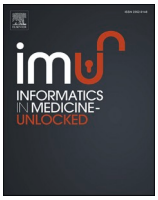
Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Informatics in Medicine Unlocked 30 (2022) 100933

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu



Supervised learning of COVID-19 patients' characteristics to discover symptom patterns and improve patient outcome prediction

Sadegh Ilbeigipour *, Amir Albadvi

Department of Information Technology Engineering, Industrial and Systems Engineering Faculty, Tarbiat Modares University, Tehran, Iran

ARTICLE INFO

Keywords:
COVID-19
Supervised learning
Machine learning
Confidence index
Support index
Association rules mining

ABSTRACT

The world today faces a new challenge that is unprecedented in the last 100 years. The emergence of a new coronavirus has led to a human catastrophe. Scientists in various sciences have been looking for solutions to this problem so far. In addition to general vaccination, maintaining social distance and adherence to government guidelines on safety precaution measures are the most well-known strategies to prevent COVID-19 infection. In this research, we tried to examine the symptoms of COVID-19 cases through different supervised machine learning methods. We solved the class imbalance problem using the synthetic minority over-sampling (SMOTE) method and then developed some classification models to predict the outcome of COVID-19 cases (recovery or death). Besides, we implemented a rule-based technique to identify different combinations of variables with specific ranges of their values that together affect disease severity. Our results showed that the random forest model with 95.6% accuracy, 97.1% sensitivity, 94.0% specification, 94.4% precision, 95.8% F-score, and 99.3% AUC-score outperforms state-of-the-art classification models. Finally, we identified the most significant rules that state various combinations of 6 features in certain ranges of their values lead to patients' recovery with a confidence value of 90%. In conclusion, the classification results in this study show better performance than recent studies, and the extracted rules help physicians consider other important factors to improve health services and medical decision-making for different groups of COVID-19 patients.

1. Introduction

In late 2019, the world suffered a great menace that influenced all viewpoints of human life. A new type of coronavirus has appeared in Wuhan, China [1]. The virus causes lung infection and has a very high rate of transmission [2]. Finally, with the extent of the virus in many countries, the World Health Organization on March 11, 2020, announced the prevalence of novel coronavirus as a fatal pandemic [3].

There are various types of coronaviruses in the world. Acute Respiratory Syndrome and Middle East Respiratory Syndrome are the most famous of these viruses [4]. The recently recognized virus is called SARS-COV-2 and is the object of COVID-19 disease [4].

Several efficient COVID-19 vaccines have been produced by well-known companies around the world so far. But, masking, adherence to government guidelines on safety precaution measures, and social distancing are still the three most effective actions in preventing COVID-19 [4]. However, computer science applications are an effective way to help physicians cope with coronavirus and improve hospital care. These applications generally include machine learning and deep learning techniques that are used to detect the patients with the disease or mathematical modeling and social network analysis (SNA) approaches to forecast the pattern of disease outbreaks. For example, regression and susceptible-exposed-infected-recovered (SEIR) analysis have been applied to predict the future prevalence of the disease [5]. Gene expression programming (GEP) and genetic algorithm (GA) were utilized to optimize the diameter of Nylon-6,6 nanofibers for coronavirus face masks [6]. Furthermore, some techniques based on mathematical modeling have been developed to forecast the spread rate of the disease [7–9]. Today, in addition to short-term estimating of the prevalence pattern and assessing the risk of infection with the COVID-19 [10,11], real-time applications are increasing in various fields of medicine, such as the diagnosis of cardiac arrhythmias [12]. Also, the social network analysis [13,14] and the partial correlation coefficients [15] approaches have been used to recognize high-risk areas of the disease and identify effective climatology factors on the prevalence of COVID-19, respectively. As a last attempt, researchers have employed different deep learning methods to distinguish positive cases based on chest X-ray and CT-Scan images [16–23]. However, although deep artificial neural

* Corresponding author.
E-mail addresses: i_sadegh@modares.ac.ir (S. Ilbeigipour), Albadvi@modares.ac.ir (A. Albadvi).

https://doi.org/10.1016/j.imu.2022.100933
Received 4 January 2022; Received in revised form 26 March 2022; Accepted 26 March 2022
Available online 12 April 2022

networks have a high capability in processing image and audio data, these methods greatly increase computational costs [23].

The previously presented methods in diagnosing the outcome of COVID_19 cases suffer from two problems. First, the model evaluation measures in this research show the relatively low performance of the classification models. Second, they do not provide any knowledge about the signs and characteristics of different groups of patients.

In this study, the aim is to address the problem of previous studies by enhancing the classification performance metrics and identifying important rules (participation of variables) that determine the outcome of patients infected with COVID-19. So, the main purpose of this study is to reduce COVID-19 mortality by improving medical decision-making. For this purpose, supervised machine learning methods are implemented to predict the *outcome of COVID-19 cases* (*recovery or death*) and identify the factors that have the greatest impact on their outcome. We trained decision tree, random forest, support vector machine (SVM), logistic regression, and k nearest neighbor (KNN) classification algorithms to diagnose death or recovery data classes. Our random forest model presented superior performance in predicting the outcome of patients infected with novel coronavirus than previous approaches. In the second part of the research, we identified the most important rules governing the set of COVID-19 symptoms through a rule-based technique. The discovered rules define various combinations of several important characteristics with a range of their values, which together determine the outcome of patients with more than 90% confidence. It helps physicians provide appropriate medical care for high-risk groups of patients.

In the next sections of the research, we first report the study area and collected data. After that, we describe the data preprocessing and developed models in the processing subsections. We explain our findings in the results section. Then we discuss different techniques, their results, and research applications. Finally, in the conclusion section, we state the goal of the research, our findings, limitations, and several suggestions for future works.

1.1. Related works

With the advent of the COVID-19 pandemic, artificial intelligence-based technologies helped humans in various areas of personal and public health. Utilizing these technologies has helped physicians and nurses improve patient services and facilitated efforts to find effective ways to combat the COVID-19 infection [25]. Most researches in this area focus on forecasting the number of infected cases, recoveries, or deaths [26–35].

Davoudi et al. [36] investigated the effect of three different types of statins on the severity of COVID-19 infection and then diagnosed the disease severity using machine learning methods. The results of research in the first step showed that the severity of the disease is lower for patients who took Simvastatin before infection. In addition, the researchers claimed in the second step that the decision tree classifier with 87.9% accuracy outperformed other classification models in diagnosing disease severity.

Sharifi et al. [37] studied the effects of the COVID-19 pandemic on several areas of society. The researchers used digital and artificial intelligence to measure the effects of the novel coronavirus on energy, industry, and medicine areas. The results showed the effect of the new coronavirus pandemic on energy-related industries and confirmed that renewable energies are significantly effective in reducing the destructive consequences of SARS-COV-2.

In [38], researchers presented a new version of the hybrid salp swarm (HSS) and genetic algorithms to improve nursing services during the COVID-19 pandemic. The authors claimed that the proposed algorithms are well able to address the nurses' scheduling and designation problems and perform better than the proposed state-of-the-art methods.

Researchers in Ref. [39] first developed the generalized logistic growth model (GML) to simulate the novel coronavirus sub-pandemic waves in Iran and then estimated the risk of inter-provincial travel in the prevalence of the COVID-19 disease. The results revealed that the GML model can well simulate the trend of two, three, and four waves. In addition, the results indicated that traveling between small cities and the capital of Iran significantly increases the risk of disease infection.

In another study, researchers examined the relationship between the sunspots number and the emergence of new viruses in the world. The researchers smoothed the sunspot cycle using a wavelet analysis method to examine the history of previous pandemics, predict the time of emergence of the new unknown virus, and identify high-risk geographical points for virus outbreak. The results confirmed that there is a significant relationship between the occurrence of the previous pandemics and the extrema of the sunspot number. In addition, the multi-step autoregression (MSAR) model estimated about 110 years for the emergence of a new pandemic [40].

The supervised machine learning applications are extensive in various fields of medicine. The researchers in Ref. [41] used decision tree, logistic regression, and support vector machine classifiers to diagnose pneumonia in limited resource and community settings. First, the researchers identified six essential features using a feature selection technique and then trained the learning models using 4500 cases. The test results showed that the decision tree model with an AUC of 93% recorded better performance than other models.

In [42], the researchers used a combined method involving supervised machine learning to identify an effective drug against coronavirus infection. Researchers used FDA-approved drugs, natural datasets in the literature, and zinc database to develop chemical libraries. These chemical libraries were used to select compounds interacting with target proteins of the COVID-19 infection such as spike and nucleocapsid proteins. The results suggest some approved drugs against hepatitis C virus, cancer, and ssRNA virus as candidate compounds to fight the novel coronavirus.

Lybarger et al. [43] developed an automated span-based framework for extracting events from the COVID-19 clinical text notes. Text data includes 1472 notes of symptoms, tests, and diagnoses provided by various patients. This research consists of two stages: 1. Construct and train the event extraction model, and 2. Predict the COVID-19 test results. The proposed framework predicted the symptom and assertion with high performance and outperformed the similar extractor developed on MetaMapLite for event extraction. In addition, the results showed that the extracted symptom annotations significantly improved the prediction performance on the structured data.

Today, a large amount of unstructured, structured, and semi-structured data is generated from various sources. Collecting, integrating, storing, and processing this amount of data with traditional data processing technologies is very time-consuming and expensive. So, it has led to the development of a new generation of big data processing technologies [12]. On the other hand, researchers have developed various feature extraction algorithms to reduce the dimension of big data, which significantly reduces computational and spatial costs [44]. Microarray data is a type of high-dimensional data generated by microarray technology to evaluate the manifestation level of genes [45]. Reducing the dimension of microarray data is critical for gene expression because it is hard to discover knowledge and identify hidden patterns from a large number of extracted genes [45].

In the new approaches proposed to reduce the dimension of microarray data, researchers developed a solution for selecting the optimal features of the microarray data. The proposed approach consists of two steps: 1. features extraction based on independent component analysis (ICA) method, and 2. selecting the optimal set of genes based on artificial bee colony (ABC) theory. The authors claimed that the proposed method reveals superior performance than the state-of-the-art approaches developed for selecting optimal genes from microarray data for the Naïve Bayes [46] and artificial neural network [47] classification methods.

2. Method and materials

2.1. Experiment data

We gathered data through interviews, questionnaires, and medical records of patients with COVID-19 admitted to Saveh Medical Center (SMC) in Iran during the year 2020. The research data contained 1,142 samples with 39 features per case. Some important features include age, sex, hospital unit, breathing conditions, fever, cough, different underlying diseases, the length of hospitalization, blood rate oxygen, intubation, and death or recovery class labels. In Table 1, we provided a detailed description and possible values of some important variables in this study. We picked out the most important attributes using a filter-based feature selection (extra-tree) method in the preprocessing stage. Besides, our data included 1131, and 111 recovered and died cases, respectively. Accordingly, our data suffered from a class imbalanced problem, which can affect learning models.

2.2. Data visualization

A useful tool for revealing hidden statistical properties in a dataset is data visualization. We visualize the data with different categorizes of visualization plots. Fig. 1 shows the age distribution of patients relative to hospital duration and the sex of patients based on their outcome (death or recovery), respectively. As can be seen, most of the patients were hospitalized for 1–10 days, and the number of patients who died (blue spots) was more than 60 years old (Fig. 1a). On the other hand, according to Fig. 1 b, the number of male patients (value 2) in this study is more than female patients (value 1). Furthermore, the mortality rate is almost the same among both sexes.

A useful way to explore the different characteristics of patients relative to each other is to present data with a bar plot. The bar diagram indicates the values assigned to the variables of a particular case. It also makes it possible to visually compare the variables of two or more samples based on the same variable. Fig. 2 represents the intubation, hospital section, blood oxygen level, and class label variables of some patients toward their age.

Visualization also provides a way to show the rate of change in the values of different features relative to each other. The line chart is the tool used for this purpose. The line chart in Fig. 3 shows the rate of change in intubation, hospital ward, blood oxygen level (ratePo2), and class label (death) characteristics for 50 COVID-19 cases.

Table 1
Definition and possible values of significant characteristics in this study.

Variable	Description
Age	Patient's age
Intubation	1; patient has undergone intubation, 2; patient has not undergone intubation.
Fever, cough, headache, chest pain	0 stands for absence of symptom, and 1 stands for presence of the symptom.
Contact coronavirus	0; no history of contact with COVID-19 cases, 1; history of contact with COVID-19 cases.
Section of hospital	the ward where the patient has been hospitalized. 1; regular ward, 2; intensive care unit, 3; no hospitalization.
Presence of underlying diseases	0 stands for absence of underlying diseases, and 1 stands for presence of the underlying diseases.
Rate of partial pressure of oxygen, Po2	0; PO2 levels are greater than 93, 2; PO2 levels are less than 93.
Shortness of breath	0 stands for absence of symptom, and 1 stands for presence of the symptom.
Hospital duration	number of hospitalization days.
Result PCR	0; negative for COVID-19, 1; positive for COVID-19, 3; test result is pending.
Condition entering the hospital	0; sever, 1; mild
CT scan manifestation	1; CT scan results for COVID-19 are negative, 2; CT scan results for COVID-19 are positive.
Death	no; patient has recovered, yes; patient has died.

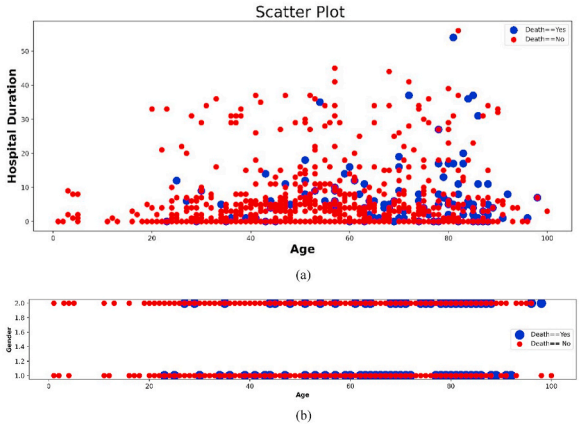


Fig. 1. Scatter plot of the age of patients relative to their (a) sex and (b) hospital duration based on death or recovery class labels.

Finally, we divide the data into separate sections using a facet chart and display it as a single plot. The facet chart in this study (Fig. 4) divides the data set into two subsets based on the class labels and shows the age distribution of patients in each subset. Accordingly, the highest age recurrence is 40 years in the recovered patients set and 80 years in the set of cases who died of infection.

2.3. Data pre-processing

In data analysis applications, data should be preprocessed before training the models. In this research, the preprocessing stage includes removing outliers, replacing null values, solving the class imbalance problem, feature selection, and random shuffle sampling. The output of the preprocessing step is a set of data that is acceptably empty of redundancy. It has a significant impact on improving the performance of learning models [24]. Additionally, preprocessing steps may require data sampling and normalization based on the learning method. The pre-processed data then enters the processing stage to train machine learning models. Fig. 5 shows an overview of the methodology used in this research.

In this study, null values in all variables are replaced with the mean value of that variable. It assures that the learning model does not tend to a specific value in a variable. Moreover, the K-means clustering method was applied to discover outliers in the data. The K-means clustering suggests data points as outliers that are further away from cluster centers than other points.

2.3.1. Feature selection

The purpose of feature selection is to find an optimal subset of characteristics by eliminating unrelated variables in the research data. So, it leads to advancing the model performance results and reduces computational complexity [48]. The most well-known methods for picking out features are Filter, Wrappers, and Embedded procedures [48]. The filter method does attribute ranking independent of the learning algorithm. Feature ranking describes the degree of influence of a feature in separating data classes [48].

In this study, we used filter-based extra tree classifier to select the ten most important variables in predicting the death or recovery classes of COVID-19 patients (Fig. 6). The extra tree is an ensemble and majority-vote-based classifier that employs a set of decision trees to distinguish class labels [49]. Fig. 6 presents the most relevant features to train the learning models. According to this figure, intubation, age, and the number of hospitalization days are the most effective properties to determine class labels, respectively.

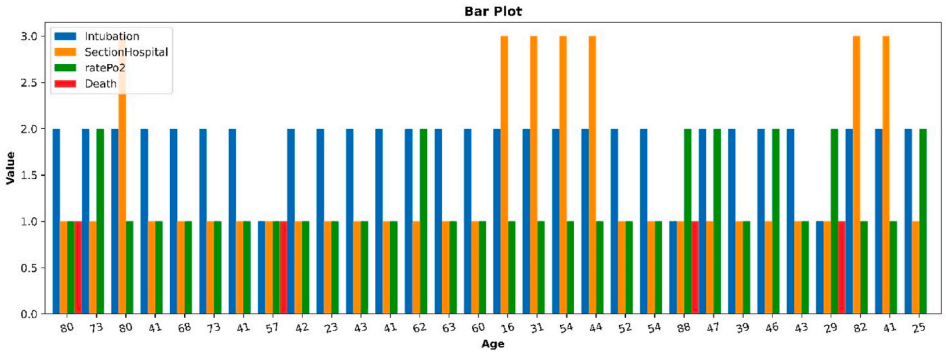


Fig. 2. Bar chart of Intubation, hospital unit, rate Po2, and the class features of the patients based on their age.

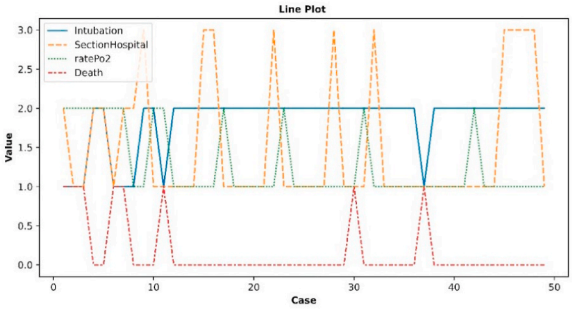


Fig. 3. Line chart of intubation, hospital ward, blood oxygen level, and class label variables for 50 COVID-19 cases.

2.3.2. *Balancing class labels*

The class imbalance problem is a common problem in supervised feature learning. Class imbalance indicates that the number of data samples with different labels is not balanced. It can greatly affect the prediction results. If the number of a particular label in a data set is very different from other labels, the data suffer from the class imbalance problem. Therefore, it is necessary to solve this problem, otherwise, the classification results will not be sufficiently reliable.

Conventional methods for solving the class imbalance problem are based on Over-sampling or Under-sampling approaches. The over-sampling methods balance class labels by replicating minority class samples. The under-sampling methods, on the other hand, balance class labels by eliminating some majority class samples.

In this study, the *cases* who *died* of *COVID-19 or recovered from the disease* constitute 9.8% and 90.2% of the data samples, respectively. Therefore, data classes are imbalanced in our research. So, we used the SMOTE [50] method to solve the class imbalance problem. The SMOTE algorithm is a method based on a data Over-sampling approach that incorporates the minority class samples to create new cases instead of duplicating them. This technique first randomly selects a sample from the minority class set and finds its best K nearest neighbor by the Euclidean distance measure. Then randomly selects some cases from its neighbors and generates a new synthetic sample by the selected samples [50]. This process is repeated for all the minority class samples until the number of the minority class cases equals the majority class cases. Equation (1) shows a way of combining samples in the SMOTE algorithm [50]. While X_n is the new sample, X is the sample selected from the minority class sample, and X_k is the sample selected from the neighbors set. Plus, the rand function generates a random number between zero and one.

$$X_n = X + \text{rand}(0, 1) * |X - X_k| \quad (1)$$

By solving the class imbalance problem in this study, the number of cases who died of new coronavirus (minority class) increased from 111 to 1131 cases, equal to recovered cases (majority class). Table 2 shows the number of patients belonging to each class label in this research after applying the SMOTE algorithm and dividing the data into a train and test set. We used the random shuffle strategy to randomize the samples, then applied the non-probable top-down sampling method to break down the dataset into train and test sets. We repeated this process several times and evaluated the classifier performance measures using the different numbers of the test samples. Accordingly, the models showed their best performance using 30% of test cases and 70% of train samples.

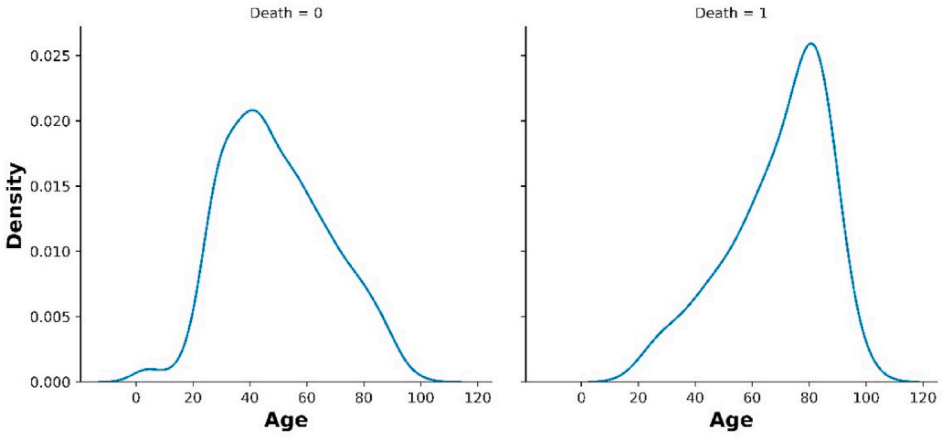


Fig. 4. Facet chart of the age distribution of patients based on different class labels.

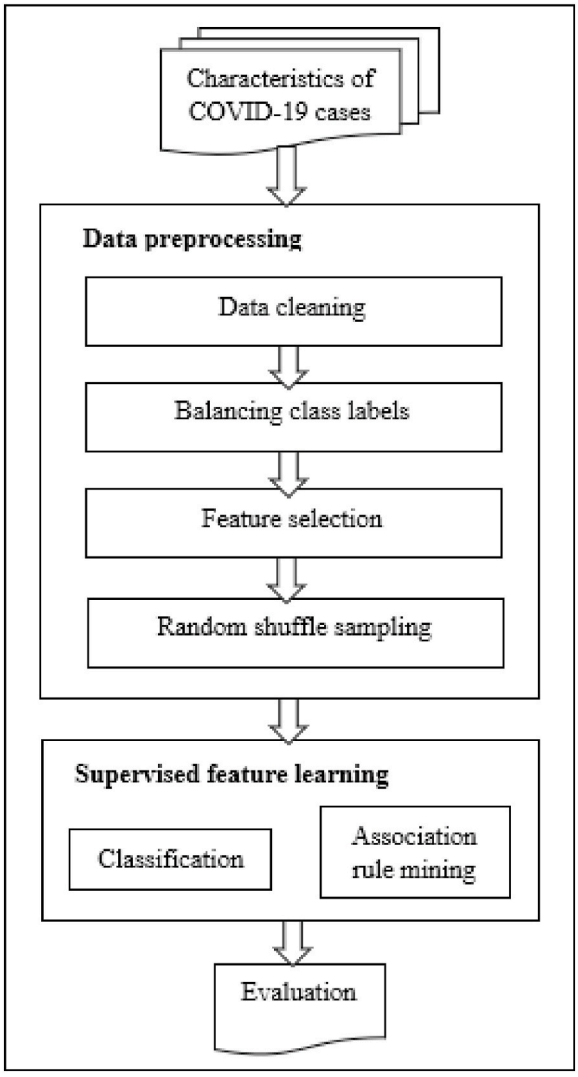


Fig. 5. The block diagram of the research methodology.

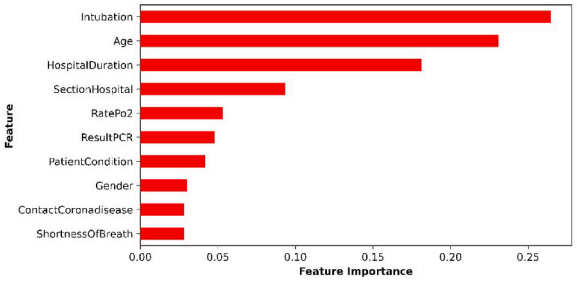


Fig. 6. Top 10 influence attributes selected using the filter-based technique.

Table 2
Number of samples in different classes in the train and test data set.

Data set	Recovery class samples	Death class samples	Total sample
Train (70%)	729	714	1443
Test (30%)	302	317	619
All	1031	1031	2062

2.4. Processing

In this stage, we train and test classification models using pre-processed data. In the literature, researchers have proposed many machine learning approaches for different artificial intelligence applications. Supervised, unsupervised, and semi-supervised feature learning approaches are the well-known machine learning categories [24].

The supervised machine learning methods can utilize what they have seen in the past to foretell the future. In these methods, the data set is divided into two train and test data sets. In the training phase, the model learns how to separate class labels (supervised learning), and in the testing phase, we evaluate how well the model can recognize the class of the samples [24].

2.4.1. Classification

Several classification models were developed for supervised feature learning to diagnose the outcome of patients infected with the coronavirus in this study. These learning models include decision tree, SVM, KNN, logistic regression, and random forest due to their low time complexity and high efficiency in classifying relational data. In the results section, the performance metrics of the models are detailed.

2.4.2. Association rule mining

Association rules extraction is a data mining operation that finds connections between the features of a data set [24]. In other words, association analysis is the study of features or characteristics that are related to each other and tries to extract rules from these features. This method seeks to discover rules to quantify the relationship between two or more properties [24]. Association rules are defined as if and then with two indices of Support and Confidence [24].

In this research, the aim is to determine rules for predicting the novel coronavirus behavior in different patients. In the next section, we describe the concepts needed to derive the rules that may not be familiar to the readers.

Important definition: Suppose that $I = [I_1, I_2, \ldots, I_m]$ is the set of total features available in the data set. Each subset of I is called a transaction, denoted by T, and D is the set of transactions in T. Then an association rule is shown as follows:

$X \rightarrow Y$ {Support, Confidence} So that, $X \subset I$, $Y \subset I$, $X \cap Y = \varnothing$.

Support: Indicates the percentage or number of D transactions that include both X and Y.

Confidence: Expresses the dependence of a particular feature on another feature and is calculated as Equation (2).

Strong rules: Strong rules are rules that have greater support and confidence than the determined threshold. In association rules analysis, the goal is to find and extract these strong rules.

$$\text{Confidence}(X \ , \ Y) = \frac{\text{support}(X \cup Y)}{\text{support}(X)} \tag{2}$$

Lift index: Lift index is a measure for evaluating association rules and shows the attractiveness of a rule [34], which is calculated as Equation (3).

$$\text{Lift}(X \rightarrow Y) = \frac{\text{support}(X \cup Y)}{\text{support}(X).\text{support}(Y)} \tag{3}$$

According to Equation (3), if the lift index for a rule is less (or greater) than one, then there is a negative (or positive) dependency between X and Y in the $X \rightarrow Y$ condition.

Many algorithms have been proposed to extract strong rules in the literature. One of the most widely used methods in this field is the Apriori algorithm, which has lower computational complexity than other methods [24]. We have used the Apriori algorithm with 30% and 90% support and confidence thresholds, respectively, to extract strong rules from the symptoms of COVID-19 disease in determining the

outcome of infected patients. The extracted rules that meet all the defined conditions include 269 rules. Figs. 7 and 8 show the scatter diagram and the grouped matrix of the extracted rules, respectively. The higher color intensity in Fig. 7 indicates a higher degree of confidence (close to one). But, the higher color intensity in Fig. 8 represents the lift value, and the node size describes the support value. The matrix elements in Fig. 8 show the symptoms of COVID-19 cases with specific intervals that lead to the patient's recovery or death. Although all 269 rules are important, the rules filtered in Fig. 7 (top left corner) are more important than the rest of the rules because they have higher lift and confidence indices. In the next section, we will examine the ten most important of these rules in more detail and state what important factors together determine the outcome of infected patients with high confidence.

2.5. Analysis environment

Our system is supported by CPU 2.3Ghz (five-core), 6 gigabytes of RAM, and one terabyte of disk space to implement algorithms in this study. We implemented visualization, preprocessing, and classification steps with the Python programming language version 3.5. Also, we use R language version 3.5.1 to discover association rules due to its capability to display results. Finally, We used dplyr (version 1.0.7), ggplot2 (version 3.3.5), plyr (version 1.8.6), kohonen (version 3.0.10), arules (version 1.7), and arulesVIZ (version 1.5-1) packages in the R statistical language, and pandas (version 1.3.5), matplotlib (version 3.5.1), seaborn (version 0.11.2), numpy (version 1.0.7), Scikit-learn (version 0.23.2), imblearn (version 0.8.1), collections (version 3.7.12), and yellowbrick (version 1.3) packages in the Python language for statistical analysis, visualization, implementation of models, and validation of results.

3. Result

In this section, we present the research results for different methods separately. First, the classification performance metrics are computed in detail for each classification algorithm and compared with the results of previous studies. Second, different sets of symptoms and the range of their values that play a significant role in determining the outcome of COVID-19 patients are identified using the association rules mining technique.

3.1. Classification performance

We adjusted the classification models with different parameters, trained with 70% of the data samples, and tested with the remaining 30%. The KNN model with parameter k = 3 and random forest model with 150 predictors recorded their best performance. To evaluate the classification performance of the models, we applied the model evaluation metrics presented in Equations (4)–(8). In these equations, true positive (TP) represents positive samples that the model accurately predicts as positive. False positive (FP) shows negative samples that the model mistakenly classifies as positive. Also, true negative (TN) is the number of negative samples that the model correctly predicts as negative. Ultimately, false negative (FN) indicates cases that the model should predict positive but wrongly considers negative [24].

Confusion Matrix is a useful tool for examining the performance of classification models in recognizing data class labels. If the data samples are grouped into M classes (where $M \geq 2$), the confusion matrix C will have at least M rows (actual value) and M columns (predicted value) for different class labels. In other words, a confusion matrix represents TP, FN, FP, and TN indexes for a classifier based on its observations. There may be additional rows or columns in the matrix to represent the sum of the samples or the percentage of recognition. Accordingly, Fig. 9 illustrates the confusion matrices of the classification models developed in this research. Besides, Table 3 provides the classification performance metrics derived from the confusion matrices in Fig. 9 for the different learning models.

$$\text{Accuracy}: \frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

$$\text{Precision}: \frac{TP}{TP+FP} \tag{5}$$

$$\text{Sensitivity}: \frac{TP}{TP+FN} \tag{6}$$

$$\text{Specificity}: \frac{TN}{TN+FP} \tag{7}$$

$$\text{F1}-\text{score}: \frac{2(Precision * Sensitivity)}{(Precision + Sensitivity)} \tag{8}$$

Fig. 10 displays the receiver operating characteristic (ROC) curve to compare the classification models. ROC curve is used to assess the performance of two or more classification models in a two-class problem [24]. This curve compares the performance of the classifier based on the rate of changes between the proportion of positive samples that the model correctly detects positive (TPR) and the ratio of negative cases that the model mistakenly labeled as positive (FPR) in different parts of the test set [24]. Meanwhile, the area under the ROC curve (AUC) for each model demonstrates a degree of the accuracy of the model. The closer the AUC score is to one, the better the performance of the model in distinguishing between positive and negative cases. Therefore, a model is selected as the final model whose AUC score is close to one and higher than the AUC score of the other models [24]. According to Table 3, the highest AUC score belongs to the random forest model that confirms its

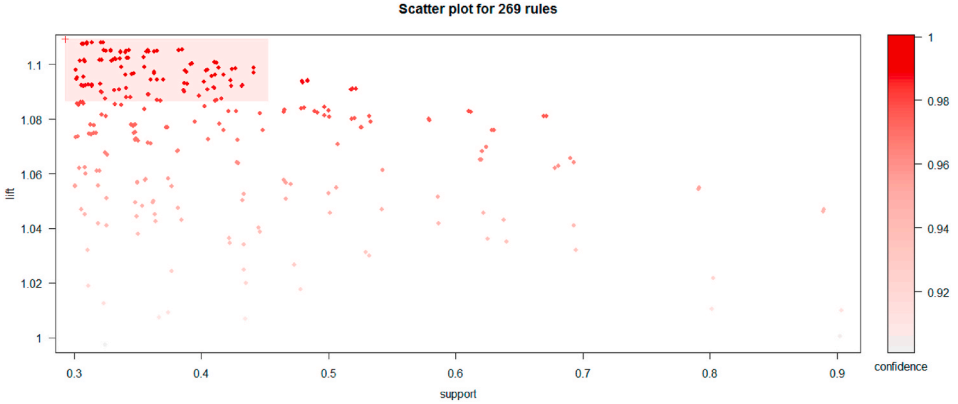


Fig. 7. The association rules extracted with the expected conditions (support = 0.3, confidence = 0.9, lift> 1).

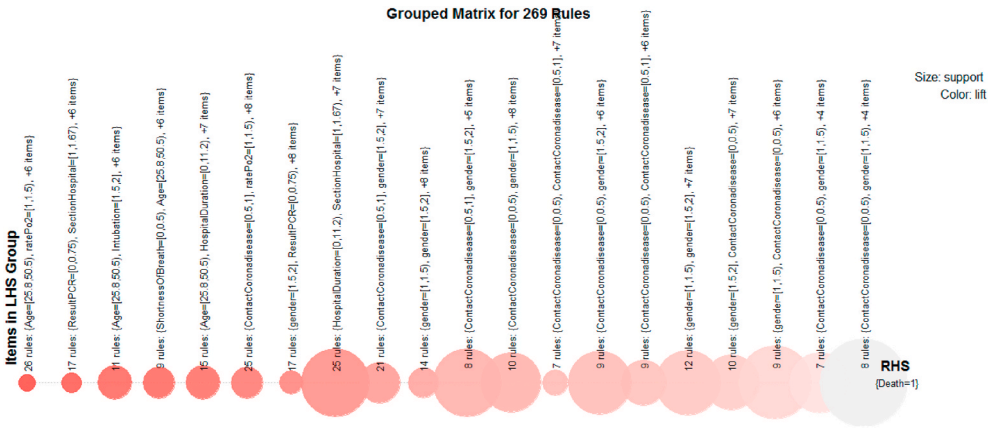


Fig. 8. Grouped matrix diagram of extracted association rules with expected conditions (support 0.3, confidence 0.9, lift> 1).

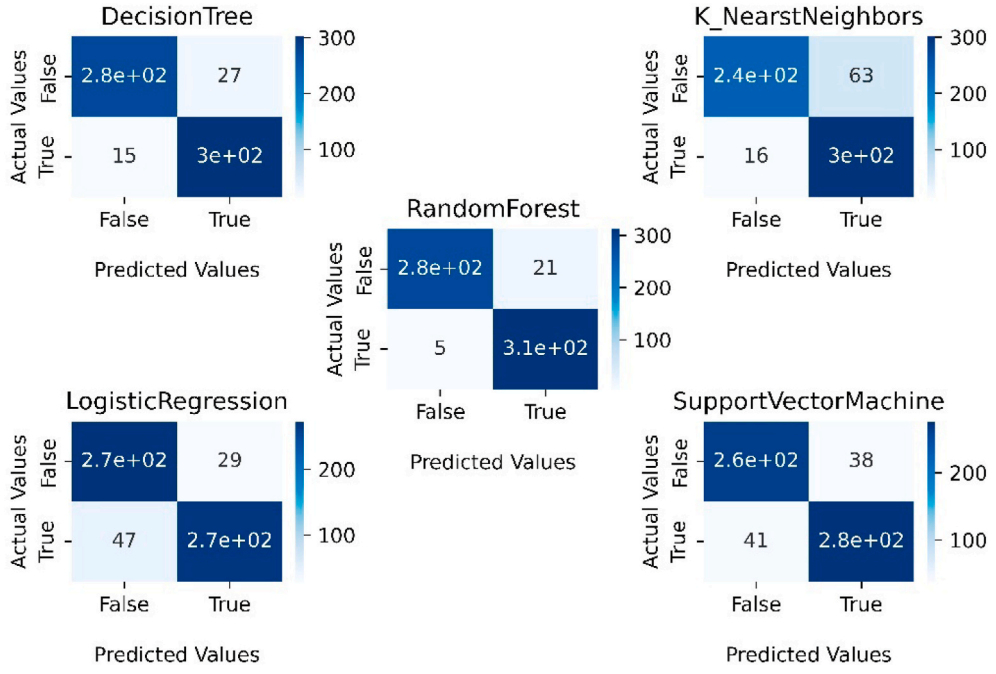


Fig. 9. Confusion matrices of the models developed in this research.

Table 3
Classification performance results to diagnosis the outcome of COVID-19 cases.

Model	Acc (%)	Pr (%)	Se (%)	Sp (%)	F1_score (%)	AUC_score (%)
Decision tree	93.21	91.29	95.89	90.39	93.53	93.14
SVM	87.07	86.23	88.95	85.09	87.57	95.66
Knn	86.91	82.24	94.95	78.47	88.14	93.02
Logistic Regression	86.59	87.74	85.80	87.41	86.76	95.48
Random forest	95.63	94.47	97.16	94.03	95.80	99.38

better performance than other models. So, the random forest classifier with 150 predictors is selected as the final model for outcome prediction of the COVID-19 cases.

Finally, Table 4 compares the classification performance results of the random forest model in this research with the results of previous state-of-the-art studies. Based on data provided in Table 4, it is obvious that our model performs better in determining the outcome of coronavirus patients than the previous researches.

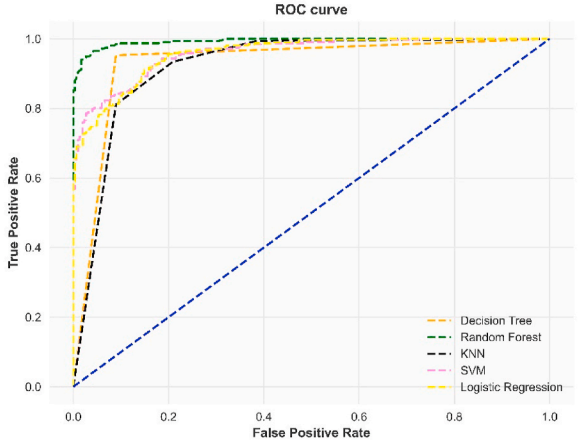


Fig. 10. ROC curve to compare the performance of the models implemented in this research.

Table 4
Classification performance of the proposed method and comparison with state-of-the-art methods.

Research	Method	Acc (%)	Pr (%)	Se (%)	Sp (%)	F1_score (%)
An, Chansik, et al. [51]	Linear SVM	91.9	25.6	92.0	91.8	40.0
Chen et al. [52]	Logistic Regression	–	–	91.4	76.0	–
Chowdhury et al. [53]	Nomogram	–	–	92.0	92.0	–
Iwendi et al. [54]	Random Forest	94.0	**100**	75.0	–	86.0
Sumayh S et al. [55]	Random Forest	95.2	95.0	94.9	93.6	95.5
Mohammad and Mahdi [56]	Neural Network	89.9	93.6	87.7	93.2	90.5
Rahila et al. [57]	Bayes Net	89.0	–	92.6	86.0	–
Proposed method	Random Forest	**95.6**	94.4	**97.1**	**94.0**	**95.8**

3.2. Association rules results

In this study, we used association rules mining by the Apriori algorithm to identify a set of symptoms that determine the outcome of COVID-19 patients with high confidence (0.9) and support (0.3). In each specific application, the rules generated are usually high. Therefore, we need to extract only the most essential rules that satisfy the value of the support and confidence thresholds, and the correlation between their items and their results is positive. We used the lift index to extract the ten most momentous rules in the data set by evaluating the relationship between the item(s) and the label of the rules. We visualized these rules with a parallel coordinates plot in Fig. 11. The vertical axis of the diagram in Fig. 11 shows the set of items, and the horizontal axis represents the position of the items in different rules. According to the result, all the rules lead to the patient's recovery (death = 1). In our data set, the class label (death) column with numbers 1 and 2 describes the *cases* who *died* of *COVID-19 or recovered from it*, respectively. Other items include hospital section, hospital duration, patient's respiratory condition, the patient's condition when visiting the hospital, blood oxygen level, need for a ventilator, and patient's age, respectively.

Before extracting the rules, the values of each attribute are discretized into different intervals. Each rule specifies the extent to which features can together determine the outcome of COVID-19 patients. In addition, it recognizes variables in which range of their values have the most occurring in the data set. The range [1,1.67) for the hospital section indicates the high range of its values and, as mentioned previously, refers to the hospital wards where patients with usual symptoms are admitted. Also, the rules record that the number of hospitalization days with a range between 0 and 11 days has an effective role in determining the patient's recovery. In this study, the set of values for lack of the patient's shortness of breath and the patient's shortness of breath is 0 and 1, respectively. Shortness of breath (interval [0,0.5)) indicates lack of shortness of breath or mild shortness of breath in patients. The

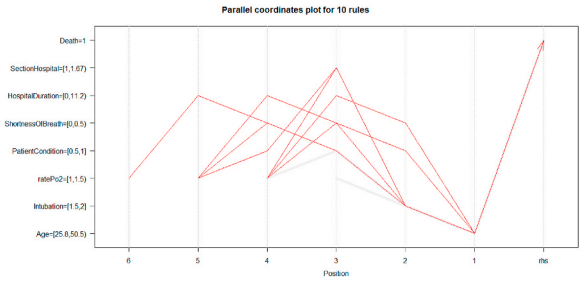


Fig. 11. The ten most essential association rules were discovered in this research.

next feature is the patient's condition when visiting the hospital. This feature with a value of zero describes the patients with normal conditions, and a value of one expresses the patients with severe conditions. So, the rules indicate that patients recover in a critical condition (interval [0.5,1]). This unusual situation in the rules occurs because almost all patients in our data set are hospitalized in unfavorable conditions. Blood oxygen level in patients is another feature that plays a significant role in their recovery. The higher the rate of pulmonary infection caused by the coronavirus, the lower the blood oxygen level in patients. In coronavirus, the values of 1 and 2 determine the oxygen level above 93% and less than 93%, respectively. As Fig. 11 shows, the oxygen level with intervals [1, 1.5) plays a substantial role in various rules. Moreover, the items that have been reviewed so far, the intubation feature is another essential feature that occupies a place in the set of rules. Intubation indicates whether a patient has needed a ventilator device or not. Patients who did not require intubation are marked with the number 2, otherwise, the value of the intubation variable is 1. Naturally, the rules indicate that intervals [1.5,2] are necessary for patients to recover because they do not need artificial respiration. Finally, the age property of patients as the last item is essential in determining the patient's recovery. Patients between the ages of 25 and 50 are more likely to recover.

Finally, it is principal to note that the extracted rules are one-sided. It means that although a combination of properties can result in the patient's recovery with high confidence, a patient may have recovered but not meet any of the conditions set out in Fig. 11. In other words, the rules do not guarantee that a patient who is not covered by any of the rules will die.

4. Discussion

In this study, we employed machine learning techniques to evaluate the clinical symptoms of patients who died of COVID-19 or recovered from it. Machine learning techniques in this research mainly included supervised methods. Classifying patients based on two-class labels (recovery and death) and discovering association rule mining were two supervised methods used.

One problem with data classification is the class imbalance problem. In this study, the initial number of patients who recovered from the infection or died of it was 1031 and 111 cases, respectively. So, our data suffered from the class imbalance problem. We used the SMOTE algorithm to solve the problem by increasing the cases who died of infection. We then developed the decision tree, KNN, SVM, logistic regression, and random forest models with 70% of the train data and tested them with the remaining 30% samples. Our results showed that the random forest classifier with 95.63% accuracy, 94.47% precision, 97.16% sensitivity, 94.03% specification, and 95.8% F1-score has better performance in diagnosing the outcome of COVID-19 patients than methods proposed in the recent studies (Table 4).

Random forest is an ensemble classification method based on the majority vote and uses a combination of decision trees to classify data classes. Although it has a higher computational complexity, past researches show that this classifier performs more beneficial than other classification models due to its ensemble nature. It can justify the higher performance of this model in this research.

Also, we presented a rule-based approach to determine the sets of factors that together affect the outcome of patients. We set the support and confidence values to 30% and 90, respectively. The purpose of adopting a high threshold for support and confidence measures is to find the most valid rules. In the next step, we calculated the quality for all generated rules by the lift metric and represented the ten highest quality rules through a parallel coordination diagram (Fig. 11). All extracted rules were related to the recovered patients and did not provide knowledge about patients who died of infection. However, the high number of recovered cases confirms the produced rules because they follow more patterns in the data set. Accordingly, the rules tell us what

symptoms and in what range of their values with high confidence will result in the patient's recovery. The items that made up the most important rules were age, intubation, hospital section, breathing condition, the patient condition when visiting the hospital, hospital duration, and blood oxygen rate. Different combinations of these attributes led to the production of various rules (Fig. 11).

Finally, our results confirm the research presented in the past. Most studies have identified patients' age as an essential factor in the outcome of coronavirus cases. According to the latest research, more than 70% of deaths in Iran are among people over 60 years old. Besides, our results present new information to physicians. For example, specialists should consider the value of clinical variables such as hospital duration, patient's respiratory condition, hospital ward, and patient condition at the time of hospitalization as other important factors to improve treatment services for high-risk patients at different stages of patients' treatment, in addition to the features identified in the past, such as age and underlying diseases.

4.1. Limitation

This research faced some constraints. Our data lacked some of the pathological and neurological features of COVID-19 cases. A further variety of features may enhance the classification performance metrics. Besides, we considered ten features as the best attributes to improve reliability, so more variables will be investigated by raising this threshold. Besides, the deep learning-based methods could not be implemented on our data because they require a large amount of data for high-level performance and our research data did not provide this number of required samples. We can provide a Big Data framework by integrating a variety (Big Data feature) of data types such as CT-scan images and electrocardiogram (ECG) signals of COVID-19 patients to utilize deep neural networks. On the other hand, it significantly raises the computational costs of the operations.

5. Conclusion

We set a target to investigate different aspects of the new coronavirus using two supervised machine learning methods. We removed redundancies from the data and solved the class imbalance problem, then developed different classification models to diagnose the outcome of COVID-19 cases. Our random forest model outperformed the previous state-of-the-art models in this area of study. Next, we developed a rule-based method to extract the essential association rules governing the COVID-19 cases. The rules identify various combinations of 6 features and the range of their values that specify the patient's recovery with a confidence value of 90%. Accordingly, this study improved the classification performance metrics in detecting the outcome of COVID-19 cases (recovery or death), identified the most effective sets of characteristics and range of their values that together determine the outcome of COVID-19 patients, distinguished high-risk groups of patients in the early stages of the disease, and improved medical decision-making and health services by separating different groups of disease cases.

Our results help health specialists consider other factors to enhance healthcare services for the COVID-19 patients. Specialists can handle different groups of patients by measuring the characteristics that have been identified as efficient in this research. It conclusively can help to decrease COVID-19 fatality.

Data availability statement

The data utilized for finding the outcomes of this research have been taken through questionnaires and patients' medical records in the SMC, Iran. Research data was approved by the SMC in Iran and was provided by figshare repository with unique identifier "http://doi.org/10.6084/m9.fihshare.12446120.v1" and under "Attribution 4.0 (CC BY 4.0)" license.

Funding statement

The authors received no financial support for the research and/or authorship of this article.

Institutional Review Board statement

Our research was confirmed by the Institutional Review Board of Department of Information Technology Engineering, Industrial and System Engineering Faculty, Tarbiat Modares University. Ethical review and approval were waived for this study due to the data samples lacked the participants' personal information, and our study did not violate participants' privacy.

Informed consent statement

Patient consent was waived due to the data samples lacked the participants' personal information, and our study did not violate participants' privacy.

Author contributions

Conceptualization, S.I. and A.A.; Methodology, S.I.; Software, S.I; Validation, A.A and S.I; Formal Analysis, S.I; Investigation, A.A, S.I; Data Curation, S.I; Writing – Original Draft Preparation, S.I; Writing - Review & Editing, S.I; Visualization, S.I; Supervision, A.A; Project Administration, A.A.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to express our full thanks to Saveh Medical Center for providing medical data.

References

[1] Chen N, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet 2020;395 (10223):507–13.
[2] Yoosefi Lebni J, et al. How the COVID-19 pandemic effected economic, social, political, and cultural factors: a lesson from Iran. Int J Soc Psychiatr 2020;67(3): 298–300. 0020764020939984.
[3] "World Health Organization. Coronavirus disease (COVID-19) pandemic. Geneva: World Health Organization; 2020.
[4] Rothan HA, Byrareddy SN. The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak. J Autoimmun 2020;109:102433.
[5] Gupta R, Pandey G, Chaudhary P, Pal SK. Machine learning models for government to predict COVID-19 outbreak. Digital Government: Research and Practice 2020;1 (4):1–6.
[6] Zeraati Malihe, et al. Optimization and predictive modelling for the diameter of nylon-6, 6 nanofibers via electrospinning for coronavirus face masks. J Saudi Chem Soc 2021;25(11):101348.
[7] Ndaïrou F, Area I, Nieto JJ, Torres DF. Mathematical modeling of COVID-19 transmission dynamics with a case study of Wuhan. Chaos, Solit Fractals 2020;135: 109846.
[8] Simsek Murat, Kantarci Burak. Artificial intelligence-empowered mobilization of assessments in COVID-19-like pandemics: a case study for early flattening of the curve. Int J Environ Res Publ Health 2020;17(10):3437.
[9] Nabi KN. Forecasting COVID-19 pandemic: a data-driven analysis. Chaos, Solit Fractals 2020;139:110046.
[10] Chakraborty T, Ghosh I. Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: a data-driven analysis. Chaos, Solit Fractals 2020; 135:109850.
[11] Roosa K, et al. Real-time forecasts of the COVID-19 epidemic in China from February 5th to February 24th, 2020. Infect. Disease Model. 2020;5:256–63.
[12] Ilbeigipour Sadegh, Albadvi Amir, Elham Akhondzadeh Noughabi. Real-time heart arrhythmia detection using Apache spark structured streaming. J. Healthcare Eng. 2021;2021.

[13] Saraswathi S, Mukhopadhyay A, Shah H, Ranganath T. Social network analysis of COVID-19 transmission in Karnataka, India. Epidemiol Infect 2020;148.
[14] Pascual-Ferr P, Alperstein N, Barnett DJ. Social network analysis of COVID-19 public discourse on twitter: implications for risk communication. Disaster Med Public Health Prep 2020:1–9.
[15] Ahmadi Mohsen, et al. Investigation of effective climatology parameters on COVID-19 outbreak in Iran. Sci Total Environ 2020;729:138705 [].
[16] Saha Prottoy, Sadi, Sheikh Muhammad, Islam, Milon Md. EMCNet: automated COVID-19 diagnosis from X-ray images using convolutional neural network and ensemble of machine learning classifiers. Inform Med Unlocked 2021;22:100505.
[17] Islam, Md Zabirul, Islam, Milon Md, Asraf Amanullah. A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images. Inform Med Unlocked 2020;20:100412.
[18] Saiz FA, Barandiaran I. COVID-19 detection in chest X-ray images using a deep learning approach. Int. J. Interact. Multimedia Artif. Intell. InPress (InPress) 2020; 1.
[19] Silva Pedro, et al. COVID-19 detection in CT images with deep learning: a voting-based scheme and cross-datasets analysis. Inform Med Unlocked 2020;20:100427.
[20] Hassantabar Shayan, Ahmadi Mohsen, Sharifi Abbas. Diagnosis and detection of infected tissue of COVID-19 patients based on lung X-ray image using convolutional neural network approaches. Chaos, Solit Fractals 2020;140:110170.
[21] Bharati Subrato, Podder Prajoy, Mondal, Hossain M Rubaiyat. Hybrid deep learning for detecting lung diseases from X-ray images. Inform Med Unlocked 2020;20:100391.
[22] Hussain E, Hasan M, Rahman MA, Lee I, Tamanna T, Parvez MZ. CoroDet: a deep learning based classification for COVID-19 detection using chest X-ray images. Chaos, Solit Fractals 2021;142:110495.
[23] Panwar H, Gupta P, Siddiqui MK, Morales-Menendez R, Singh V. Application of deep learning for fast detection of COVID-19 in X-Rays using nCOVnet. Chaos, Solit Fractals 2020;138:109944.
[24] Han J, Kamber M, Pei J. Data mining concepts and techniques third edition. Morgan Kaufmann Series Data Manag. Syst. 2011;5(4):83–124.
[25] Sharma, Sandeep Kr, et al. Artificial intelligence-based systems for combating COVID-19. Applications of artificial intelligence in COVID-19. Springer, Singapore, 2021. 19-34.
[26] Agbehadji, Edem Israel, et al. Review of big data analytics, artificial intelligence and nature-inspired computing models towards accurate detection of COVID-19 pandemic cases and contact tracing. Int J Environ Res Publ Health 2020;17(15): 5330.
[27] Bragazzi, Nicola Luigi, et al. How big data and artificial intelligence can help better manage the COVID-19 pandemic. Int J Environ Res Publ Health 2020;17(9):3176.
[28] Yeşilkanat CM. Spatio-temporal estimation of the daily cases of COVID-19 in worldwide using random forest machine learning algorithm. Chaos, Solit Fractals 2020;140:110210.
[29] Wang P, Zheng X, Li J, Zhu B. Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics. Chaos, Solit Fractals 2020;139: 110058.
[30] Mollalo A, Rivera KM, Vahedi B. Artificial neural network modeling of novel coronavirus (COVID-19) incidence rates across the continental United States. Int J Environ Res Publ Health 2020;17(12):4204.
[31] Santosh K. AI-driven tools for coronavirus outbreak: need of active learning and cross-population train/test models on multitudinal/multimodal data. J Med Syst 2020;44(5):1–5.
[32] Yadav M, Perumal M, Srinivas M. Analysis on novel coronavirus (COVID-19) using machine learning methods. Chaos, Solit Fractals 2020;139:110050.
[33] Alimadadi A, Aryal S, Manandhar I, Munroe PB, Joe B, Cheng X. Artificial intelligence and machine learning to fight COVID-19. Bethesda, MD: American Physiological Society; 2020.
[34] Leeuwenberg AM, Schuit E. Prediction models for COVID-19 clinical decision making. Lancet Digit. Health 2020;2(10):e496–7.
[35] Santosh K. COVID-19 prediction models and unexploited data. J Med Syst 2020;44 (9):1–4.
[36] Davoudi Alireza, et al. Studying the effect of taking statins before infection in the severity reduction of COVID-19 with machine learning. BioMed Res Int 2021;2021.
[37] Sharifi Abbas, Ahmadi Mohsen, Ali Ala. The impact of artificial intelligence and digital style on industry and energy post-COVID-19 pandemic. Environ Sci Pollut Control Ser 2021;28(34):46964–84.
[38] Moein Qaisari Hasan Abadi, et al. HSSAGA: designation and scheduling of nurses for taking care of COVID-19 patients using novel method of Hybrid Salp Swarm Algorithm and Genetic Algorithm. Appl Soft Comput 2021;108:107449.
[39] Ahmadi Mohsen, Sharifi Abbas, Khalili Sarv. Presentation of a developed sub-epidemic model for estimation of the COVID-19 pandemic and assessment of travel-related risks in Iran. Environ Sci Pollut Control Ser 2021;28(12):14521–9.
[40] Nasirpour, Hossein Mohammad, et al. Revealing the relationship between solar activity and COVID-19 and forecasting of possible future viruses using multi-step autoregression (MSAR). Environ Sci Pollut Control Ser 2021:1–11.
[41] Stokes Katy, et al. A machine learning model for supporting symptom-based referral and diagnosis of bronchitis and pneumonia in limited resource settings. Biocybern Biomed Eng 2021;41(4):1288–302.
[42] Kadioglu Onat, et al. Identification of novel compounds against three targets of SARS CoV-2 coronavirus by combined virtual screening and supervised machine learning. Comput Biol Med 2021;133:104359.
[43] Lybarger Kevin, et al. Extracting COVID-19 diagnoses and symptoms from clinical text: a new annotated corpus and neural event extraction framework. J Biomed Inf 2021;117:103761.
[44] Huang Xuan, Wu Lei, Ye Yinsong. A review on dimensionality reduction techniques. Int J Pattern Recogn Artif Intell 2019;33(10):1950017.
[45] Aziz Rabia, Verma CK, Srivastava Namita. Dimension reduction methods for microarray data: a review. AIMS Bioeng. 2017;4(2):179–97.
[46] Musheer Rabia Aziz, Verma CK, Srivastava Namita. Novel machine learning approach for classification of high-dimensional microarray data. Soft Comput 2019;23(24):13409–21.
[47] Aziz Rabia, et al. Artificial neural network classification of microarray data using new hybrid gene selection method. Int J Data Min Bioinf 2017;17(1):42–65.
[48] Chandrashekar G, Sahin F. A survey on feature selection methods. Comput Electr Eng 2014;40(1):16–28.
[49] Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Mach Learn 2006;63 (1):3–42.
[50] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321–57.
[51] An Chansik, et al. Machine learning prediction for mortality of patients diagnosed with COVID-19: a nationwide Korean cohort study. Sci Rep 2020;10(1):1–11. .
[52] Hu Chuanyu, et al. Early prediction of mortality risk among patients with severe COVID-19, using machine learning. Int J Epidemiol 2020;49(6):1918–29 [].
[53] Chowdhury Muhammad EH, et al. An early warning tool for predicting mortality risk of COVID-19 patients using machine learning. Cognitive Computation; 2021. p. 1–16. .
[54] Iwendi Celestine, et al. COVID-19 patient health prediction using boosted random forest algorithm. Front Public Health 2020;8:357.
[55] Aljameel Sumayh S, et al. Machine learning-based model to predict the disease severity and outcome in COVID-19 patients. Sci Program 2021:2021.
[56] Pourhomayoun Mohammad, Shakibi Mahdi. Predicting mortality risk in patients with COVID- 19 using machine learning to help medical decision-making. Smart Health 2021;20:100178.
[57] Sardar Rahila, Sharma Arun, Gupta Dinesh. Machine learning assisted prediction of prognostic biomarkers associated with COVID-19, using clinical and proteomics data. Front Genet 2021;12:.