Prediction of COVID-19 disease severity using machine learning techniques

Alaa H. Ahmed[1], Mokhaled N. A. Al-Hamadani[2], Ihab A. Satam[2]
[1]Department of Network, College of Computer Science and Information Technology, University of Kirkuk, Kirkuk, Iraq
[2]Department of Electronic Techniques, Technical Institute/Alhawija, Northern Technical University, Adan, Kirkuk, Iraq

Article Info

Article history:

Received Oct 19, 2021
Revised Mar 4, 2022
Accepted Mar 13, 2022

Keywords:

Chronic diseases
Classification
COVID-19
Machine learning
Severity

ABSTRACT

A terrifying spread of COVID-19 (which is also known as severe acute respiratory syndrome coronavirus 2 or SARS-COV-2) led scientists to conduct tremendous efforts to reduce the pandemic effects. COVID-19 has been announced pandemic discovered in 2019 and affected millions of people. Infected people may experience headache, body pain, and sometimes difficulty in breathing. For older people, the symptoms can get worse. Also, it can cause death because of the huge effect on some parts of the human body, particularly for those who have chronic diseases like diabetes. Machine learning algorithms are applied to patients diagnosed with Corona Virus to estimate the severity of the disease depending on their chronic diseases at an early stage. Chronic diseases could raise the severity of COVID-19 and that is what has been proved in this paper. This paper applies different machine learning techniques such as random forest, decision tree, linear regression, binary search, and k-nearest neighbor on Mexican patients' dataset to find out the impact of lifelong illnesses on increasing the symptoms of the virus in the human body. Besides, the paper demonstrates that in some cases, especially for older people, the virus can cause inevitable death.

Corresponding Author:

Mokhaled N. A. Al-Hamadani
Department of Electronic Techniques, Northern Technical University
Adan, Kirkuk 36001, Iraq
Email: Mokhaled_hwj@ntu.edu.iq

1. INTRODUCTION

At the end of 2019 and at the beginning of 2020, suddenly the whole world was isolated and the people started to be afraid of everything and everyone around them. The reason was because of coronavirus (COVID-19) terrifying spread which derived to be considered as a new pandemic [1], [2]. The virus affected almost all the countries around the world since it is a contagious virus. The term COVID-19 is titled by the World Health Organization (WHO), depending on the first affirmative test which was at the end of 2019 [2], [3]. According to WHO, the incubation time of the virus ranges from 2 to 14 days in the human body [3], [4]. However, the problem with corona virus is that it shows different symptoms from one person to another. For example, for some people, it shows minor symptoms such as headache and loss of sense of smell and taste only. However, others may suffer from very severe symptoms such as shortness of breath, persistent cough, and physical exhaustion. Those symptoms can get worse if the person has chronic diseases such as diabetes and in some cases can lead to death. Therefore, in this paper we showed that chronic diseases increase COVID-19 symptoms dramatically by using machine learning (ML) techniques.

ML is one of the subdomains of artificial intelligence which concerns in training machines to work as a human being [5]-[7]. A lot of scientists and researchers tried to enhance ML algorithms both in terms of classification and regression. For those reasons, ML algorithms have been used widely in many fields and applications especially in healthcare fields [2], [3], [8], [9]. Thus, we used several ML algorithms in this paper like random forest (RF), linear regression (LR), decision tree (DT), k-nearest neighbor (KNN), and support vector machine (SVM) algorithms. We applied those algorithms on the extracted patents' dataset which is taken from the Mexican government. We used python programming language to develop those algorithms in order to prove that chronic diseases can raise corona virus symptoms especially for elder people.

This paper is organized as section 2 provides information about the methodology that we used and the programming language that we utilized to obtain the required result. Also, there is a detailed explanation of dataset that is taken from Mexican patients and the extracted dataset that we depend on it in our experiments. In section 3, we explained ML techniques that we applied to the extracted dataset. Moreover, it also contains a demonstration of the result and experiences that we conducted. Finally, the conclusion and discussion is taking place in section 4.

2. RESEARCH METHOD

2.1. Program

We used python programming language to apply ML algorithm on COVID-19 patients' dataset. Python language supports object-oriented, functional, and multiple paradigm programming [10], [11]. Therefore, it can be used to develop complex and massive software programs. It is one of the best well-known languages that implement in several fields like machine learning, data mining, and the internet of things [9]. For those reasons, the popularity of python is growing so rapidly nowadays [10].

2.2. Dataset

The dataset that we used in our paper is taking from kaggle website which is provided by the Mexican government [12]. The dataset contains enormous information of COVID-19 patients. It comprises 23 features and 563201 instance with unique patient ID's. In each row, there is specific patient information which are a unique id, sex (1-female, 2-male), patient type (outpatient-1, inpatient-2), date of entry to hospital, date of symptoms, date died (no-1, yes-2 died), intubed (yes-1, no-2), pneumonia (yes-1, no-2), age , pregnancy (yes-1, no-2), diabetes (yes-1, no-2), copd (yes-1, no-2, data missing-97,98,99, asthma (yes-1, no-2), inmsupr (yes-1, no-2), hypertension (yes-1, no-2), other disease (yes-1, no-2), cardiovascular (yes-1, no-2), obesity (yes-1, no-2), renal chronic (yes-1, no-2), tobacco (yes-1, no-2), contact other COVID-19 patients (yes-1,no-2), COVID-19 result (positive-1, negative-2), awaiting results-3), and icu (yes-1, no-2).

2.3. Dataset extraction

In our work, we extracted a sample of 10,000 data instances along with 6 features which are age, sex, date died (to know if the person has died or not), diabetes, obesity, and COVID-19 result (to verify if it is positive or negative). Figures 1 and 2 show the frequency of age and gender effects on the patient. Figures 3 and 4 show the frequency of diabetes and obesity effects and how it can lead to increase death cases. Table 1 shows an uncomplicated sample of the dataset that we utilized in this paper.

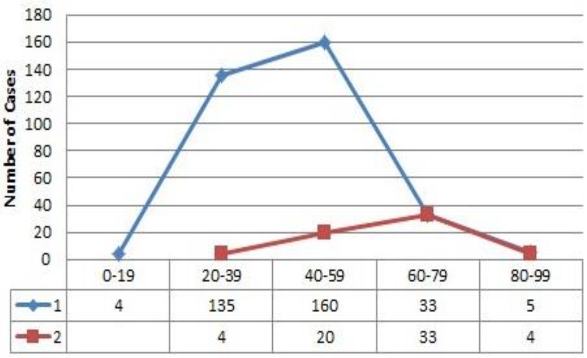

Figure 1. Age frequency where blue is for infected cases and red is for death cases

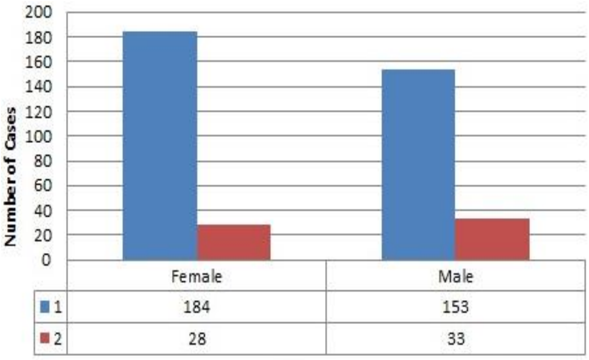

Figure 2. Gender frequency where blue is for infected cases and red is for death cases

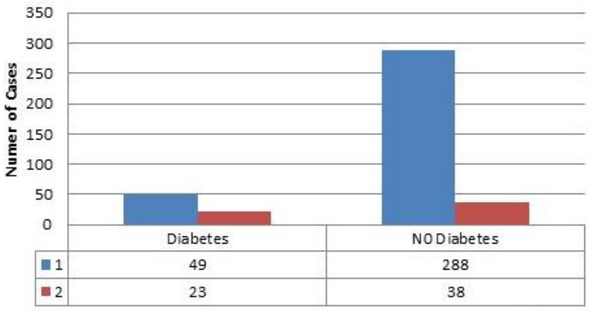

Figure 3. Diabetes frequency where blue is for infected cases and red is for death cases

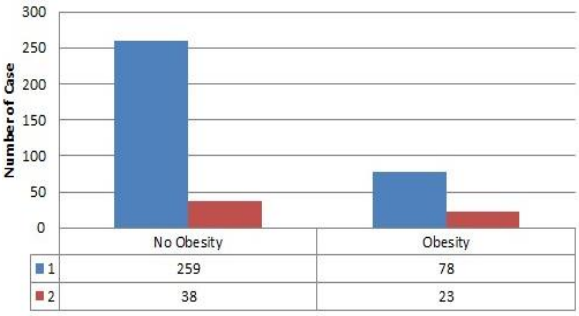

Figure 4. Obesity frequency where blue is for infected cases and red is for death cases

Table 1. A sample from the extracted dataset

No.	Age	Gender	Covid_result	Death	Obesity	Diabetes
1	27	Male	Positive	1	No obesity	No diabetes
2	24	Male	Positive	1	No obesity	No diabetes
3	54	Female	Positive	1	Obesity	No diabetes
4	30	Male	Positive	1	No obesity	No diabetes
5	60	Female	Positive	2	No obesity	Diabetes
6	47	Male	Positive	2	No obesity	Diabetes
7	63	Male	Positive	1	No obesity	No diabetes

3. RESULTS AND DISCUSSION

3.1. Machine learning algorithms

ML has four techniques depending on the required data nature. These four categories are supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning [6], [13], [14]. We used supervised learning algorithms since we have labeled dataset as shown in Figure 5. This flowchart shows the proposed system (ML models) that have been used over the Mexican dataset.

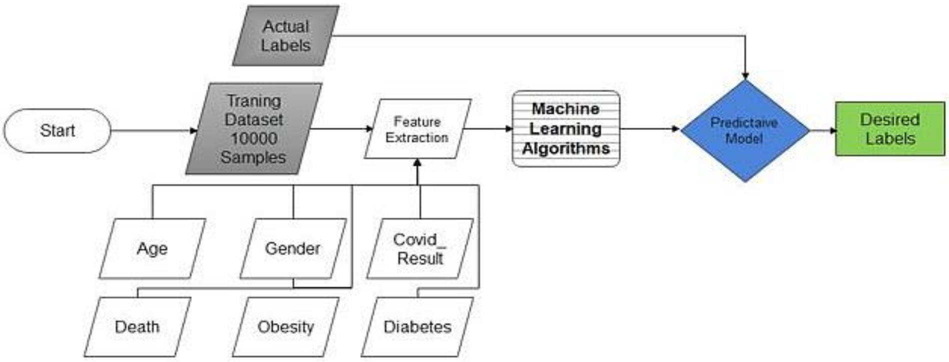

Figure 5. Flowchart of supervised ML algorithms

3.1.1. Random forest

RF is one of the most easily applied supervised ML algorithms [15], [16]. It uses for both regression and classification tasks. It establishes the classification by generating many decision trees. The algorithm chooses prediction from each one of those trees. The more decision trees, the better algorithm works. After that, RF determines the best result using the voting technique [17]. Using RF algorithm obtains a very high accurate result even though it consumes time in choosing prediction and tracing each decision tree that it comprises.

3.1.2. Decision tree

Decision tree is also known as a decision making tree that uses in many fields such as data mining, statistics, and machine learning. It has two main types which are classification tree and regression tree. DT steps are straightforward by classifying instances after sorting them depending on feature values [8], [18].

The classification starts from the root node where each node exemplifies a feature; the branch exemplifies the rules of the tree and the leaf node exemplifies the result. The root node (which is the first or top node in the tree) is trained to partition the tree recursively depending on the instance values. Each internal node considers as a part of the tree partition process; because each one of them makes a decision and has many branches. However, the case is different for the leaf node, because it stores the final outcome and does not make a decision or have branches.

3.1.3. Support vector machine

It is one of the most powerful self-learning algorithms; which is developed from statistic-learning to be used for big data regression and classification [19]-[21]. SVM method avoids the overfitting problems that happen in most of ML techniques [22]. It can perform both linear and nonlinear classification. Also, it performs more efficiently when there is enough separation between the classes. SVM algorithm establishes by representing each data as a point with space n from each other (where n is the number of features and each one of them has a coordinator value). After that, the classification performs by constructing one or multi hyper-plane which separates the data groups. Choosing the best hyper-plane for SVM depends on the distance between it and any point of the data. In another word, classifying the largest number of points correctly depends on the large distance between the hyper-plane and data points.

3.1.4. K-nearest neighbor

The logic behind KNN is intuitive to be comprehended and easy to be implemented [23]. It can be used for both regression and classification but mostly for classification jobs. KNN is called so because it classifies a given data relying on its neighbors' classification. The basic key behind KNN algorithm is that it depends on the similarity between the features [24]. That means the algorithm assumes similarity in order to find the category of the new data. The algorithms start by choosing K value which can be any integer, then it measures the distance of k nearest neighbor. After that, it assigns the new data to the category which has the maximum number of data points. Thus, K value and distance measurement (which can be done by using Euclidean distance and Hamming distance) consider the core of the KNN algorithm's work.

3.1.5. Linear regression

LR is a supervised ML algorithm which is also well known as a statistical algorithm. LR has two types which are simple LR and multiple LR [25]. It predicts the outcome by observing each feature constantly instead of making categories to classify [17]. It is called LR because it finds a relationship between dependent variable (y) and independent variable (x). That means it discovers the changes in dependent variable values according to the independent variable values. The relation between dependent and independent variables can be positive; if y-axis shows dependent variable increase and x-axis shows independent variable increase; otherwise, it will be negative. LR simply can be mathematically represented as:

$$Y = b1 + b2X + \varepsilon$$

Where y is dependent variable, x is independent variable, b1 is line intercept, is b2 is LR coefficient, and ε is a random error.

3.2. Experiments and results

We applied ML algorithms on the extracted dataset of COVID-19 patients to find the severity of the symptoms. We used python programming language to implement ML algorithms which are RF, decision tree, support vector machine, LR, and k nearest neighbors to classify the dataset. We utilized precision, recall, f1 score, and accuracy to show the effectiveness and trustworthiness of each technique as it's shown in Table 2. Where precision is the fraction of true positive to the summation of true positive and false positive. Recall is the fraction of true positive to the summation of true positive and false negative; whereas, f1 score is the mean of recall and precision. We obtained sufficient outcomes from those algorithms in classifying patients' cases depending on their chronic diseases. Table 2 shows the details of each algorithm showing their outcome of precision, recall, F1 score, and accuracy.

Table 2. Precision, recall, and F1 score of ML algorithms

No.	ML techniques	Accuracy	Precision	Recall	F1-score
1	RF	0.88	0.88	0.99	0.93
2	DT	0.88	0.88	0.99	0.93
3	SVM	0.87	0.87	1	0.93
4	KNN	0.86	0.89	0.96	0.92
5	LR	0.88	0.88	1	0.93

4. CONCLUSION

The entire globe has been affected by the corona virus. It causes more than 4 million death cases as of August 31, 2021. A lot of scientists and researchers tried to find the best solution to fight this virus. ML techniques were one of the methods to combat the virus in the early stage. In this paper, we used ML algorithms to predict the severity of virus symptoms, especially for people who have chronic disease. We applied RF, KNN, SVM, LR, and DT algorithms over COVID-19 patients dataset which was taken from the Mexican government. We obtained sufficient results classifying patients' cases depending on their lifelong diseases. Our proposed system using five ML algorithms (RF, DT, SVM, KNN, and LR) classified the dataset with an accuracy of (0.88%, 0.88%, 0.87, 0.86, and 0.88%) respectively. For future work, more enhancements can be done to get more accurate outcomes.

REFERENCES

[1] F. Ahouz and A. Golabpour, "Predicting the incidence of COVID-19 using data mining," *BMC Public Health*, vol. 21, no. 1087, pp. 1-12, 2021, doi: 10.1186/s12889-021-11058-3.

[2] S. M. Ayyoubzadeh, S. M. Ayyoubzadeh, H. Zahedi, M. Ahmadi and S. R. N. Kalhori, "Predicting COVID-19 Incidence Through Analysis of Google Trends Data in Iran: Data Mining and Deep Learning Pilot Study," *JMIR public health and surveillance*, vol. 6, no. 2, pp. 1-13, 2020, doi: 10.2196/18828.

[3] D. Painuli, D. Mishra, S. Bhardwaj and M. Aggarwal, "Forecast and prediction of COVID-19 using machine learning," *Data Science for COVID-19, Academic Press*, pp. 381-397, 2021, doi: 10.1016/B978-0-12-824536-1.00027-7.

[4] E. Hersh and J. Seladi-Schulman, "After Exposure to the Coronavirus, How Long Before Symptoms Appear?," *Healthline*, 2020. [Online]. Available: https://www.healthline.com/health/coronavirus-incubation-period.

[5] P. P. Shinde and S. Shah, "A Review of Machine Learning and Deep Learning Applications," *Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, 2018, pp. 1-6, doi: 10.1109/ICCUBEA.2018.8697857.

[6] M. Mohammed, M. B. Khan, and E. B. M. Bashie, "Machine learning: Algorithms and applications," in Machine Learning: Algorithms and Applications, 2016, no. July, pp. 1–204, doi: 10.1201/9781315371658.

[7] K. Kersting, "Machine Learning and Artificial Intelligence: Two Fellow Travelers on the Quest for Intelligent Behavior in Machines," *Frontiers Big Data*, vol. 1, no. 6, pp. 1-4, 2018, doi: https://doi.org/10.3389/fdata.2018.00006.

[8] F. Y. Osisanwo, J. E. T. Akinsola, O. Awodelea, J. O. Hinmikaiye, O. Olakanmi and J. Akinjobi, "Supervised Machine Learning Algorithms: Classification and Comparison," *International Journal of Computer Trends and Technology (IJCTT)*, vol. 48, no. 3, pp. 128-138, 2017.

[9] L. J. Muhammad, M. M. Islam, S. S. Usman and S. I. Ayon, "Predictive Data Mining Models for Novel Coronavirus (COVID-19) Infected Patients' Recovery," *SN Computer Science*, vol. 1, no. 4, pp. 1-7, 2020, doi: 10.1007/s42979-020-00216-w.

[10] K. R. Srinath, "Python – The Fastest Growing Programming Language," *International Research Journal of Engineering and Technology (IRJET)*, vol. 4, no. 12, pp. 354-357, 2017.

[11] S. Raschka, J. Patterson and C. Nolet, "Machine Learning in Python: Main Developments and Technology Trends in Data Science, Machine Learning, and Artificial Intelligence," *Information*, vol. 11, no. 4, pp. 1-44, 2020, doi: https://doi.org/10.3390/info11040193.

[12] T. Mukherjee, "COVID-19 patient pre-condition dataset," *Kaggle*, 2020. [Online]. Available: https://www.kaggle.com/tanmoyx/covid19-patient-precondition-dataset.

[13] G. Edwards, "Machine Learning | An Introduction," *Towards Data Science*, 2018. [Online]. Available: https://towardsdatascience.com/machine-learning-an-introduction-23b84d51e6d0.

[14] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Computer Science*, vol. 2, no. 160, pp. 1-21, 2021.

[15] A. Navlani, "Understanding Random Forests Classifiers in Python," *Datacamp*, 2018. [Online]. Available: https://www.datacamp.com/community/tutorials/random-forests-classifier-python.

[16] L. Rafea, A. Ahmed and W. D. Abdullah, "Classification of a COVID-19 dataset by using labels created from clustering algorithms," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 21, no. 1, pp. 164-173, 2021, doi: 10.11591/ijeecs.v21.i1.

[17] J. Ulaganathan and K. M. Sadyojatha, "A Review On Maintenance Techniques For Industrial Equipment And Its Machine Learning Algorithms," *International Journal of Advanced Research in Engineering and Technology (IJARET)*, vol. 12, no. 4, pp. 183-194, 2021, doi: 10.34218/IJARET.12.4.2021.022.

[18] M. M. Saad, N. Jamil and R. Hamzah, "Evaluation of Support Vector Machine and Decision Tree for Emotion Recognition of Malay Folklores," *Bulletin of Electrical Engineering and Informatics*, vol. 7, no. 3, pp. 479-486, 2018, doi: https://doi.org/10.11591/eei.v7i3.1279.

[19] D. K. Srivastava and L. Bhambhu, "Data classification using support vector machine," *Journal of Theoretical and Applied Information Technology*, vol. 12, no. 1, pp. 1-7, 2010.

[20] S. Suthaharan, "Machine Learning Models and Algorithms for Big Data Classification," New York: Springer US, vol. 36, pp. 1-12, 2016, doi: 10.1007/978-1-4899-7641-3.

[21] E. A. Mahareek, A. S. Desuky and H. A. El-Zhni, "Simulated annealing for SVM parameters optimization in student's performance prediction," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 3, pp. 1211-1219, 2021, doi: 10.11591/eei.v10i3.2855.

[22] M. Awad and R. Khanna, "Support Vector Machines for Classification," *Efficient Learning Machines*, Berkeley, CA, Apress, pp. 39-66, 2015, doi: 10.1007/978-1-4302-5990-9_3.

[23] S. Kang, “k-Nearest Neighbor Learning with Graph Neural Networks,” *Mathematics*, vol. 9, no. 8, pp. 1-12, 2021, doi: 10.3390/math9080830.

[24] Y.-l. Cai, D. Ji and D.-f. Cai, “A KNN Research Paper Classification Method Based on Shared Nearest Neighbor,” *Proceedings of NTCIR-8 Workshop Meeting NTCIR-8*, Tokyo, Japan, 2010.

[25] D. H. Maulud and A. M. Abdulazeez, “A Review on Linear Regression Comprehensive in Machine Learning,” *Journal of Applied Science and Technology Trends*, vol. 1, no. 4, pp. 140-147, 2020, doi: 10.38094/jastt1457.

BIOGRAPHIES OF AUTHORS

Alaa H. Ahmed B.Sc. Computer Science-College of Science University of Kirkuk-Iraq. M.Sc Computer Science College of Arts and Sciences The University of North Carolina at Greensboro-USA. Assistant lecturer University of Kirkuk, Iraq. Research interest: data mining, data fusion, database, machine learning, networking, and any new techniques and subjects in computer science. She can be contacted at email: alaa.ahmed@uokirkuk.edu.iq.

Mokhaled N. A. Al-Hamadani B.Sc. Computer Science-College of Science-University of Kirkuk-Iraq. M.Sc Computer Science College of Arts and Sciences The University of North Carolina at Greensboro, USA. Lecturer at Northern Technical University, Iraq. Research interest: big data, deep learning, machine learning, database, networking, and any new techniques and subjects in computer science. He can be contacted at email: mokhaled_hwj@ntu.edu.iq.

Ihab Abdulrahman Satam B.Sc. Mechatronics-Engineering College University of Mosul, Iraq. M.Sc Mechatronics Alkhwarizmi Engineering College University of Baghdad, Iraq. Lecturer-Northern Technical University, Iraq. Research interest: robotics-neural network-control-autopilot. He can be contacted at email: ihab_hwj@ntu.edu.iq.