Determining the efficiency of data analysis systems in predicting COVID-19 infected cases

Pegah Kalantar Shahpoori¹, Abaset Mirzaei²

¹Department of Health Care Management, Faculty of Health, ²Health Economic Policy Research Center, Tehran Medical Science, Islamic Azad University, Tehran, Iran

ABSTRACT

After the outbreak of the novel coronavirus disease (2019) (COVID-19), a lot of people have been affected around the world. Due to the large number of affected patients in the world, the global health care system has been disrupted and nearly all hospitals around the world has faced a shortage of bed spaces. As a consequence, being able of prediction of the number of COVID-19 cases is extremely important for taking appropriate decision for management of the affected patients. An accurate prediction of the number of COVID-19 cases can be obtained using the historical data of reported cases as well as some other data affecting the virus outbreak. However, most of the literature has used only historical data to provide a method of predicting COVID-19 cases and has neglected other influential factors. This has led to inaccurate estimates of the number of infected cases with COVID-19. Thus, the present study tries to provide a more accurate estimation of the number of COVID-19 cases by considering both historical data and other effective factors on the virus. For this purpose, data analysis including the development of a network-based neural algorithm [i.e., nonlinear autonomous exogenous input (NARX)] can be adopted. To examine the viability of this algorithm, experiments were conducted using data collected for the number of COVID-19 cases in the five most affected countries on each continent. Our method led to a more accurate prediction than those obtained by the existing methods. Moreover, we performed experiments to extend our method to predict the number of COVID-19 cases in the future during a period between August 2020 and September 2020. Such predictions can be utilized by the government or people in the affected countries to take precautionary measures against the pandemic.

Keywords: COVID-19, data analysis, neural network, pandemic

INTRODUCTION

The outbreak of COVID-19 has caused concern around the world, because this virus not only causes disease but also leads to the death of many people.¹ The World Health Organization (WHO) declared COVID-19 disease as a global pandemic on March 11.² There is currently no effective treatment, vaccines have recently been reported whose effectiveness is being debated. In addition, the symptoms of this disease are nonspecific or may not be seen at all. Different communities have presented challenges to control and prevent this disease. A recent systematic and meta-analysis of 212 epidemiological studies and 281,461 reported a mortality rate of about 6%, with Italy having the highest mortality rate (14%).³ In Iran, cases were first reported in Qom in February 2020.⁴ Immediately after that, more cases were identified in other provinces of the country, some provinces were greatly affected in the early outbreak of the disease, including Tehran, Gilan, Mazandaran, Qom, and Golestan.⁵ Since the outbreak of the epidemic, Iran has pursued a wide range of control policies to reduce the prevalence of the epidemic.⁶ Collective immunity is the indirect protection given to vulnerable individuals in the presence of relatively large numbers of immune individuals.⁷ This immunity can be achieved naturally or through a vaccination program, if a vaccine is available. So far, however, no drug has been reported for treatment.⁸⁹ Due to the shortage of vaccines

How to cite this article: Shahori PK, Mirzaei A. Determining the efficiency of data analysis systems in predicting COVID-19 infected cases. J Family Med Prim Care 2022;11:2405-10.
and the long duration of injections in the entire population of the world, and the fact that as of April 3, 2021, only 124,000 people in the population of Iran have been vaccinated,[8] other non-clinical treatment techniques to control and prohibit further spread of COVID-19, like data mining approaches, requires machine learning.[9]

The development of artificial intelligence (AI) has gradually changed medical practices. Recent advances in machine learning (ML), digital data collection, and computing infrastructure have expanded the use of AI in areas previously practiced only by human.[9] DM is an advanced AI technique used to discover hidden patterns or new, useful, and valid data sets.[10] This technique demonstrates knowledge, associations, or patterns in a single or multiple datasets. Also, it has been a common technique used to predict and diagnose a variety of viral illnesses, including SARS-CoV and MERS-CoV, which occurred in 2003 and 2012, respectively.[11]

Because the vast body of data generated worldwide on the comprehensive corona epidemic is a valuable resource for extracting knowledge, useful, valid, and innovative models are essential in making better decisions to prevent the spread of the disease. Data mining has found various applications in healthcare sectors, for example, health outcome modelings, disease outcome predictions, patient ranking in the hospitals, and treatment efficacy assessments in infection stability, control, and recovery.[12] Introducing forecasting systems that are capable to provide an accurate prediction and diagnosis of such a viral disease remains a challenge. The development of AI-based techniques for identifying epidemiological hazards will be critical in achieving improved preventions, predictions, and detections of future global hazards.[13]

Medical data mining has its own characteristics that distinguish it from data mining in other issues and applications. Researchers who practice data mining in other fields may encounter limitations and problems such as exploring heterogeneous, confidential or private data, bulk and distributed, and less ethical, legal, and security aspects.[14] In the present study, several data mining models are presented to predict the recovery of patients with COVID-19. Perhaps these models can predict when patients infected with COVID-19 will recover and be released from treatment centers or quarantine, as well as patients who may not recover and die from the epidemic of COVID-19.

Meanwhile, Alsunaidi et al. performed a literature review to underline the contributions of several COVID-19 studies by big data analysis. In this study, a taxonomy of several applications used for controlling and managing the epidemic is provided.

Besides, this paper confers about the various challenges while analyzing the data on COVID-19. This study provided valuable directions for further research in the future.[15]

Yahya et al.[16] investigated the spread of COVID-19 in next 6 months in Iraq. The results of this study showed an increased spread of 17.1% in the 6-month period, resulting in 8.3% increase in the average death cases. In this study, spatial distribution maps were generated to simulate virus spread using a Geographic Information System (GIS) environment. It was concluded that this is a useful technique to develop plans for confronting COVID-19 in Iraq.

Elroukhly et al. performed experiments to extend their method for predicting the number of COVID-19 cases in the future during a period between August 2020 and September 2020.[17] This method led to a more accurate prediction than those obtained by other methods. They argued that such predictions might be utilized by the government or people in the affected countries to take precautionary measures against the pandemic.

Although the ability of data analysis systems in predicting the number of COVID-19 cases have been studied in the research, the existing problems regarding the use of this technique remained unchanged and finding practical solutions for these problems is a necessity. In response, the present study aimed to present new solutions to these problems.

These models help healthcare providers in determining stability and recovery of newly affected people with COVID-19 pandemic. These models have been developed using data sets from Iranian hospitals and data samples from the history of death and epidemic recovery of COVID-19 infection have been considered. In order to develop models, data mining algorithms including naive Bayesian, support vector machine, decision tree, nearest neighbor K and logistic regression random forest are directly applied on the data set using Python programming language.[18,19]

Today, the volume of data stored electronically in the medical field is increasing day by day. To make sense of this data, it must be analyzed and turned into knowledge. Given such a volume of patterns and the use of humans as pattern identifiers and data analysts, it is not possible to respond to such a large amount of information, and therefore data mining is of great importance in the medical field. Data mining in medicine is used in the prevention or diagnosis of diseases and the choice of treatment methods. Thus, the purpose of the present study was to reduce the prevalence of coronary heart disease by predicting the incidence of the disease using data mining techniques.

**Method and Materials**

In this cross-sectional study conducted in 1399, the main objective was to perform a comparison between different data mining algorithms and combine these algorithms, a new, efficient and high-precision method and the ability to implement on local data. These three techniques included the J48 independent decision tree, the Adaboost and the collective Bagging techniques using J48 as basic learner.

Each technique analyzed the data based on a specific algorithm. After collecting and arranging the data, the data were interpreted and explained and all three variables were summarized together.
In the present study, two data mining techniques were compared and then a new model for predicting coronary heart disease was presented.

The techniques used in this study are:

a. Decision Tree J48: is an open-source Java classifier that implements the C4.5 algorithm in WEKA. It is an extension of ID3 algorithm suggested by Quinlan. This is part of the information acquisition and data sharing Quinlan. 4 provided comprehensive details on the J48 decision tree.

\[
\text{Gain_Ratio}(D, A) = \frac{\text{Entropy}(D) \sum_{j=1}^{t} p_j \times \text{Entropy}(p_j)}{\text{Spliting_Info}}
\]

b. Bagging: is a simple but strong method of designing independent group for improving the accuracy of learning algorithms that are unstable, such as decision tree and rule learning algorithms. It is derived from bootstrap collection. The bagging data set is distributed in different bootstrap. Each version is designed independently of the original replacement data set. Each iteration accounts for 63.2% of the original data in average. This process is done by frequently performing a weak learner in different bootstrap. The resulted classification is combined in each repetition for the strong composite classification to achieve high accuracy from each separate classification done separately.

In the present study, a researcher-made questionnaire was used to collect data, which included gender, patient age, weight, length of hospital stay, physician specialty, readmission, travel history, underlying illness, and length of stay in the hospital. It was outside the home that was completed from the patients file.

Statistical Analysis conducted using IBM SPSS 23.0 (IL Chicago, USA). Significance was held on 0.05.

## Results

The data of 600 patients with COVID-19 infection was collected and analyzed as shown in [Table 1-6]. If patients were over 60, the probability of death prognosis increases by 37.5%.

According to Table 3, 18.4 of expired patients and 19.8% of asymptomatic patients were among men. The risk assessment success was 72% as reported in [Table 3].

The results of risk assessment for foreign trip history revealed that most prevalent cases of death prognosis was among patients without foreign trip history. As shown in [Table 5 and 6], the risk assessment success was. The results on BMI showed that 83% of slim patients infected with COVID-19 were asymptomatic and also 26.5% of patients with death prognosis were among patients with high BMI.

About daily outward career, the results of analysis showed that most of patients with death prognosis were working more than 8 h daily.

The results on systemic disease showed that in patients with any previous systemic diseases, the prevalence of death prognosis was 0% and critical prognosis was 16%. But in patients with systemic disease the rate of death prognosis was 22% and critical prognosis was 34% and shown in [Table 6].

Evaluation: The performance of the decision tree and bagging classification was evaluated on COVID-19. [Table 7] presents the data set classification results. The best performance is related to bagging and then the decision tree. The efficacy of these algorithms in classifying experimental data set is shown in [Table 4 and 5]. According to [Table 7], bagging uses the best performance among the classification algorithm when selecting a classification based on objectives.

## Discussion

In the present study, COVID-19 data set was prepared from the records of 600 patients. The dataset was used to aid scholars with last advances in vaccine development to gain new insights that could support competition for a COVID-19 pandemic control solution.

The application of decision tree for classification has led to successful results in different areas. The structure of decision tree resembles the hierarchical analysis of data space.20 Through performing inductive learning, decision tree acquires a tree-like structure from data that provides a set of equivalent decision regulations.21 This algorithm aims to provide a tree-like structure of the properties of the classified data. Although the process of learning and predicting is very fast in decision tree, this technique is susceptible to overuse and small disturbances in the data.
However, such shortcoming can be prevented through pruning and validation. In the present study, the decision tree algorithm was combined with various attributes. In the present study, the decision tree had a sensitivity of 97% and an accuracy of 95% among other classification algorithms. The classification of documents and textual datasets was performed using bagging in different bootstrap instances, each of which creating a uniform instance of the collection. This technique has been employed to enhance the performance of standard algorithms in ML. The classifications obtained from bagging were incorporated into a composite classifier dedicated for predicting a weighted composition of individual classifiers. It is different, including computational complexity and loss of interpretation. The parameters of random_state (= 1) and n_estimators (= 100) are used in this study. In the present study, bagging had the best performance among the classification algorithms.
### Table 7: Comparison between Decision Tree and Bagging System

| Models    | Decision Tree | Bagging |
|-----------|---------------|---------|
| Accuracy  | 97.75         | 98.07   |
| Conductivity | 95.70         | 96      |
| Recall    | 95.88         | 96.1    |

In the present study, the accuracy of the decision tree was 95%, which is a good accuracy. These results are consistent with differences in the type of disease, despite differences such as the research population, and the reason for this is that this algorithm manages the missing values in the data compared to other decision tree algorithms. It also performs pruning steps after the tree is made. Due to this feature, it has been used extensively in disease prediction. Also, the improved model of J48 algorithm has been able to show higher sensitivity in predicting disease traits.

A very important feature of the J48 algorithm is its flexibility and scalability, meaning that it is suitable for both large-scale data processing and small-scale data processing. A major problem with this algorithm is the complexity of its execution time, and it also has the potential for classification errors in predicting disease for which no solution has been proposed so far. In the present study, the probability of error was estimated to be about 39%. Fan et al. (2015) reviewed clinical data mining and reviewed data mining techniques in diabetes. The results showed that data mining can be an effective method in preventing, treating and discovering the relationship between disease and its risk factors, leading to significant advances in research and providing better health care for patients.[26]

### Conclusion

In the present study, by examining the records of patients with COVID-19 and using two methods of data tree extraction, decision making and bagging to classify and determine which of these algorithms works best, it was found that the use of bagging in diagnosis and prediction of coronary heart disease based on the existing traits in the diagnosis and treatment of this disease is more effective than the decision tree method, although the accuracy and sensitivity of these two methods differ by only 1%, in the end we can say that both methods J48 tree and bagging are effective in diagnosing and predicting coronary heart disease based on existing traits in diagnosing and treating this disease.

### Financial support and sponsorship

Nil.

### Conflicts of interest

There are no conflicts of interest.

### References

1. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. China Novel Coronavirus Investigating and Research Team. A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med 2020;382:727–33. Available from: https://doi.org/10.1056/NEJMoa2001017.
2. Cucinotta D, Vanelli M. WHO declares COVID-19 a pandemic. Acta Biomed 2020;91:157‑60.
3. Li J, Huang DQ, Zou B, Yang H, Hui WZ, Rui F, et al. Epidemiology of COVID-19: A systematic review and meta-analysis of clinical characteristics, risk factors, and outcomes. J Med Virol 2021;93:1449‑58.
4. Doosti-Irani A, Haghdoot AA, Najafi F, Eybpoosh S, Moradi G, Amiri FB, et al. How can the epidemic curve of COVID-19 in Iran be interpreted? J Res Health Sci 2020;20:e00491.
5. Moradi Y, Eshrat B. Estimation of the net reproductive number of COVID-19 in Iran. Med J Islamic Repub Iran 2020;34:34.
Shahori and Mirzaei: The efficiency of data analysis systems

6. Abdi M. Coronavirus disease 2019 (COVID-19) outbreak in Iran: Actions and problems. Infect Control Hosp Epidemiol 2020;41:754-5.

7. Randolph HE, Barreiro LB. Herd immunity: Understanding COVID-19. Immunity 2020;52:737-41.

8. Saba N, Akram W, Ahmed T. Predicting COVID-19 incidence using data mining techniques: A case study of Pakistan. Brain 2021;11:168-84.

9. Muhammad I, Islam MM, Usman SS, Ayon SI. Predictive data mining models for novel coronavirus (COVID-19) infected patients’ recovery. SN Comput Sci 2020;1:1-7.

10. Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng 2018;2:719-31.

11. Elsi ZRS, Pratiwi H, Efendi Y, Rusdina R, Alfah R, Windarto AP, et al., editors. Utilization of data mining techniques in national food security during the Covid-19 pandemic in Indonesia. Journal of Physics: Conference Series. J. Phys 2020;1594:012007.

12. Reigal RE, Pastrana-Brincones JL, González-Ruiz SL, Hernández-Mendo A, Morillo-Baro JP, Morales-Sánchez V. Use of data mining to determine usage patterns of an online evaluation platform during the COVID-19 pandemic. Front Psychol 2020;11:1-16.

13. Long JB, Ehrenfeld JM. The role of artificial intelligence (AI) in detecting and preventing the spread of novel coronavirus. J Med Syst 2020;44:44-59.

14. 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.

15. Alsunaidi SJ, Almuhaideb AM, Ibrahim NM, Shaikh FS, Alqudaifi KS, Alhaidari FA, et al. Applications of big data analytics to control COVID-19 pandemic. Sensors (Basel) 2021;21:2282.

16. Yahya BM, Yahya FS, Thannoun RG. COVID-19 prediction analysis using artificial intelligence procedures and GIS spatial analysis: a case study for Iraq. Applied Geomatics 2021;13:481-91. doi:10.1007/s12518-021-00365-4.

17. Eltoukhy AEE, Shaban FTS, Abdel-Aal MAM. Data analytics for predicting COVID-19 cases in top affected countries. Int J Environ Res Public Health 2020;17:7080.

18. Lee SJ, Siau K. A review of data mining techniques. Ind Manag Data Syst 2001;101:41-6.

19. Li D, Chaudhary H, Zhang Z. Modeling spatiotemporal pattern of depressive symptoms caused by COVID-19 using social media data mining. Int J Environ Res Public Health 2020;17:4988.

20. Le TT, Andreadakis Z, Kumar A, Román RG, Tollefsen S, Saville M, et al. The COVID-19 vaccine development landscape. Nat Rev Drug Discov 2020;19:305-6.

21. Rafea L, Ahmed A, Abdullah WD. Classification of a COVID-19 dataset by using labels created from clustering algorithms. Indones J Electr Eng Comput Sci 2021;21:164-73.

22. Amador Penichet L, Magdaleno Guevara D, Garcia Lorenzo MM. New similarity function for scientific articles clustering based on the bibliographic references. Computacion y sistemas 2018;22:93-102.

23. Han T, Liu C, Yang W, Jiang D. A novel adversarial learning framework in deep convolutional neural network for intelligent diagnosis of mechanical faults. Knowl Based Syst 2019;165:474-87.

24. Wu X, Zhu X, Wu G, Q, Ding W. Data mining with big data, IEEE transactions on knowledge and data. eng 2014;26:97-107.

25. Magerman DM. Statistical decision-tree models for parsing. arXiv preprint cmp-lg/9504030. 1995.

26. Fan H, Yue F, Li H. Project-based as-needed information retrieval from unstructured AEC documents. J Manag Eng 2015;31:A4014012.