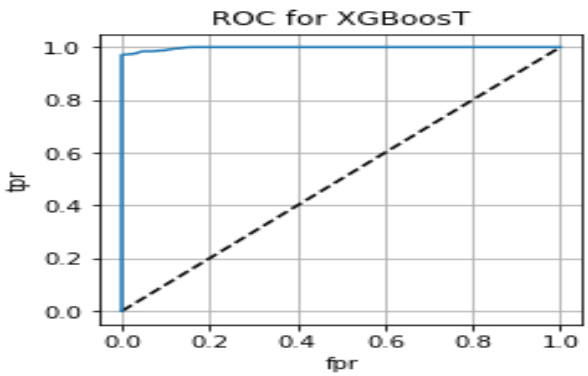
Techniques	Optimized Parameter
RFT	N_estimators=150, Min_samples_leaf= 1, Min_samples_split =5 , Max_depth= 15,
XGBoost	max_depth =3, learning_rate 0.05, max_features =0.5, random_state 40
K-NN	N_neighbors=7, leaf_size=30, metric='minkowski'
LR	Var_smoothing=0.2234035678

Table 3. Performance Evaluation Results of ML Models in Prediction of Patients Suffering from COVID-19

Techniques	Accuracy	Precision	Recall	F1_Score	ROC_AUC	MSE
RFT	0.983	0.96	0.97	0.96	0.9743	2.203
K-NN	0.98	0.95	0.97	0.96	0.969	2.257
XGBoost	0.97	0.95	0.96	0.97	0.971	2.483
LR	0.96	0.97	0.93	0.95	0.932	3.035

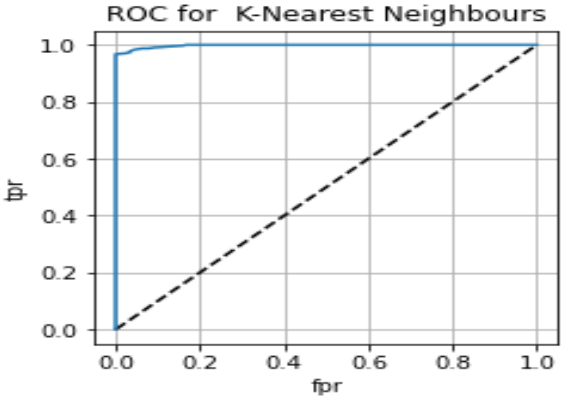


(a)

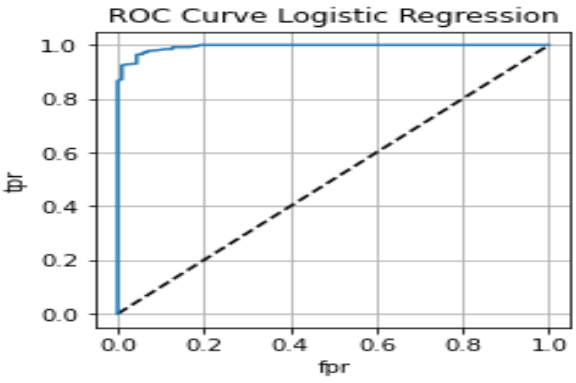


(b)

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 12, No. 1, December : 2023



(c)



(d)

Figure 3. ROC_AUC Curve Representation for Different ML Techniques

Performance results of different ML techniques are depicted in Figure 3. The performance results show that amongst all ML techniques, RFT achieved the highest ROC_AUC score of 97.43% with the lowest MSE of 2.203. Followed by XGBoost with a 97.1% ROC_AUC score with 2.483 MSE. Finally from the experimental outcomes it is observed that RFT achieved the best accuracy score with the lowest MSE.

```
COVID PREDICTION BASED ON ML ALGORITHMS
Enter 1 for Yes and 0 for No
Does the patient have breathing problem ? 1
Does the patient have fever ? 1
Does the patient have dry cough ? 0
Does the patient have sore throat ? 1
Does the patient have running nose ? 0
Does the patient have any record of asthma ? 1
Does the patient have any records of chronic lung disease ? 1
Is the patient having headche ? 1
Does the patient have any record of any heart disease ? 1
Does the patient have diabetes ? 1
Does the patient have hyper tension ? 1
Does the patient experience fatigue ? 0
Does the patient have any gastrointestinal disorders ? 0
Has the patient travelled abroad recently ? 0
Was the patient in contact with a covid patient recently ? 0
Did the patient attend any large gathering event recently ? 0
Did the patient visit any public exposed places recently ? 0
Does the patient have any family member working in public exposed places ? 0

Results :  [1]
You may be affected with COVID-19 virus! Please get RTPCR test ASAP and stay in Quarantine for 14 days!
```

Figure 4. Covid-19 Prediction by Machine Learning Model Based on Given Input

Figure 4 shows how the presented model predicts the patient based on their given input. The model asks the user to make entries of their health symptoms. In the above-stated example, the model predicts that the patient is affected by COVID-19 and an alert is given to the patient that he/she must proceed to “RTPCR test ASAP and stay in Quartine for 14 days”.

V. Conclusion

Accurate prediction of COVID-19 is highly effective to slow down the transmission of this pandemic. Highly relevant information provided to healthcare professionals assists them in making effective decisions about patient health status as well as managing available healthcare resources and staff. This work aim to develop and effective data driven model to predict the patient health status. This paper introduces a model with optimized hyperparameters via Bayesian optimization to predict whether the patient is affected by COVID-19 or not. The model includes the most promising machine learning techniques like k-nearest neighbor, random forest, XGBoost and logistic regression to make predictions. The results showed that the optimized RF model outperformed other models and achieved

the highest AUC_ROC score of 97.43 with the lowest MSE of 2.203. As the developed model is limited to symptoms datasets, we plan to develop a more efferent forecasting model that considers more relevant COVID-19 data. Additional information or diagnoses from hospital records, individuals who have contracted the virus, COVID-19 survivors, and patients undergoing assessment, or management can all be incorporated for future research.

References

[1] Vedaraj, M.; Saravanan, K.; Srinivasan, V. P.; Balachander, K.; Jaithunbi, A. K. Prevalence and Characteristics of Fever in Adult and Pediatric Patients With. ijhs 2022, 9467–9474.
[2] Joshi, R. K.; Mehendale, S. M. Prevention and Control of COVID-19 in India: Strategies and Options. Medical Journal Armed Forces India 2021, 77, S237–S241.
[3] Saniasiaya, J.; Islam, M. A.; Abdullah, B. Prevalence and Characteristics of Taste Disorders in Cases of COVID-19: A Meta-analysis of 29,349 Patients. Otolaryngol.--head neck surg, 2021, 165 (1), 33–42.
[4] Kelly, M. P. Digital Technologies and Disease Prevention. American Journal of Preventive Medicine, 2016, 51 (5), 861–863.
[5] Safi, M. India's Shocking Surge in Covid Cases Follows Baffling Decline. The Guardian. Retrieved 2021.
[6] Medina, J.; Espinilla, M.; García-Fernández, Á. L.; Martínez, L. Intelligent Multi-Dose Medication Controller for Fever: From Wearable Devices to Remote Dispensers. Computers & Electrical Engineering, 2018, 65, 400–412.
[7] Vinod, D. N.; Prabaharan, S. R. S. COVID-19-The Role of Artificial Intelligence, Machine Learning, and Deep Learning: A Newfangled. Arch Computat Methods Eng. , 2023, 30 (4), 2667–2682.
[8] Yan, L.; Zhang, H.-T.; Goncalves, J.; Xiao, Y.; Wang, M.; Guo, Y.; Sun, C.; Tang, X.; Jin, L.; Zhang, M.; Huang, X.; Xiao, Y.; Cao, H.; Chen, Y.; Ren, T.; Wang, F.; Xiao, Y.; Huang, S.; Tan, X.; Huang, N.; Jiao, B.; Zhang, Y.; Luo, A.; Mombaerts, L.; Jin, J.; Cao, Z.; Li, S.; Xu, H.; Yuan, Y. A Machine Learning-Based Model for Survival Prediction in Patients with Severe COVID-19 Infection; preprint; Epidemiology, 2020.
[9] Yao, H.; Zhang, N.; Zhang, R.; Duan, M.; Xie, T.; Pan, J.; Peng, E.; Huang, J.; Zhang, Y.; Xu, X.; Xu, H.; Zhou, F.; Wang, G. Severity Detection for the Coronavirus Disease 2019 (COVID-19) Patients Using a Machine Learning Model Based on the Blood and Urine Tests. Front. Cell Dev. Biol. 2020, 8, 683.
[10] Hu, C.; Liu, Z.; Jiang, Y.; Shi, O.; Zhang, X.; Xu, K.; Suo, C.; Wang, Q.; Song, Y.; Yu, K.; Mao, X.; Wu, X.; Wu, M.; Shi, T.; Jiang, W.; Mu, L.; Tully, D. C.; Xu, L.; Jin, L.; Li, S.; Tao, X.; Zhang, T.; Chen, X. Early Prediction of Mortality Risk among Patients with Severe COVID-19, Using Machine Learning. International Journal of Epidemiology 2021, 49 (6), 1918–1929.
[11] Bertsimas, D.; Lukin, G.; Mingardi, L.; Nohadani, O.; Orfanoudaki, A.; Stellato, B.; Wiberg, H.; Gonzalez-Garcia, S.; Parra-Calderón, C. L.; Robinson, K.; Schneider, M.; Stein, B.; Estirado, A.; A Beccara, L.; Canino, R.; Dal Bello, M.; Pezzetti, F.; Pan, A.; The Hellenic COVID-19 Study Group. COVID-19 Mortality Risk Assessment: An International Multi-Center Study. PLoS ONE 2020, 15 (12), e0243262.
[12] Pourhomayoun, M.; Shakibi, M. Predicting Mortality Risk in Patients with COVID-19 Using Machine Learning to Help Medical Decision-Making. Smart Health 2021, 20, 100178.
[13] Hosmer, D. W.; Lemeshow, S. Applied Logistic Regression, 1st ed.; Wiley, 2000.
[14] Zhang, Z. Introduction to Machine Learning: K-Nearest Neighbors. Ann. Transl. Med. 2016, 4 (11), 218–218.

[15] Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016; pp 785–794.
[16] Nguyen, V.-H.; Le, T.-T.; Truong, H.-S.; Le, M. V.; Ngo, V.-L.; Nguyen, A. T.; Nguyen, H. Q. Applying Bayesian Optimization for Machine Learning Models in Predicting the Surface Roughness in Single-Point Diamond Turning Polycarbonate. Mathematical Problems in Engineering 2021, 2021, 1–16.