Original Article
http://mjiri.iums.ac.ir
Medical Journal of the Islamic Republic of Iran (MJIRI)

Med J Islam Repub Iran. 2022 (24 Sep);36.110. https://doi.org/10.47176/mjiri.36.110

Supervised Machine Learning Approach to COVID-19 Detection Based on Clinical Data

Azita Yazdani[1], Maryam Zahmatkeshan[2,3]*, Ramin Ravangard[4], Roxana Sharifian[5], Mohammad Shirdeli[6]

Received: 7 Jan 2021 Published: 24 Sep 2022

Abstract

Background: The new coronavirus has been spreading since the beginning of 2020, and many efforts have been made to develop vaccines to help patients recover. It is now clear that the world needs a rapid solution to curb the spread of COVID-19 worldwide with non-clinical approaches such as artificial intelligence techniques. These approaches can be effective in reducing the burden on the health care system to provide the best possible way to diagnose the COVID-19 epidemic. This study was conducted to use Machine Learning (ML) algorithms for the early detection of COVID-19 in patients.

Methods: This retrospective study used data from hospitals affiliated with Shiraz University of Medical Sciences in Iran. This dataset was collected in the period March to October 2020 andcontained 10055 cases with 63 features. We selected and compared six algorithms: C4.5, support vector machine (SVM), Naive Bayes, logistic Regression (LR), Random Forest, and K-Nearest Neighbor algorithm using Rapid Miner software. The performance of algorithms was measured using evaluation metrics, such as precision, recall, accuracy, and f-measure.

Results: The results of the study show that among the various used classification methods in the diagnosis of coronavirus, SVM (93.41% accuracy) and C4.5 (91.87% accuracy) achieved the highest performance. According to the C4.5 decision tree, "contact with a person who has COVID-19" was considered the most important diagnostic criterion based on the Gini index.

Conclusion: We found that ML approaches enable a reasonable level of accuracy in the diagnosis of COVID-19.

Keywords: COVID-19, Data mining, Machine Learning, Artificial Intelligence, Classification

Conflicts of Interest: None declared
Funding: This study was supported by Shiraz University of Medical Sciences.

Cite this article as: Yazdani A, Zahmatkeshan M, Ravangard R, Sharifian R, Shirdeli M. Supervised Machine Learning Approach to COVID-19 Detection Based on Clinical Data. *Med J Islam Repub Iran.* 2022 (24 Sep);36:110. https://doi.org/10.47176/mjiri.36.110

Introduction

One of the most serious global public health threats is emerging infectious diseases (1). The Coronavirus Disease 2019 (COVID-19), a public health emergency of international concern, is thought to have originated in Wuhan, China (2). COVID-19 as a human pathogen, is spreading around the world rapidly (3). The main symptoms of Coronavirus include fever, cough, and shortness of breath, which in many cases appear to be similar to the influenza virus (4). According to the Iranian Ministry of Health, the COVID-19 pandemic has affected 31 provinces in Iran and as of January 2021, it has infected 1,385,706 people and left

↑*What is "already known" in this topic:*
Effective screening of coronavirus enables fast and efficient diagnosis and can discount the burden on healthcare systems. Diagnosis models based on ML algorithms that combine several attributes to assess the risk of infection are useful to assist healthcare teams in triaging patients, especially in the context of limited healthcare resources.

→*What this article adds:*
We utilized the different supervised ML algorithms and compared their efficiency in diagnosing COVID-19. Our findings show that the SVM algorithm can be used as a potential diagnostic classifier for earlier detection of COVID-19.

Corresponding author: Dr Maryam Zahmatkeshan, m.zahmatkeshan@fums.ac.ir

[1] Department of Health Information Management, Clinical Education Research Center, Health Human Resources Research Center, School of Health Management and Information Sciences, Shiraz University of Medical Sciences, Shiraz, Iran
[2] Noncommunicable Diseases Research Center, Fasa University of Medical Sciences, Fasa, Iran
[3] School of Allied Medical Sciences, Fasa University of Medical Sciences, Fasa, Iran
[4] Department of Health Services Management, Health Human Resources Research Center, School of Health Management and Information Sciences, Shiraz University of Medical Sciences, Shiraz, Iran
[5] Department of Health Information Management, Health Human Resources Research Center, School of Health Management and Information Sciences, Shiraz University of Medical Sciences, Shiraz, Iran
[6] Department of Health Information Management, Student Research Committee, Health Human Resources Research Center, School of Health Management and Information Sciences, Shiraz University of Medical Sciences, Shiraz, Iran

57,560 dead (5). It is necessary to use non-clinical or non-medical treatment techniques to control and prevent the further spread of COVID-19 epidemic diseases. Over the past few decades, advances in modern technology are gradually changing medical practice (6, 7). Data science are diverse scopes that are actively used for COVID-19 detection, prognosis, prediction, and prevalence forecasting (8, 9). In recent years, we have seen progress in electronic data collection. Disease registration systems provide a useful tool for public health surveillance (10). An electronic record of patient health information, while providing a dynamic and flexible structure for reporting (11, 12), Allows the discovery of hidden knowledge from stored data. With the expansion of electronic data registration in health care, there is a large and complex amount of data that cannot be analyzed by traditional methods. Because of this, the need for data mining in health care is essential. Machine learning and data mining approaches have been widely used in health care, including predicting outcomes, evaluating treatment effectiveness, controlling infection, and diagnosing disease (13). Data mining is an advanced artificial intelligence(AI) technique that is used to detect hidden, useful, new, and valid patterns from datasets (14). In a short time after the outbreak of the new coronavirus, countries began to electronically record inpatient and outpatient data for COVID-19. With the availability of data sources from patients with COVID-19, it has been possible to analyze this data for knowledge discovery by data mining techniques. Data mining techniques have also been widely used in the prognosis and diagnosis of coronaviruses (CoV) diseases, including the Acute Respiratory Syndrome Coronavirus (SARS-CoV) and the Middle East Respiratory Syndrome Coronavirus (MERS-CoV) (15).

Research results show that to estimate and predict recovery rates from COVID infections, data mining techniques have been successful ways (16). COVID-19 has spread among humans and has threatened human life. So, several types of research have been directed to develop an intelligent medical diagnosis system using AI techniques to manage and control this virus and its effects (17-20). Generally, three different COVID-19 diagnosis methodologies are available as named RT-PCR test, CT scan test, and numerical laboratory test (21). In most of the existing articles on artificial intelligence application in COVID-19 detection, CT scan images used as datasets (22-25) and fewer studies have used laboratory data and clinical-PCR datasets for this purpose. This study was conducted to use Machine Learning (ML) algorithms for early detection of COVID-19 in patients that utilize a clinical dataset and PCR test results.

Methods

Dataset Description and Preprocessing

This retrospective study used a dataset from hospitals affiliated with Shiraz University of Medical Sciences in Iran. These instances contain the records of patients with COVID-19 as well as suspected coronavirus. Finally, based on the results of the PCR test, they were classified into two classes (COVID-19 and non-COVID-19) in the dataset. This dataset was collected in the period March to October 2020 and contained 10055 records with 63 features. To improve the quality of the classification methods, various data preprocessing techniques were used. In the first phase of data cleaning, to find the noises and outliers, we use heuristic methods. Records that were further distant than others were identified as outliers and five percent of them were deleted. The missing values in the dataset reduce the predictive power and produce biased estimates that lead to invalid conclusions (26). Therefore, we used two methods: eliminating data objects and estimating missing ones to control the missing values in the dataset. In order to estimate, the K-NN algorithm with size k = 5 and Euclidean distance criterion was used. In addition, the data is converted to numeric data because the SVM algorithm deals only with numeric data. The dataset was prepared and cleaned so that only the relevant features could be extracted from the original dataset. Since the number of features is large (63 attributes), we must select the effective features. In this study, the Gini index and PCA were used for this purpose. The Gini index was used to weigh the features. In this study, stratified random sampling was used. Stratified random sampling is a way of sampling that involves dividing a population into smaller groups. These groups are named strata. The strata are organized based on the shared attributes of the members of the group. Stratification is the process of classifying the population into groups

In this study, the ten-fold cross-validation strategy was used to find the group with the highest risk of COVID-19. In every ten repetitions, nine parts are used to train, and one part is used to test the performance of the classifier.

In the collected dataset, there is a feature called PCR that shows the result of the PCR test. In this feature, we considered two classes, which include positive and negative PCR test results. Table 1 shows attributes and their data type and some instances of the dataset.

Figures 1, 2, and 3 show the replication of each feature in the dataset.

According to Figure 1, in this study, 4992 records were male (4992/9208, 54%), and 4216 were female (4216/9208, 46%).

Patient ages ranged from a few months to 103 years, with a median age of 56 years and 7 months (Fig. 2). 2689 patients (2689/9208, 29.2%) had an identified history of close contact with family members diagnosed with COVID-19. Fever (4993/9208, 54.2%) and cough (4416/9208, 47.9%) were the most common symptoms.

Supervised Classification Methods

Different types of supervised classification methods in the Rapid Miner software were employed to diagnose COVID-19.

Logistic Regression Classifier

The LR is used to determine the relationship between classified variables versus independent variables (27). LR is used when the dependent variable has two values, such as zero and one, yes or no or true and false. Therefore, it is called binary logistic regression (28). However, when the dependent variable has more than two values, polynomial

[DOI: 10.47176/mjiri.36.110]

Table 1. The dataset features after preprocessing and some instances

Number	Attribute	Data type	Instance 1	Instance 2	Instance 3
1	Age	Integer	43	58	36
2	Gender	Binominal	Male	Female	Female
3	contact with a person who has COVID-19	Binominal	No	No	Yes
4	Fever	Binominal	No	No	No
5	Cough	Binominal	No	Yes	Yes
6	Muscular pain	Binominal	Yes	Yes	Yes
7	Respiratory distress	Binominal	No	No	No
8	Decreased consciousness	Binominal	No	No	No
9	Loss of smell	Binominal	No	No	No
10	Taste loss	Binominal	No	No	No
11	Seizures	Binominal	No	No	No
12	Headache	Binominal	No	Yes	No
13	Dizziness	Binominal	Yes	Yes	No
14	Paresis	Binominal	No	No	No
15	Quadriplegia	Binominal	No	No	No
16	Chest pain	Binominal	No	No	No
17	Skin lesions	Binominal	No	No	No
18	Abdominal pain	Binominal	No	No	No
19	Nausea	Binominal	No	No	No
20	Vomiting	Binominal	No	No	No
21	Diarrhea	Binominal	No	No	No
22	Eating disorder	Binominal	Yes	Yes	No
23	Po2	Binominal	More than 93	More than 93	Less than 93
24	Status	Binominal	Dead	Alive	Alive
25	PCR	Binominal	Positive	Positive	Negative

logistic regression is used. To predict the transformation of dependent variables, a mathematical model using a set of explanatory variables for LR is used.

Suppose the dependent values are numerical one and zero, where zero represents the negative value, and one represents the positive value as a binary variable. Therefore, the mean of the binary variable will be the ratio of the positive values. If p is the ratio of observations to the result 1, 1-P is the probability of the result. The ratio is called p / (1 - p) chance. All 24 variables were entered into the LR classifier as independent variables and the dependent variable is a positive and negative PCR test.

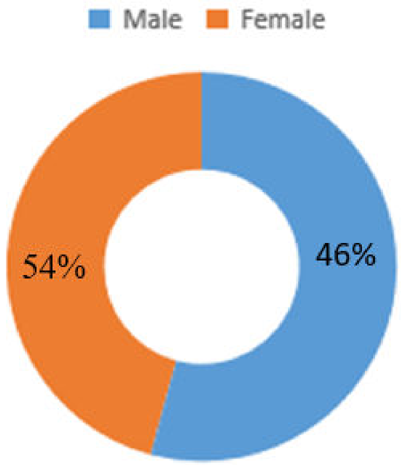

Fig. 1. Frequency of gender attribute in the dataset

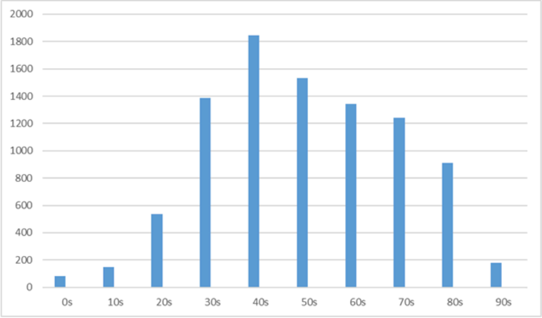

Fig. 2. Frequency of age attribute in the dataset

Support Vector Machine Classifier

SVM is one of the supervised ML algorithms used for classification and regression. The classification task in SVM involves testing and training data that contains some data instances. Each instance in the training set contains one or more target values. Therefore, the main purpose of SVM is to produce a model that predicts the amount of val-

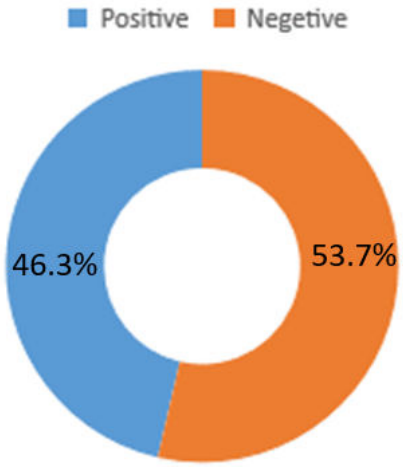

Fig. 3. Frequency of PCR test in the dataset

[DOI: 10.47176/mjiri.36.110]

ues of the target. The basis of the SVM classifier is the linear classification of data, and in the linear segmentation of data, we try to choose the line that has the most reliable margin (29).We divided the dataset into two sets, train and test, with a ratio of 7: 3. SVM separates data linearly using linear kernels and hyper-planes. The purpose of SVM is to find a smaller margin hyper-plane to diagnose patients with Covid19 with appropriate accuracy.

Decision Tree Classifier (C4.5)

The decision tree is used for the classification process in data mining due to its capability to manage batch and continuous data, simplicity, and comprehensibility, and is considered a successful technique. The decision tree consists of two stages growth and pruning. In the first stage, a tree is created by splitting the data into a smaller set until each partition is pure, but the split data type depends only on the data type. The bifurcations for a numerical property C form the value (C) ≤ y, where y is the value in the domain C. To divide the classification D, form the values (D), B ∈ G, where G is a subset of the domain (D). To remove noise in the dataset, the pruning method is used to create the final tree when it is fully grown. The growth stage is computationally more expensive than the decision tree pruning stage (30). Data were divided into training and test sets with a ratio of 7:3.

Naïve Bayes Classifier

In machine learning, a group of simple categorizers based on probabilities is said to be based on Bayes' theorem, assuming the independence of random variables. The Bayesian method is a simple method of classifying phenomena based on the probability of occurrence or non-occurrence of a phenomenon. This method is one of the simplest predicting algorithms with acceptable accuracy. In data mining, the Naive Bayes algorithm is used to separate dataset instances based on specified features (31).

Formula 2: $P\ (c|x) = \frac{P(X|C)P(c)}{P(x)}$

c indicates the desired class.

x indicates the features, each of which must be calculated separately.

P (c | x): Posterior probability class c with predictor x.

P(c): class prior probability.

P (x | c) is the Likelihood probability criterion, which indicates the probability of predicting x having class c.

P(x) .predictor prior probability

Random Forest Classifier

Random forest is a supervised learning algorithm that makes decision trees on instances, then predicts each of them, and finally selects the best solution by voting. This algorithm generates a large number of decision trees during training (32). This method is better than a single decision tree because by averaging the result, it reduces overfitting. Random forest is a method for averaging to reduce variance using deep decision trees created from different parts of educational data.

K-nearest Neighbor Classifier

K-Nearest Neighbor (K-NN) is a nonparametric supervised classification used for regression and classification tasks. In both cases, K contains the closest instructional example in the data space, and its output varies depending on the type used in the classification and regression. K-NN relies on labeled input data to learn to generate good output when entering unlabeled data. In the classification mode, according to the value specified for K, it calculates the distance of the instance we want to label with the nearest neighbors and according to the maximum number of votes of these neighboring points, decides on the label of the desired instance. Different methods are used to calculate this distance, one of the most important of which is the Euclidean distance. In the K-NN classification, the output is a class membership in which instances are classified by a majority vote of neighbors (33).

Classifier Evaluation Metrics

To measure the performance of the classifiers, we applied some evaluation metrics, including accuracy, recall, precision and f-measure. Finally, all these evaluation metrics were compared in terms of performance to get the best algorithm for the diagnosis of COVID-19. Confusion matrix performance metrics are shown in Table 2.

Results

After data cleaning, the final dataset contains 9208 records (4270 positive and 4938 negative). At preprocessing stage, of 63 clinical features, 39 features were excluded from the dataset, and 25 predictors were selected as the input for the ML algorithms. According to the GI index, the most important diagnostic features are: age, gender, contact with an infected, fever, cough, muscular pain, respiratory distress, decreased consciousness, loss of smell, taste loss, seizures, headache, dizziness, paresis, quadriplegia, chest pain, inflammation or skin lesion, abdominal pain, nausea, vomiting, diarrhea, eating disorder, Po2, status and PCR test (Table 1). Based on the C4.5 drawn, the root node is "contact with a person who has COVID-19", which was considered the most important diagnostic criterion based on the Gini index. This classifier has identified the most important symptoms in patients with muscle pain, fever, and cough, loss of smell and chest pain in COVID-19 patients. Also, it has been shown that another important feature for predicting a recovery in COVID-19 patients is the "age"

Table 2. The performance evaluation metrics

Measure	Formula	Intuitive meaning
Precision (P)	TP/(TP + FP)	The percentage of positive predictions those are correct.
Recall/Sensitivity	TP/(TP + FN)	The percentage of positive labeled instances that were predicted as positive.
Specificity	TN/(TN+FP)	The proportion of actual negatives which got predicted as the negative
Accuracy (A)	(TP + TN)/(TP + TN + FP + FN)	The percentage of predictions those are correct.
F-measure	2 * PR/(P + R)	The weighted harmonic mean of Precision and Recall.

Table 3. Performance evaluation of predictive data mining models

Number	Predictive data mining models	Accuracy (%)	Precision (%)	Recall (%)	Sensitivity (%)	Specificity (%)	F-measure (%)
1	Support vector machine	93.41	95	92	92	94.34	91.5
2	Decision tree (C4.5)	91.87	90.85	90.7	90.7	92.2	89.1
3	K-nearest neighbor	89.06	88.2	88.2	88.2	90.1	88.2
4	Logistic regression	88.42	87	87	87	88.91	87.5
5	Random forest	85.69	84.55	85.1	85.1	86.26	85.1
6	Naïve Bayes	83.21	82.7	82.2	82.2	83.95	82.2

feature. Patients between the ages of 65-85 years are at high risk of not recovering from the COVID-19 epidemic. These patients had acute symptoms of COVID-19, while patients between the ages of 26-64 years had milder symptoms of COVID-19. According to C4.5 classifiers, elderly patients are at risk for complications of COVID-19 that may lead to death. The structure of the decision tree provides us with a set of if-then rules which leads to its popularity compared to other classification methods. We interpreted the 3 rules extracted from the C4.5 algorithms as follows:

Rule 1: IF (contact with a person who has COVID-19 == yes AND muscular pain==yes AND fever=yes AND cough=Yes AND loss of smell=yes) THEN COVID-19=1.

Rule 2: IF (contact with a person who has COVID-19 == no AND muscular pain==yes AND fever=yes AND cough=no AND loss of smell=no) THEN COVID-19=1.

Rule 3: IF (contact with a person who has COVID-19 == yes AND muscular pain==yes AND fever=no AND cough=no) THEN COVID-19=0.

Table 3 shows the performance of each classifier.

As shown in Table 3, the SVM with 91.5% f-measure, 92% precision, 92% recall, and 93.41% accuracy yielded better capability in the diagnosis of COVID-19 than other ML algorithms. The Naïve Bayes (F-measures=82.2%) had the worst performance in this respect. The C4.5 decision tree with an accuracy of 91.87%, has the best performance in categorizing the unknown cases after the SVM algorithm.

Discussion

COVID-19 infection is growing rapidly and is still threatening the lives of people; therefore, early detection of COVID-19 patients is critical to the control and treatment of the disease. The literature review shows that no optimal method can be determined so far (34). Increasing emphasis on ML techniques and data mining in the medical scope can deliver a fertile ground for revolution and enhancement. Most of the recent research done has been using ML techniques to detect COVID-19 with CT images. for example; The DarkCovidNet is a COVID-19 automatic detection model based on a deep learning technique that was presented as a new detection method based on the use of chest X-ray images (35). in the study (36), corona patients' detection strategy proposed that based on the most effective and significant features and using enhanced KNN classifier can detect COVID-19 patients. Proposed Convolutional Neural Network(CNN) as a detection model was proposed to accurately detect COVID-19 patients (23). In the study, (6) data mining and ML techniques for Coronavirus (SARS, MERS, and COVID-19) prediction were reviewed. The findings showed that previous researchers used various algorithms. The decision tree algorithm (j48) is the most widely used. Naive Bayes algorithms and SVM are in second place. k-NN was ranked third, and the rest of the algorithms (LR, Latent Dirichlet allocation, Natural language processing, Bayesian belief network, Apriority, Word2Vec) were ranked fourth (6). The results of our study also show that among the various widely used classification methods in the diagnosis of COVID-19, SVM (93.41% accuracy) and C4.5 (91.87% accuracy) have the best results in our data set. Laboratory data show that infected people spread the disease just before they develop symptoms (namely 2 days before they develop symptoms), while those who have never had symptoms can also spread the virus to others (34). According to the C4.5 decision tree, "contact with a person who has COVID-19" was determined as the first divider in identifying the infected person. Among the research that have been done with the aim of diagnosing COVID-19 based on ML algorithms in Iran, we can mention the research of Shanbehzadeh M et al. (37). In order to select the best detection model, they compared 7 decision tree algorithms. According to their findings J-48, with an accuracy of 0.85 has the best performance for diagnosing COVID-19. Based on their dataset, they identified 13 effective features in COVID-19 diagnosis. Based on the Gini index, the lung lesion existence, fever, and history of contact with suspected people are more effective factors. The accuracy of the random forest algorithm (82.5%) in this research is less than our research (87.5%). In our study, "contact whit a person who has COVID-19", was recognized as the most important factor in diagnosing COVID-19. But in Shanbehzadeh M's research, "the lung lesion existence" is the most important factor identified. One of the strengths of our research compared to(37)is high volume dataset. Also, since our dataset is related to a province (Fars province), our results have more integrity than Shanbehzadeh M's research, which used only the data from one hospital. Nopour R et al. (38) also conducted a study in Iran with the aim of detecting COVID-19 based on eight different ML algorithms. Based on their results, J - 48 is more efficient (F - score = 85%) in diagnosis. In this study, they used much less dataset than our research and their dataset is related to a one corona center that has affected the integrity of their model. According to their results, "Lung lesions" with the highest Chi-square are the most important diagnostic criteria. Bayes classifiers in our study (F - score = 82.2%) are more efficient than Nopour R's research (F - score = 76.1%). The LR algorithm is less efficient in their research (F - score = 82%) than in our findings (F - score =

87.5%).Other studies have been performed in Iran to diagnosis COVID-19 using ML algorithms that have used CT images dataset (39) or routine blood tests (40). Other studies based on ML algorithms in Iran in the field of predicting mortality (41, 42) and intubation prediction (43) have been performed. The strength of our research compared to other research is the multi-center data set and the large volume of data set.

Limitation

The integrity of models based on ML algorithms depends on the comprehensiveness of the dataset. Because all analyzes are based on datasets from Fars province, the results of this study are not comprehensive enough to be used nationally. Therefore, intelligent analysis of a national dataset is necessary for the development of intelligent COVID-19 detection systems. As another limitation, our using dataset has nominal data (from PCR test). Recent studies have shown that the use of such nominal data may suffer from false positives or false negatives, which reduces the accuracy of COVID-19 diagnosis.

Conclusion

By performing different ML algorithms on infectious disease datasets, identification of the factors affecting the incidence of infection and the possibility of recovery from various infectious diseases will be feasible.

Because colds, influenza, and other seasonal illnesses are common in the cold season and make it harder to diagnose the coronavirus, artificial intelligence diagnostic models can be effective in saving patients' lives. Due to the fact that in the present study, in the existing dataset, the presence or absence of influenza was not collected in the samples, we were not able to develop a classifier to distinguish COVID-19 from influenza. We suggest that this disease be recorded in the collection of samples so that in future work, we can see the development of models for differentiation and diagnosis of COVID-19 disease from influenza. As another future work, we have decided to use numerical laboratory experiments to develop a fuzzy expert system for COVID-19 diagnosis.

Ethical Statement

This article is taken from the research entitled "Proposed a Covid19 detection model using data mining techniques" Evaluated by: Shiraz University of Medical Sciences, Approval Date: 2020-08-22; Approval ID: IR.SUMS.REC.1399.691.

URL:https://ethics.research.ac.ir/ProposalCertificateEn.php?id=149048&Print=true&NoPrintHeader=true&NoPrintFooter=true&NoPrintPageBorder=true&LetterPrint=true In order to support privacy and confidentiality, we concealed the unique identifying information of people in the data gathering.

Conflict of Interests

The authors declare that they have no competing interests.

References

1. Pourahmadi M, Delavari S, Delavari S. The Role of Empathy in Full-Scale Battle of Medical and Paramedical Learners Against COVID-19. Iran J Med Sci. 2020;45(6):491.
2. Prem K, Liu Y, Russell TW, Kucharski AJ, Eggo RM, Davies N, et al. The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study. Lancet Public Health. 2020;5(5):e261-e70.
3. Salehinejad S, Niakan Kalhori SR, Hajesmaeel Gohari S, Bahaadinbeigy K, Fatehi F. A review and content analysis of national apps for COVID-19 management using Mobile Application Rating Scale (MARS). Inform Health Soc Care. 2020:1-14.
4. Ding Q, Lu P, Fan Y, Xia Y, Liu M. The clinical characteristics of pneumonia patients coinfected with 2019 novel coronavirus and influenza virus in Wuhan, China. J Med Virol. 2020;92(9):1549-55.
5. 2021 [updated 27/1/2021. Available from: http://ird.behdasht.gov.ir/.
6. Albahri A, Hamid RA, Alwan JK, Al-Qays Z, Zaidan A, Zaidan B, et al. Role of biological data mining and machine learning techniques in detecting and diagnosing the novel coronavirus (COVID-19): a systematic review. J Med Syst. 2020;44:1-11.
7. Erfannia L, Amraei M, Arji G, Yazdani A, Sabzehgar M, Yaghoobi L. Reviewing and Content Analysis of Persian Language Mobile Health Apps for COVID-19 Management. Stud Health Technol Inform. 2022;289:106-9.
8. Islam M, Karray F, Alhajj R, Zeng J. A review on deep learning techniques for the diagnosis of novel coronavirus (covid-19). arXiv preprint arXiv: 200804815. 2020.
9. Yazdani A, Sharifian R, Ravangard R, Zahmatkeshan M. COVID-19 and information communication technology: a conceptual model. J Adv Pharm Res 2021;11(S1).
10. Mojarrab S, Rafei A, Akhondzadeh S, Jeddian A, Jafarpour M, Zendehdel K. Diseases and health outcomes registry systems in IR Iran: successful initiative to improve public health programs, quality of care, and biomedical research. Arch Iran Med. 2017;20(11):696-703.
11. Hajesmaeel-Gohari S, Bahaadinbeigy K, Tajoddini S, R Niakan Kalhori S. Minimum data set development for a drug poisoning registry system. Digit Health. 2019;5:2055207619897155.
12. Zahmatkeshan M, Zakerabasali S, Farjam M, Gholampour Y, Seraji M, Yazdani A. The use of mobile health interventions for gestational diabetes mellitus: a descriptive literature review. J Med Life. 2021;14(2):131.
13. Islam MS, Hasan MM, Wang X, Germack HD, editors. A systematic review on healthcare analytics: application and theoretical perspective of data mining. Healthcare; 2018: Multidisciplinary Digital Publishing Institute.
14. Larose DT, Larose CD. Discovering knowledge in data: an introduction to data mining: John Wiley & Sons; 2014.
15. Jang S, Lee S, Choi S-M, Seo J, Choi H, Yoon T, editors. Comparison between SARS CoV and MERS CoV Using apriori algorithm, decision tree, SVM. MATEC Web of Conferences; 2016: EDP Sciences.
16. Al-Turaiki I, Alshahrani M, Almutairi T. Building predictive models for MERS-CoV infections using data mining techniques. J Infect Public Health. 2016;9(6):744-8.
17. Khan AI, Shah JL, Bhat MM. CoroNet: A deep neural network for detection and diagnosis of COVID-19 from chest x-ray images. Comput Methods Programs Biomed. 2020;196:105581.
18. Wang L, Lin ZQ, Wong A. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. Sci Rep. 2020;10(1):1-12.
19. Ye Y, Hou S, Fan Y, Qian Y, Zhang Y, Sun S, et al. $\alpha $-Satellite: An AI-driven System and Benchmark Datasets for Hierarchical Community-level Risk Assessment to Help Combat COVID-19. arXiv preprint arXiv: 200312232. 2020.
20. Muhammad L, Islam MM, Usman SS, Ayon SI. Predictive data mining models for novel coronavirus (COVID-19) infected patients' recovery. SN comput sci. 2020;1(4):1-7.
21. Mansour NA, Saleh AI, Badawy M, Ali HA. Accurate detection of Covid-19 patients based on Feature Correlated Naïve Bayes (FCNB) classification strategy. J Ambient Intell Humaniz Comput. 2022;13(1):41-73.
22. El Asnaoui K, Chawki Y. Using X-ray images and deep learning for automated detection of coronavirus disease. J Biomol Structs 2020:1-12.
23. Marques G, Agarwal D, de la Torre Díez I. Automated medical diagnosis of COVID-19 through EfficientNet convolutional neural

network. Appl Soft Comput. 2020;96:106691.
24. Wang B, Jin S, Yan Q, Xu H, Luo C, Wei L, et al. AI-assisted CT imaging analysis for COVID-19 screening: Building and deploying a medical AI system. Appl Soft Comput. 2021;98:106897.
25. Zhou T, Lu H, Yang Z, Qiu S, Huo B, Dong Y. The ensemble deep learning model for novel COVID-19 on CT images. Appl Soft Comput. 2021;98:106885.
26. Kang H. The prevention and handling of the missing data. Korean J. Anesthesiol. 2013;64(5):402.
27. Kleinbaum DG, Dietz K, Gail M, Klein M, Klein M. Logistic regression: Springer; 2002.
28. De Menezes FS, Liska GR, Cirillo MA, Vivanco MJ. Data classification with binary response through the Boosting algorithm and logistic regression. Expert Syst Appl. 2017;69:62-73.
29. Wang L. Support vector machines: theory and applications: Springer Science & Business Media; 2005.
30. Priyam A, Abhijeeta G, Rathee A, Srivastava S. Comparative analysis of decision tree classification algorithms. Int J Curr Eng Technol. 2013;3(2):334-7.
31. Saritas MM, Yasar A. Performance analysis of ANN and Naive Bayes classification algorithm for data classification. Int J Intell Syst. 2019;7(2):88-91.
32. Rodriguez JJ, Kuncheva LI, Alonso CJ. Rotation forest: A new classifier ensemble method. IEEE Trans Pattern Anal Mach Intell. 2006;28(10):1619-30.
33. Viswanath P, Sarma TH, editors. An improvement to k-nearest neighbor classifier. 2011 IEEE Recent Advances in Intelligent Computational Systems; 2011: IEEE.
34. Shaban WM, Rabie AH, Saleh AI, Abo-Elsoud M. Detecting COVID-19 patients based on fuzzy inference engine and Deep Neural Network. Appl Soft Comput. 2021;99:106906.
35. Ozturk T, Talo M, Yildirim EA, Baloglu UB, Yildirim O, Acharya UR. Automated detection of COVID-19 cases using deep neural networks with X-ray images. Comput Biol Med. 2020;121:103792.
36. Shaban WM, Rabie AH, Saleh AI, Abo-Elsoud M. A new COVID-19 Patients Detection Strategy (CPDS) based on hybrid feature selection and enhanced KNN classifier. Knowl Based Syst. 2020;205:106270.
37. Shanbehzadeh M, Kazemi-Arpanahi H, Nopour R. Performance evaluation of selected decision tree algorithms for COVID-19 diagnosis using routine clinical data. Med J Islam Repub Iran. 2021;35:29.
38. Nopour R, Kazemi-Arpanahi H, Shanbehzadeh M, Azizifar A. Performance analysis of data mining algorithms for diagnosing COVID-19. J Educ Health Promot. 2021;10(1):405.
39. Afshar P, Heidarian S, Enshaei N, Naderkhani F, Rafiee MJ, Oikonomou A, et al. COVID-CT-MD, COVID-19 computed tomography scan dataset applicable in machine learning and deep learning. Sci Data. 2021;8(1):1-8.
40. Mehralian S, Jalaeian Zaferani E, Shashaani S, Kashefinishabouri F, Teshnehlab M, Sokhandan HA, et al. Rapid COVID-19 Screening Based on the Blood Test using Artificial Intelligence Methods. Int J Control. 2021;14(5):131-40.
41. Moulaei K, Ghasemian F, Bahaadinbeigy K, Sarbi RE, Taghiabad ZM. Predicting mortality of COVID-19 patients based on data mining techniques. Biomed Phys Eng Express. 2021;11(5):653.
42. Moulaei K, Shanbehzadeh M, Mohammadi-Taghiabad Z, Kazemi-Arpanahi H. Comparing machine learning algorithms for predicting COVID-19 mortality. BMC Med Inform Decis Mak. 2022;22(1):1-12.
43. Varzaneh ZA, Orooji A, Erfannia L, Shanbehzadeh M. A new COVID-19 intubation prediction strategy using an intelligent feature selection and K-NN method. Inform Med Unlocked. 2021:100825.

[DOI: 10.47176/mjiri.36.110]