Predicting Mortality of COVID-19 Patients based on Data Mining Techniques

Khadijeh Moulaei[1], Fahimeh Ghasemian[2], Kambiz Bahaadinbeigy[3]*, Roghayeh Ershad Sarbi[4], Zahra Mohamadi Taghiabad[5]

[1]PhD Candidate, Medical Informatics Research Center, Institute for Futures Studies in Health, Kerman University of Medical Sciences, Kerman, Iran
[2]PhD, Department of Computer Engineering, Faculty of Engineering, Shahid Bahonar University Kerman, Kerman, Iran
[3]MD, PhD, Medical Informatics Research Center, Institute for Futures Studies in Health, Kerman University of Medical Sciences, Kerman, Iran
[4]PhD, Faculty of Management and Medical Information Sciences, Kerman University of Medical Sciences, Kerman, Iran
[5]MSc, School of Health Management and Information Sciences, Iran University of Medical Sciences, Tehran, Iran

ABSTRACT

If Coronavirus (COVID-19) is not predicted, managed, and controlled timely, the health systems of any country and their people will face serious problems. Predictive models can be helpful in health resource management and prevent outbreak and death caused by COVID-19. The present study aimed at predicting mortality in patients with COVID-19 based on data mining techniques. To do this study, the mortality factors of COVID-19 patients were first identified based on different studies. These factors were confirmed by specialist physicians. Based on the confirmed factors, the data of COVID-19 patients were extracted from 850 medical records. Decision tree (J48), MLP, KNN, random forest, and SVM data mining models were used for prediction. The models were evaluated based on accuracy, precision, specificity, sensitivity, and the ROC curve. According to the results, the most effective factor used to predict the death of COVID-19 patients was dyspnea. Based on ROC (1.000), accuracy (99.23%), precision (99.74%), sensitivity (98.25%) and specificity (99.84%), the random forest was the best model in predicting of mortality than other models. After the random forest, KNN5, MLP, and J48 models were ranked next, respectively. Data analysis of COVID-19 patients can be a suitable and practical tool for predicting the mortality of these patients. Given the sensitivity of medical science concerning maintaining human life and lack of specialized human resources in the health system, using the proposed models can increase the chances of successful treatment, prevent early death and reduce the costs associated with long treatments for patients, hospitals and the insurance industry.

Citation: Moulaei Kh, Ghasemian F, Bahaadinbeigy K, Ershad Sarbi R, Mohamadi Taghiabad Z. Predicting Mortality of COVID-19 Patients based on Data Mining Techniques. *J Biomed Phys Eng*. 2021;11(5):653-662. doi: 10.31661/jbpe.v0i0.2104-1300.

Keywords

Mortality; COVID-19; Data Mining; Prediction

Introduction

The rapid outbreak of Coronavirus (COVID-19) disease in December 2019 in China became a global health emergency [1]. The virus can cause a wide range of illnesses, ranging from cold to acute respiratory symptoms, and can lead to death due to pneumonia and respiratory problems [2]. Also, COVID-19 patients have a low psychological tolerance capacity and are highly prone to psychological disorders such as anxiety, fear, depression and negative thoughts [3].

On the other hand, COVID-19 imposes great health, economic and social challenges for countries [4]. In addition to putting much pressure on health care providers, it has also led to increased health care costs [5]. Thus, if the outbreak of this disease is not managed and controlled

*Corresponding author:
Kambiz Bahaadinbeigy
Department of Health Information Management and Technology, Medical Informatics Research Center, Institute for Futures Studies in Health, Kerman University of Medical Sciences, Kerman, Iran
E-mail: kambizb321@gmail.com

Received: 6 April 2021
Accepted: 20 May 2021

timely, health systems will face serious problems due to the lack of medical staff and medical equipment [6]. Given these challenges and the rapid spread of the disease worldwide, it should be noted that predictive models can be helpful in the management of health resources and planning for the prevention of COVID-19 [7]. Early diagnosis of diseases can lead to timely intervention and reduction of patient mortality. Also, these models can be used as a guideline in prioritizing of patients, supporting the clinical decision, evaluating care quality, controlling care quality, and standardizing and optimizing care [8].

Data mining models and techniques are well-known tools for developing predictive and data analysis models. These techniques can implicitly extract useful information from raw data [7]. Some classification techniques that are used in data mining for prediction include Decision Tree (J48), random forest, K Nearest Neighborhood (KNN), Multi-Layer Perceptron (MLP), and Support Vector Machine (SVM) [9]. These data mining tools analyze data and create practical models in various fields such as knowledge base and determine strategy for the business and scientific medical research [10]. These models can help health policymakers and managers plan health resources and prevent the spread of epidemics such as COVID-19 [7]. The results of a study conducted by Mengistie showed that data mining techniques and related technologies have greatly influenced our daily lives and also have been effective in helping humans fight against COVID-19 [11]. Due to the availability of large amounts of data, the urgent need to extract knowledge from this data, and the high costs imposed on health institutions, it is expected that in addition to a clinical evaluation system, a dedicated prediction system based on data mining methods is needed to optimize care for patients with COVID-19. Thus, after identifying the effective factors involved in the mortality of patients with COVID-19, the results of five models based on data mining techniques(decision tree (J48), multilayer perceptron, KNN, random forest, and SVM) were compared. Finally, the most optimal model for predicting the mortality rate of patients with COVID-19 was identified and introduced. Also, this study tries to estimate the mortality rate among COVID-19 patients by extracting appropriate features with higher accuracy and precision. Since the diagnosis of the disease by a human is a difficult, time-consuming, and error-prone process, conducting this study can increase the speed, reduce error, and facilitate diagnostic and therapeutic decision-making processes, that makes it possible to plan for service provision in the Intensive Care Units (ICU) of patients with COVID-19 (for example, the allocation of ICU beds to patients with more acute conditions). It finally results in a reduction in treatment costs and improved health.

Material and Methods

The present study was conducted based on the proposed model, i.e. the Cross-Industry Standard Process (CRISP) methodology (Figure 1), which includes steps of system recognition, data understanding, data preparation, modeling, and deployment.

Step1: Data Identification

In the first step, to identify the predictive factors of mortality of COVID-19 patients, various studies in this field were reviewed [1, 2, 4, 6, 7]. These five studies were approved in accordance with the opinion of two infectious disease specialists. They believed that these studies provided complete and comprehensive information on the predictive factors of mortality of COVID-19 patients. Thus, factors identified in each study were recorded through a data extraction form. Then the researchers examined 50 record samples of COVID-19 patients in Valiasr Hospital affiliated to Ilam University of Medical Sciences, Iran (Darreh Shahr city) to be sure about these factors. Data were extracted from medical records based on

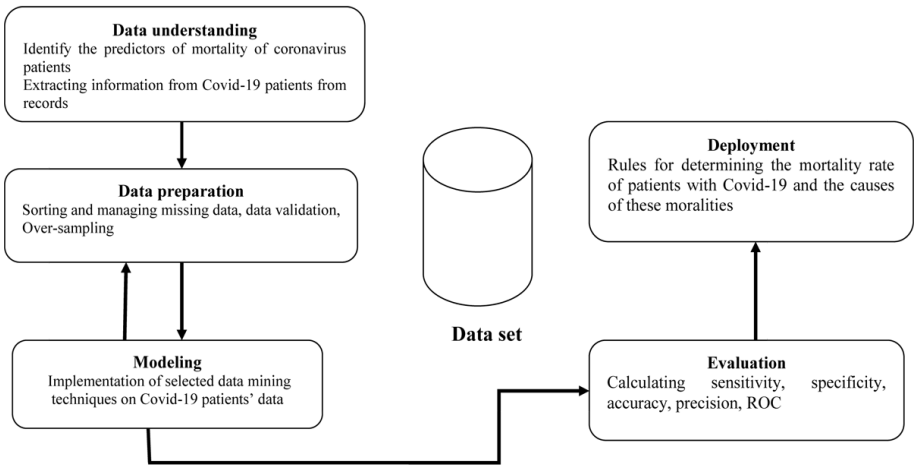

Figure 1: Proposed model of research steps based on the Cross-Industry Standard Process (CRISP) model.

a data extraction form.

This data extraction form consisted of fields of the row, medical record number, COVID-19 mortality prediction factors, ward name, and hospital name. After extracting the effective factors, duplicate factors were removed and factors with different names were homogenized. Finally, a final list of factors was prepared. In the next step, for final confirming of factors, ten invitations were sent via e-mail to specialists in three fields of internal medicine, pulmonary and infectious diseases, working in the hospitals of Ilam University of Medical Sciences. The invitation was sent to physicians who had experience of working on COVID-19 patients. Six physicians responded to our invitation to participate in the study (internal medicine (n=3), pulmonary (n=1), and infectious diseases (n=2)). In the next step, for final confirming of factors, two brainstorming sessions were held through WhatsApp. The sessions lasted one hour on September 30 and October 1.

In the next step, a total of 850 records (650 living patients and 250 dead patients) were examined from three hospitals affiliated to Ilam University of Medical Sciences, Iran (including Shahid Mostafa Khomeini, Hazrat Valiasr hospital, and Imam Hossein hospital). The medical records belonged to the patients with a positive COVID-19 test. They referred to hospitals from March 5 to September 22 of 2020. The data collection tool was a researcher-made checklist designed according to the research objectives. Data were collected according to this checklist, which consisted of fields of the row, medical record number, mortality prediction factors for COVID-19 patients, ward name, and hospital name. Necessary data were collected from the hospitals by the researcher. The inclusion criteria for this study were patients with positive COVID-19 tests and who living in the Ilam province. All the data collected through the checklist were entered into an Excel file.

Step 2: Pre-processing the data

After collecting data, they were sorted and managed according to the type and mode of data. Records that had missing values or had no resemblance to other data (Outlier) were deleted. Also, the variable of patients' medical record numbers was deleted. Patients' names were anonymously placed in the database (with the number one to 850). Then the data were prepared for analysis, and the data set created for data preprocessing, which is missing value management and data valida-

tion, was re-examined in more detail. Data entered outside the normal or possible range were re-examined and corrected. Also, since the sample size consisted of 200 deceased patients and 650 living patients, the data set was imbalanced, i.e. the number of records in one class was higher than in the others (the number of samples in a larger class could be up to twice as many as in the other classes). As a result, the prediction of classification models is biased towards the class that has a larger sample size. Thus, the values of the performance indicators of the classification models will decrease. To solve this problem, two methods of under-sampling or over-sampling are defined. In under-sampling, the larger class size decreases to be equal with a smaller class, and in over-sampling, the smaller class size increases to be equal with the larger class. In this study, over-sampling was used to balance the records of the deceased class with the records of the living class, so that the class with the lowest number of samples is oversampled and balanced. Then data mining models were performed on the balanced samples.

Step 3: Implementing the selected data mining techniques

In this step, to select data mining models for data analysis, various studies in this field were reviewed [12-14]. According to the studies and the type and quality of data, appropriate models were selected. In this study, decision tree (J48), multilayer perceptron, KNN, random forest, and SVM models were used. It should be noted that to reduce the minimum distance between the query instance and the training samples, the KNN model was applied to the data set three times [15].

17 leaves were used to create a decision tree. Finally, a tree with a size of 31 was formed. The test option selected for the decision tree was Validation-Fold Cross-10. Experiments have shown that the best choice for getting the most accurate estimate is Validation-Fold Cross-10 [16]. To create the neural network, its usual structure, MLP, was used. Neural network inputs and outputs were effective factors in predicting mortality and the target variable or patient death, respectively. Therefore, the neural network used in the present study consisted of 1 input, 11 nodes in the hidden layer, and 1 output. Samples of clinical data mining studies were cited to determine the hidden layers [17, 18]. These studies emphasized the ability of two-layer neural networks. It should be noted that one of these layers was the output layer and another one was the hidden layer. For KNNs, 1, 3, and 5 neighbors were used. For random forest analysis, bagging with 100 iterations and base learner were used. As the decision tree algorithm in this model, 10 Fold Cross Validation was used to obtain the most accurate estimate [16]. To create each of these models, sample studies that used these models in different clinical areas were reviewed.

Step 4: Evaluating and validating the performance of the models

Sensitivity, specificity, accuracy, precision, and area under the ROC curve indices were used to evaluate the performance of predictive models. These criteria will be defined and calculated using the confusion matrix compo-

Table 1: Confusion matrix

Output		Prediction value	
		Death(+)	Living(-)
Real value	Death(+)	TP	FN
	Living(-)	FP	TN

Note: True positive (TP): The number of deaths that the model has correctly identified.

False positive (FP): The number of living people but the model has incorrectly identified them as dead.

True negative (TN): The number of people who are living and the model correctly identified them as living.

True positive (TP): The number of deaths that the model has correctly identified.

False negative (FN): The number of people who are dead but the model has identified them as living incorrectly

nents (Table 1). Accuracy refers to the number of living and dead people who have been diagnosed as living or dead correctly. Precision refers to the number of people who have died and the model has correctly identified them. Sensitivity refers to the proportion of people who have died and the model has correctly identified them as dead people. Therefore, the larger value indicates a more accurate diagnosis of the dead people. Specificity is the proportion of people who are living and the model has correctly identified as living [16]. Receiver operating characteristic (ROC) is also often used as an indicator to determine the power of a model. Also, in the medical field, the area under the ROC curve is used to evaluate the accuracy of diagnostic tests [17].

Finally, for all the selected models, their performance was reported separately. We used Weka v3.9.2 software to analyze the data, identify the importance of each factor in predicting patient mortality, implement patient mortality prediction models, and draw a confusion matrix. Weka is a set of machine learning algorithms for data mining and data analysis. On the other hand, this software program includes tools for data preparation, classification, regression, clustering, rule extraction and visualization.

Step 5: Development

This step includes presenting the rules created in determining the probability of mortality of COVID-19 patients and the factors related to the mortality of these patients.

To conduct the present study (code: 99000245), the code of ethics of IR.KMU.REC.1399.329 was obtained from the ethics committee of Kerman University of Medical Sciences. The anonymity of COVID-19 patients was maintained by deleting the medical record number and using the number one to 850.

Technical presentation

The research results are presented based on the research steps.

Identifying and pre-processing data

In total, according to the reviewed studies and the opinion of experts, 16 factors were finally confirmed to predict the mortality of COVID-19 patients. Table 2 shows the importance of the 16 factors used. Dyspnea was the most effective factor to predict mortality in COVID-19 patients. The gender factor also had the least effect compared to other factors.

The required data set was extracted from 850 medical records of patients with coronary artery disease in four hospitals. The mean age of dead and living patients was 53.2 and 44.8, respectively. In terms of gender, 63.9% were male and 36.1% were female. The dead people accounted for 23.5% and living people accounted for 76.5% of the study population. Died males and females were 15.2 and 8.3 per-

Table 2: Effective factors used in the mortality prediction database of patients with Covid 19.

Row	Factors name	Degree of importance
1.	Dyspnea	0.594
2.	Underlying diseases	0.470
3.	Headache	0.450
4.	Weakness and lethargy	0.398
5.	Body pain	0.314
6.	Fatigue	0.261
7.	Sore throat	0.252
8.	Age	0.214
9.	Dry cough	0.166
10.	Diarrhea	0.148
11.	Pain or pressure in the chest	0.073
12.	High fever	0.066
13.	Loss of sense of smell and taste	0.056
14.	Nausea and Vomiting	0.047
15.	Anorexia	0.032
16.	Gender	0.004

cent, respectively. Also, 48.7% and 27.8% of males and females, respectively, were living.

Evaluating and validating the performance of models

In this step, the performance of the decision tree (J48), MLP, KNN, and random forest models was evaluated. Then their sensitivity, specificity, accuracy, precision, and ROC curve indicators were reported. A comparison of these indicators for each model is presented in Table 3. Sensitivity, accuracy, and ROC curve indicators of random forest were higher than other models. Also, the specificity and accuracy in KNN2 were better than other models. The specificity and accuracy of this model were reported at 100%. J48 model was found as the weakest model based on ROC curve.

According to the Table 3, among the models, KNN and KNN2 performed better than the two models of KNN1 and KNN3.

The performance of the selected models based on sensitivity, specificity, accuracy, precision, and ROC is shown in Figure 2.

Discussion

In the present study, based on retrospec-

Table 3: Performance evaluation of selected models.

Models	Performance of each model				
	Sensitivity (%)	Specificity (%)	Accuracy (%)	Precision (%)	ROC
Decision Tree (J48)	98	97.38	97.61	95.84	0.980
Multilayer perceptron	95.25	98.76	97.42	97.94	0.989
KNN1	95. 25	97.84	96.85	96.45	0.992
KNN2	97.75	100	99.14	100	0.987
KNN3	95	95.23	95.14	92.45	0.993
Random forest	98.25	99.84	99.23	99.74	1.000
SVM	98	96	96.47	93.73	0.966

ROC: Receiving Operating Characteristics, KNN: K Nearest Neighborhood, SVM: Support Vector Machine

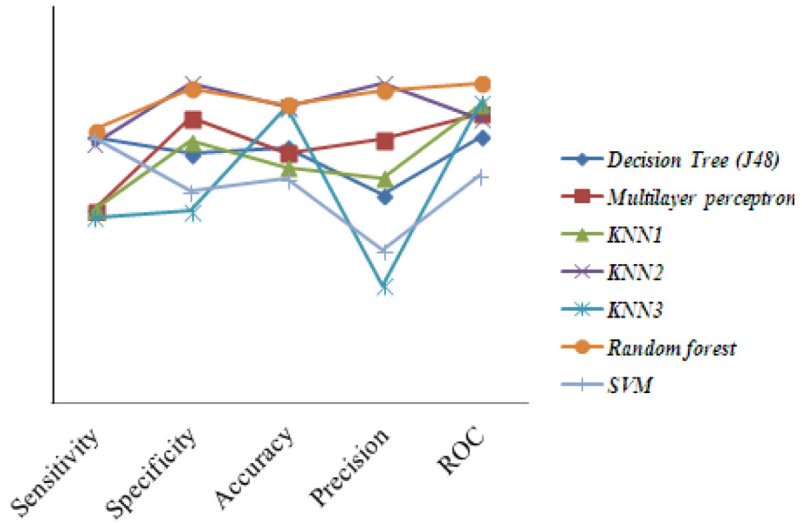

Figure 2: Performance evaluation of selected models

tive data, the mortality of COVID-19 patients was predicted. The total number of predictors of mortality in patients with COVID-19 disease was 16. Patients' mortality was predicted based on a data set extracted from 850 medical records (200 dead and 650 living patients) of patients with a positive corona test. Four data mining models of the decision tree (J48), multilayer perceptron, KNN, and random forest were used to predict patient mortality. The random forest was the best model in predicting mortality than other models. KNN5, MLP, and J48 models were ranked next, respectively. Based on the results, COVID-19 mortality prediction models can be presented with high accuracy.

Muhammad et al. [13] identified patients with COVID-19 infection based on data mining. In this study, in addition to different models of support vector machine, naive Bayes, logistic regression, as in the present study, random forest, decision tree, and KNN were used. The data set of this study (1505) [13] was larger than the present study and formed based on five factors of gender, age, infectious cases, and patients' condition (living or dead). In this study, the decision tree with 99.88% accuracy had the highest accuracy in identifying patients with COVID-19 infection. The highest accuracy in the present study belonged to KNN2. Muhammad et al. [13] believed that SVM controls nonlinear input spaces and separates data points using a hyperplane with the highest amount of margins. They also stated that SVM as a discriminative classifier was able to find an optimal hyperplane for their data and help classify new unannotated data points. Due to some efficient characteristics of k-NN in the predictions, some studies have used it. Weinberger and Saul [19] believed that k-NN improves classification accuracy. Kubota et al. [20] introduced a hierarchical model developed based on k-NN. The efficiency and high sensitivity of this model in discriminating small classes was noteworthy.

In contrast to the present study, which predicted mortality of COVID-19 patients based on symptoms, signs, laboratory tests, and demographic characteristics, Mousavi et al. [14] identified high-risk COVID-19 patients based on only laboratory tests. In this study, the sample size (n=4542) was higher than the present study. However, unlike the present study, in which all people with COVID-19 disease were included at any age, in this study, only the COVID-19 patients aged over 18 years were studied. Also, in contrast to the present study, which used four different data mining models, in the study conducted by Mousavi et al. [14], logistic regression was used to compare recovered and dead patients in the SPSS environment. The highest accuracy in the present study was reported at 99.14% using the KNN2 algorithm. The accuracy of the regression model used in the study conducted by Ahouz et al. [18] was obtained at 83% and in the study conducted by Li et al. [21], it was obtained at 0.8698 with the SVM (Linear kernel) model. In the study conducted by Ahouz et al. [18], 17136 records and 4 variables (latitude, longitude, history, and background) were analyzed. Li et al. [21] also identified stress symptoms in people due to the COVID-19 outbreak using data mining. In this study, SVM (Linear kernel), logistic regression, naive Bayes, and simple neural network models were used. The collected data contained 80 million tweets.

In reviewing the above studies, we found that none of these studies focused on predicting mortality in patients with COVID-19. These studies have been conducted with different aims of identifying patients with COVID-19 infection [13], predicting death in patients with suspected sepsis [22], and identifying stress symptoms in people due to outbreak of COVID-19 [21], and patients with high-risk COVID-19 [14] and predicting the rate of COVID-19 in the next two weeks [18]. Regarding the importance of predicting patients' mortality, Deschepper et al. [23] stated that early prediction of mortality in hospitals can improve the patient outcomes and enable health care

providers to take adequate and timely action to save lives. Bhattacharya also believed that identifying patients at risk of death could lead to vital decisions such as discontinuing treatment, using the necessary equipment, assessing medical risks, tracking the resources needed by the intensive care unit, and reducing the length of hospital stay in the intensive care unit [24].

Also, as mentioned in the studies [13, 18, 21], some different predictors have been used in the development of the models. All of these studies used fewer factors to develop their models compared to the present study. These studies have shown that different algorithms have different performance in different conditions (sample size and number of different predictors). Therefore, based on the results of the mentioned studies, it can be stated that with decreasing the number of predictors, the accuracy of the models decreases. The present study also introduced dyspnea and underlying diseases as the most effective in predicting mortality in COVID-19 patients. Li et al. [25], as in the present study, identified clinical factors over 50 years of age, underlying diseases, and dyspnea as three risk factors for severe/critical COVID-19 pneumonia. Liguoro et al. [26] examined the SARS-COV-2 infection and found that dyspnea was the most commonly reported symptom in infancy that can be life-threatening. Shi et al. [27] also showed in their study that shortness of breath was positively associated with an increased risk of mortality in COVID-19 patients. Some studies on COVID-19 patients have also shown that people with underlying diseases are not only at higher risk for developing the disease, but also more likely to die from the virus infection compared to others [28]. Therefore, based on the results of the present study and other studies mentioned, it can be stated that with increasing the severity of shortness of breath and underlying diseases, the mortality rate among COVID-19 patients will be higher.

One of the limitations of the study is the small sample size in evaluating the models. It is recommended to use a larger sample size in future studies or conduct studies in other provinces. Also, in this study, only five data mining models and 16 effective factors were used to predict mortality in patients with COVID-19. Thus, it is recommended that more effective factors along with more data mining models be used to predict the mortality of COVID-19 patients. Another limitation of the study was the non-use of CT scan data, and a study should be thus conducted to use a combination of symptoms, sign, and CT scan data. Also, like all data mining models, these models can be viewed as "black box" models, i.e. there is little knowledge of the way factors, playing a role in predicting patient mortality. This problem can be reduced by ranking predictors after their contribution to the total AUC.

Conclusion

In the present study, five mortality prediction models of COVID-19 patients based on data mining techniques were compared. The results showed that the use of data mining techniques can be an efficient way to predict mortality in patients with COVID-19. Thus, considering the critical role of medical science in maintaining human life and the lack of specialized human resources in the health system, the proposed models can provide the necessary services to patients by diagnosing death earlier according to the identified factors. Thus, the chances of successful treatment can be increased, early death can be prevented, service providers can help to prioritize and allocate hospital resources, and the costs of long-term treatment imposed on patients, hospital, and insurance industry can be reduced.

Acknowledgment

The authors would like to thank all experts who freely participated in this study.

Conflict of Interest

None

References

1. Kazemi-Arpanahi H, Moulaei K, Shanbehzadeh M. Design and development of a web-based registry for Coronavirus (COVID-19) disease. *Med J Islam Repub Iran*. 2020;**34**:68. doi: 10.34171/mjiri.34.68. PubMed PMID: 32974234. PubMed PMCID: PMC7500427.
2. Mullins E, Evans D, Viner RM, O'Brien P, Morris E. Coronavirus in pregnancy and delivery: rapid review. *Ultrasound Obstet Gynecol.* 2020;**55**(5):586-92. doi: 10.1002/uog.22014. PubMed PMID: 32180292.
3. Yao H, Chen JH, Xu YF. Patients with mental health disorders in the COVID-19 epidemic. *Lancet Psychiatry.* 2020;**7**(4):e21. doi: 10.1016/S2215-0366(20)30090-0. PubMed PMID: 32199510. PubMed PMCID: PMC7269717.
4. Kuwahara K, Kuroda A, Fukuda Y. COVID-19: Active measures to support community-dwelling older adults. *Travel Med Infect Dis*. 2020;**36**:101638. doi: 10.1016/j.tmaid.2020.101638. PubMed PMID: 32205272. PubMed PMCID: PMC7270647.
5. Moazzami B, Razavi-Khorasani N, Moghadam AD, Farokhi E, Rezaei N. COVID-19 and telemedicine: Immediate action required for maintaining healthcare providers well-being. *J Clin Virol*. 2020;**126**:104345. doi: 10.1016/j.jcv.2020.104345. PubMed PMID: 32278298. PubMed PMCID: PMC7129277.
6. Keshvardoost S, Bahaadinbeigy K, Fatehi F. Role of telehealth in the management of COVID-19: lessons learned from previous SARS, MERS, and Ebola outbreaks. *Telemed J E Health*. 2020;**26**(7):850-2. doi: 10.1089/tmj.2020.0105. PubMed PMID: 32329659.
7. Ayyoubzadeh SM, Ayyoubzadeh SM, Zahedi H, Ahmadi M, Kalhori SR. Predicting COVID-19 incidence through analysis of google trends data in iran: data mining and deep learning pilot study. *JMIR Public Health Surveill*. 2020;**6**(2):e18828. doi: 10.2196/18828. PubMed PMID: 32234709. PubMed PMCID: PMC7159058.
8. Grissom CK, Brown SM, Kuttler KG, Boltax JP, et al. A modified sequential organ failure assessment score for critical care triage. *Disaster Med Public Health Prep*. 2010;**4**(4):277-84. doi: 10.1001/dmp.2010.40. PubMed PMID: 21149228. PubMed PMCID: PMC3811929.
9. Hadi WE, El-Khalili N, AlNashashibi M, Issa G, AlBanna AA. Application of data mining algorithms for improving stress prediction of automobile drivers: A case study in Jordan. *Comput Biol Med*. 2019;**114**:103474. doi: 10.1016/j.compbiomed.2019.103474. PubMed PMID: 31585402.
10. Bramer M. Principles of data mining. London: Springer; 2007.
11. Mengistie TT. COVID-19 Outbreak Data Analysis and Prediction Modeling Using Data Mining Technique. *International Journal of Computer (IJC)*. 2020;**38**(1):37-60.
12. Rivo E, De La Fuente J, Rivo Á, et al. Cross-Industry Standard Process for data mining is applicable to the lung cancer surgery domain, improving decision making as well as knowledge and quality management. *Clin Transl Oncol.* 2012;**14**(1):73-9. doi: 10.1007/s12094-012-0764-8. PubMed PMID: 22262722.
13. Muhammad LJ, Islam MM, Usman SS, Ayon SI. Predictive data mining models for novel coronavirus (COVID-19) infected patients' recovery. *SN Comput Sci.* 2020;**1**(4):206. doi: 10.1007/s42979-020-00216-w. PubMed PMID: 33063049. PubMed PMCID: PMC7306186.
14. Mousavi A, Rezaei S, Salamzadeh J, Mirzazadeh A, Peiravian F, Yousefi N. Value of laboratory tests in COVID-19 hospitalized patients for clinical decision-makers: a predictive model, using data mining approach. *Research Square.* 2020. doi: 10.21203/rs.3.rs-56252/v1.
15. Daberdaku S, Tavazzi E, Di Camillo B. A combined interpolation and weighted K-nearest neighbours approach for the imputation of longitudinal ICU laboratory data. *Journal of Healthcare Informatics Research*. 2020;**4**(2):1-15. doi: 10.1007/s41666-020-00069-1.
16. Liu J, Lan H, Fu Y, Wu H, Li P. Analyzing electricity consumption via data mining. *Wuhan University Journal of Natural Sciences.* 2012;**17**(2):121-5. doi: 10.1007/s11859-012-0815-6.
17. Kim S, Kim W, Park RW. A comparison of intensive care unit mortality prediction models through the use of data mining techniques. *Healthc Inform Res.* 2011;**17**(4):232-43. doi: 10.4258/hir.2011.17.4.232. PubMed PMID: 22259725. PubMed PMCID: PMC3259558.
18. Ahouz F, Golabpour A. Predicting the incidence of COVID-19 using data mining. *BMC Public Health.* 2020. doi: 10.21203/rs.3.rs-21247/v3.
19. Weinberger KQ, Saul LK. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research.* 2009;**10**(2):207-44.
20. Kubota R, Uchino E, Suetake N. Hierarchical

k-nearest neighbor classification using feature and observation space information. *IEICE Electronics Express.* 2008;**5**(3):114-9. doi: 10.1587/elex.5.114.

21. Liu D, Li L, Wu X, Zheng D, Wang J, Yang L, Zheng C. Pregnancy and perinatal outcomes of women with coronavirus disease (COVID-19) pneumonia: a preliminary analysis. *AJR Am J Roentgenol.* 2020;**215**(1):127-32. doi: 10.2214/AJR.20.23072. PubMed PMID: 32186894.
22. Jaimes F, Farbiarz J, Alvarez D, Martínez C. Comparison between logistic regression and neural networks to predict death in patients with suspected sepsis in the emergency room. *Crit Care.* 2005;**9**(2):R150-6. doi: 10.1186/cc3054. PubMed PMID: 15774048. PubMed PMCID: PMC1175932.
23. Deschepper M, Waegeman W, Vogelaers D, Eeckloo K. Using structured pathology data to predict hospital-wide mortality at admission. *PLoS One*. 2020;**15**(6):e0235117. doi: 10.1371/journal.pone.0235117. PubMed PMID: 32584872. PubMed PMCID: PMC7316243.
24. Bhattacharya S, Rajan V, Shrivastava H. ICU mortality prediction: a classification algorithm for imbalanced datasets. *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence.* 2017;**31**(1):1288–94.
25. Li K, Wu J, Wu F, Guo D, Chen L, Fang Z, Li C. The clinical and chest CT features associated with severe and critical COVID-19 pneumonia. *Invest Radiol.* 2020;**55**(6):327-31. doi: 10.1097/RLI.0000000000000672. PubMed PMID: 32118615. PubMed PMCID: PMC7147273.
26. Liguoro I, Pilotto C, Bonanni M, Ferrari ME, et al. SARS-COV-2 infection in children and newborns: a systematic review. *Eur J Pediatr.* 2020;**179**(7):1029-46. doi: 10.1007/s00431-020-03684-7. PubMed PMID: 32424745. PubMed PMCID: PMC7234446.
27. Shi L, Wang Y, Wang Y, Duan G, Yang H. Dyspnea rather than fever is a risk factor for predicting mortality in patients with COVID-19. *J Infect.* 2020;**81**(4):647-79. doi: 10.1016/j.jinf.2020.05.013. PubMed PMID: 32417316. PubMed PMCID: PMC7228739.
28. Verity R, Okell LC, Dorigatti I, Winskill P, et al. Estimates of the severity of coronavirus disease 2019: a model-based analysis. *Lancet Infect Dis*. 2020;**20**(6):669-77. doi: 10.1016/S1473-3099(20)30243-7. PubMed PMID: 32240634. PubMed PMCID: PMC7158570.