ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Prediction of CoVid-19 mortality in Iraq-Kurdistan by using Machine learning

Brzu T. Muhammed[1], Ardalan H. Awlla[2], Sherko H. Murad[3], Sabah N. Ahmad[4]

[1]Department of Computer Science, Kurdistan Technical Institute, Sulaymaniyah, Iraq, [2]Department Information Technology, University of Human Development, Sulaymaniyah 0778-6, Iraq, [3]Department of Computer Science, Kurdistan Technical Institute, Sulaymaniyah, Iraq, [4]General Manager of Health in Sulaymaniyah, Iraq

ABSTRACT

This research analyzed different aspects of coronavirus disease (COVID-19) for patients who have coronavirus, for find out which aspects have an effect to patient death. First, a literature has been made with the previous research that has been done on the analysis dataset of coronavirus using Machine learning (ML) algorithm. Second, data analytics is applied on a dataset of Sulaymaniyah, Iraq, to find factors that affect the mortality rate of coronavirus patients. Third, classification algorithms are used on a dataset of 1365 samples provided by hospitals in Sulaymaniyah, Iraq to diagnose COVID-19. Using ML algorithm provided us to find mortality rate of this disease, and detect which factor has major effect to patient death. It is shown here that support vector machine (SVM), decision tree (DT), and naive Bayes algorithms can classify COVID-19 patients, and DT is best one among them at an accuracy (96.7 %).

Index Terms: Coronavirus disease, Coronavirus, Forecasting, Machine learning, Kurdistan-IRAQ

1. INTRODUCTION

The coronavirus disease (COVID-19) is the family of viruses including SARS, ARDS. W.H.O declared this outbreak as a public health emergency [1] and mentioned the following; the virus is being transmitted through the respiratory tract when a healthy person comes in contact with the infected person.

In December 2019, Wuhan, Hubei region, China, has been accounted for as the focal point of the COVID-19 episode [2]. A quarter of a year later, that outbreak was pronounced as a worldwide pandemic by the World Health Organization (WHO) [3]. More than 54.40 million confirmed COVID-19 cases and more than 1.32 deaths worldwide have been officially reported in 16 November, 2020. Therefore, it has been considered as the most critical universal crisis since the World War-II [4]. The coronavirus has spread in Kurdistan – Iraq like all the country in the world, and it has expanded fast in Sulaymaniyah city. The mortality of this disease expands day by day and this infection becomes as a major danger to the mankind of whole world. Alongside the clinical explores, the examination of related information will support the humanity. Recent studies identified that machine learning (ML) and artificial intelligence (AI) are promising technology employed by various health-care providers as they result in better scale-up, speed-up processing power, reliable, and even outperform human in specific health-care tasks [5]. In this paper, we established three ML algorithm for the prediction of coronaviruses' diseased patients' mortality. The models forecast when COVID-19 infected patients would be death or recovered. The proposed algorithms are designed with the dataset found from Sulaymaniyah city for coronavirus and dataset cases of the death and recovery records of the infected coronavirus's pandemic. ML algorithm which includes decision tree (DT), support vector

Access this article online

DOI:10.21928/uhdjst.v5n1y2021.pp66-70	**E-ISSN:** 2521-4217 **P-ISSN:** 2521-4209

Corresponding author's e-mail: Brzu T. Muhammed, Department of Computer Science, Kurdistan Technical Institute, Sulaymaniyah, Iraq. E-mail: brzu.tahir@kti.edu.krd

Received: 22-11-2020 Accepted: 19-05-2021 Published: 23-05-2021

machine (SVM), and naive Bayes (NB) was implemented directly on the dataset using Weka Tool which is a data mining tool.

2. LITERATURE REVIEW

Development of AI changed the world in all fields. ML a subset of AI causes the human to discover answers for exceptionally complex issues and furthermore assumes an imperative part in making human life refined. The application zones of ML incorporate business applications, clever robots, medical services, atmosphere demonstrating, picture handling, natural language preparing, and gaming [6]. According to Al Sadig *et al.* [7], depend on the dataset as given by the various site developed by digital science in cooperation with over 100 leading research organizations all over the world. Create a model using J48 algorithm to predict the most common symptoms causing death is acute kidney injury and coronary heart disease.

Arun and Iyer [8] propose some of the ML techniques such as rough set (SVM), Bayesian Ridge and Polynomial Regression, SIR model, and RNN to examination of the transmission of COVID 19 disease and predict the scale of the pandemic, the recovery rate as well as the fatality rate.

According to Bullock *et al.* [9], ML and deep learning can replace humans by giving an accurate diagnosis. The perfect diagnosis can save radiologists' time and can be cost-effective than standard tests for COVID-19. X-rays and computed tomography scans can be used for training the ML model.

Wang and Wong [10] created COVID-Net, which is a profound convolutional neural network, which can analyze COVID-19 from chest radiography pictures.

Alibaba Cloud 2020 [11] exploit ML to set up an adjusted Susceptible - Exposed - Infectious - Recovered model to anticipate the commonness of COVID-19 and evaluate the expanded danger of defilement in a particular territory.

Kemenkes [12] finding diabetes utilizing AI and ML methods result demonstrated that ensemble technique guaranteed exactness of 98.60%. These reasons can be advantageous to analyze and foresee COVID-19. According to Muhammad *et al.* [13] use several ML algorithms which includes DT, SVM, NB, LR, RF, and K-NN are applied directly on the dataset which include COVID-19 infected patients' recovery, the model invented with DT algorithm was discovered to be the most precise with 99.85% exactness which has all the earmarks of being the most noteworthy among others.

3. DATA PREPARATION

Collection of data is the vital step to induce data over corona virus. The information was collected from the distinction health care center in Sulaymaniyah City in the Kurdistan Region of IRAQ. The dataset comprises 1376 patients which have appeared side effects of crown infection. The data collection comprises seven factors (Gender, Age, status, O_2, ventilate, Day of hospitalization, and death patient). The informational index contained data about hospitalized patients with COVID-19. After informational index start another stage data preprocessing. Information preparing is a significant cycle being developed of ML model. The information gathered is frequently approximately controlled with out-of-range esteems, missing values; and so, on such information can deceive the consequence of the examination. Weka, one of the expansively utilized data mining computer program, is utilized for the classification. The processing of data preparation illustrated in Figure 1.

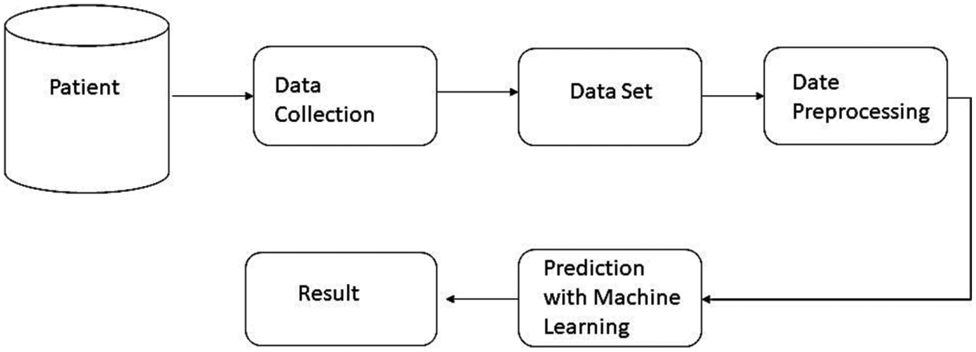

Fig. 1. The workflow diagram for coronavirus disease.

After preprocessing stage of the dataset, the collected variables are divided into two classes "Death" assigned as "Yes" and "No Death" assigned as "No." The selected data samples are transferred to a spreadsheet file for further processing to be suitable for data mining approaches. The dataset were normalized to minimize the effect of scaling on the data and saved as a commas separated value file format. In Table 1, all the attributes explained with their description.

4. METHODOLOGY

Recently, ML techniques have been used to medical prediction; there are different types of ML algorithms that can be applied to different types of applications in various fields [14]. Many different types of research have demonstrated that algorithms of ML had given better help to clinical backings moreover for decision-making on the basis of the patient information. In the medical services field, illness predictive examination is one of the valuable and strong uses of ML forecast algorithms. In this paper proposed a machine-learning algorithm to analyze unusual COVIDE-19 disease datasets. In this paper, rate of death across the region analyzed based on the factors explained in Table 1.

4.1. SVM

SVM is one supervised classification algorithm which is commonly utilized for linear classification and regression problem. It means SVM can solve both linear and nonlinear problems. SVM provides unique and optimal solution, the kernel function is selected based on the points of the variables in the hyperplane. The best separating hyper plane can be written as, $W.X + b = 0$, Where w is a weight vector, the value of the attributes is referred as x, and b is scalar often referred as bias [15].

TABLE 1: Attribute's description used for predication does the patient recovered or died

Variables	Description	Possible values
Gender	It is a social definition of men and women.	Male, Female
Age	Patient age	Date
Status	Situation of the patients' status.	Bad, Severe, Good, Critical
O_2	It indicates does the patient need oxygen or not.	Yes, No
Ventilate	A ventilator uses pressure to blow air or air with extra oxygen	Yes, No
Date of hospital admission	Day of hospitalization	Date
Death	Does the patient died or recovered?	Yes, No

4.2. DT

DT is a supervised learning algorithm that can be utilized for both classification and regression issues, however generally it is ideal for attempting Classification issues. It is a DT classifier; the structure of this algorithm is divided by three parts: Internal node which is features of dataset, branches are demonstrating rules, and leaf is represent outcome for each leaf.

4.3. Naïve Bias

NB classifier is the simple and powerful supervised machine-learning algorithm used for predictive modeling. It considers all variables contribute in the direction of arrangement and they are equally connected [16]. The algorithm is based on a theorem called Bayesian Theorem and used when the coordination of the inputs is high, which assumes that features are statistically independent.

5. EXPERIMENT RESULTS

For the experiment Weka tool have been used, the dataset was collected used to train the above algorithms using the Weka tool. In this paper, the dataset is divided for two parts, for the classification algorithms first part which is 80% used for training the classification algorithms and the second part which is 20% used as a test set and the results are illustrated in Table 2. The achievement of every algorithm was assessed at phases of the training set. Every algorithm was trained with the record sets having 1100 records. This examination is carried out to achieve which algorithm can be the most appropriate for the prediction of COVID-19.

The accuracy of the forecast algorithm in almost all of the research work has exploited like one of the regular measurability while working on the forecast algorithm. In

TABLE 2: Accuracy classification algorithms

S. No.	Classification algorithms	Accuracy
1	Decision tree	96.07
2	Support vector machines	95.27
3	Naive Bayes	94.47

this paper, the accuracy forecast is whether the patient is recovered or deceased while the patient infected by the COVID-19. Base on the above algorithms mentioned in Table 2 and in Figure 2. Each classification algorithm has an alternate expectation precision dependent on its hyper parameters.

TABLE 3: Error metrics for the classification algorithms

S. No.	Algorithm	Kappa statistics	Mean absolute	Root mean square
1	Decision Tree	0.29	0.07	0.19
2	Support Vector Machine s	0.42	0.10	0.21
3	Naive Bayes	0.32	0.12	0.22

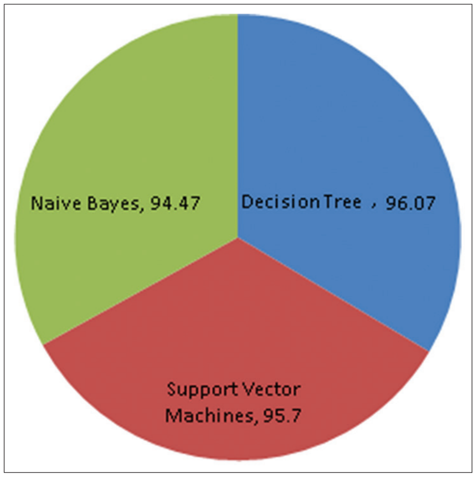

Fig. 2. Accuracy classification algorithms.

Table 3. Describe the performance error measurement for each algorithm; the error metrics which are kappa statistics, mean absolute error, and root mean square error for each algorithm is assessed.

As shown in Table 3. The decision tree has lowest error rates compared to other algorithms.

According to Figure 3, which is the visualization tree for the DT algorithm and Figure 4, the main factor which is the ventilator has the maximum effect on patients, which made it the beginning tree and this cause has the most effect on patients to recover or not. If the patient is not recovered depend on the most second-factor attribute which is status and the rest of the other attributes showed in the DT has a type of impact on the patient is recovery or died. However, the main factor attributes are ventilator, status as shown in Figure 3. This means that if a patient attribute ventilator is yes and the status are bad the patient died otherwise status factors Severe and Critical mostly recovered. In addition, the interesting Status attribute is good which is depending on the patient's date of hospital admission, as shown in Figure 3. It means some patients are in a good status, but they are dead. Because epidemic COVID-19 has more influence in cold weather as shown in Figure 5, it means weather conditions can increase cause death because of COVID-19 in the winter season.

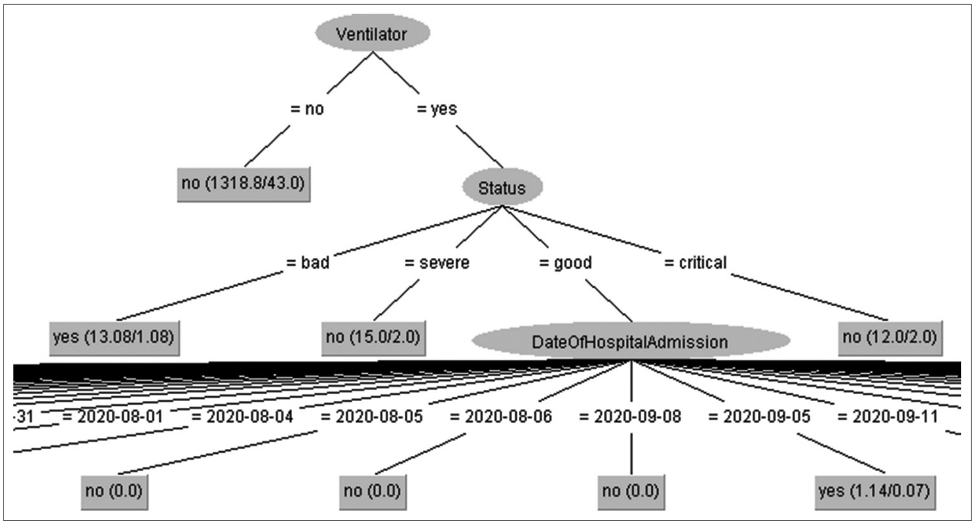

Fig. 3. A decision tree generated by the C4.5 algorithm for predicting COVID-19.

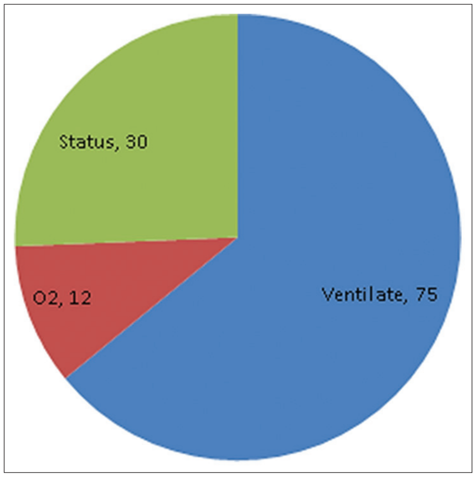

Fig. 4.: Factors with a significant effect on patient mortality.

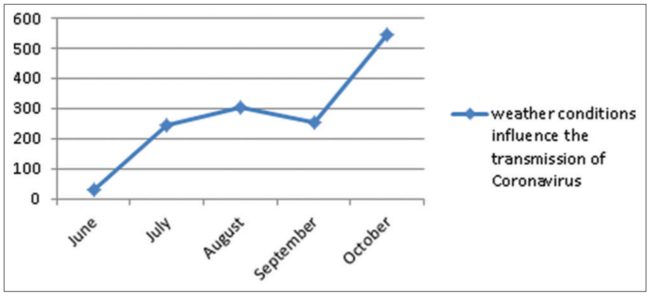

Fig. 5. The role of weather condition on transmission rates of the coronavirus.

6. CONCLUSION

The COVID-19 pandemic lay-down medical care systems in the entire world into a difficult situation. Computer algorithms and ML can help humanity to finding best solution to overcome the coronavirus epidemic. In this paper, data mining technique was used for the predication of coronaviruses infected patient's using dataset of coronaviruses patients of Iraqi-Kurdistan region. DT, support vector machine and NB were used directly on the dataset using Weka ML tool. To identify the accuracy suggested algorithms, the accuracy of the algorithms has been calculated based on the dataset features that have been used. The experiment result showed that the DT has the highest percentage of accuracy which is 96.7% followed by Support Vector Machine which is 95.27 accuracy and Naïve Bayes which is 94.47% accuracy. The experiment result showed that the most effective reason for the patient to recover or not is ventilator and other factors have their effect on the patients to recover or not. In addition, the weather condition means with the coming of the cold weather the virus's effects will increase.

REFERENCES

[1] Medscape Medical News. *The WHO Declares Public Health Emergency for Novel Coronavirus*, 2020. Available from: https://www.medscape.com/viewarticle/924596.

[2] M. C. Collivignarelli, C. Collivignarelli, M. Carnevale Miino, A. Abbà, R. Pedrazzani and G. Bertanza. "SARS-CoV-2 in sewer systems and connected facilities". *Process Safety and Environmental Protection*, vol. 143, pp. 196-203, 2020.

[3] P. Shi, Y. Dong, H. Yan, C. Zhao, X. Li, W. Liu, M. He, S. Tang and S. Xi. "Impact of temperature on the dynamics of the COVID-19 outbreak in China". *Science of the Total Environment*, vol. 728, p. 138890, 2020.

[4] S. Boccaletti, W. Ditto, G. Mindlin and A. Atangana, A. "Modeling and forecasting of epidemic spreading: The case of Covid-19 and beyond". *Chaos Solitons Fractals*, vol. 135, p. 109794, 2020.

[5] T. Davenport and R. Kalakota. "The potential for artificial intelligence in healthcare". *Future Healthcare Journal*, vol. 6, no. 2, pp. 94-98, 2019.

[6] P. Theerthagiri, I. J. Jacob, A. U. Ruby and Y. Vamsidhar. *An Investigation of Machine Learning Algorithms on COVID-19 Dataset*, 2020.

[7] M. Al Sadig and K. N. Abdul Sattar. "Developing a prediction model using j48 algorithm to predict symptoms of COVID-19 causing death". *International Journal of Computer Science and Network Security*, vol. 20, no. 8, p. 80, 2020.

[8] S. S. Arun and G. N. Iyer. "*On the Analysis of COVID19-Novel Corona Viral Disease Pandemic Spread Data Using Machine Learning Techniques*. 4th International Conference on Intelligent Computing and Control Systems, pp. 1222-1227, 2020.

[9] J. Bullock, A. Luccioni, K. H. Pham, C. S. N. Lam and M. Luengo-Oroz. Mapping the landscape of artificial intelligence applications against COVID-19". *Journal of Artificial Intelligence Research*, vol. 69, pp. 807-845, 2020.

[10] L. Wang and A. Wong. "COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 Cases from chest radiography images". *Scientific Reports*, vol. 10, p. 19549, 2020.

[11] Alibaba Cloud e-Magazine. "*Alibaba Cloud Helps Fight COVID-19 through Technology*". Alibaba Cloud, 2020.

[12] Kementerian Kesehatan RI. Pedoman Pencegahan dan Pengendalian Coronavirus Disease (COVID-19). In: L. Aziza, A. Aqmarina and M. Ihsan (Eds.), Revisi Ke4. Kementerian Kesehatan RI, Direktorat Jenderal Pencegahan dan Pengendalian Penyakit (P2P), 2020. Available from: https://www.infeksiemerging.kemkes.go.id.

[13] L. J. Muhammad, M. M. Islam, U. S. Sharif and S. I. Ayon. "Predictive data mining models for novel coronavirus (COVID-19) infected patients recovery". *SN Computer Science*, vol. 1, no. 4, p. 206, 2020.

[14] Sirwan. M. Aziz and Ardalan. H. Awlla. "Performance to build effective student using data mining techniques". *UHD Journal of Science and Technology*, vol. 3, no. 2, p. 10, 2019.

[15] R. Sukanya and K. Prabha. "Comparative analysis for prediction of rainfall using data mining techniques with artificial neural network". *International Journal of Computational Science and Engineering*, vol. 5, pp. 1-5, 2017.

[16] S. D. Jadhav and H. Channe. "Comparative study of K-NN, naive Bayes and decision tree classification techniques". *International Journal of Science and Research*, vol. 5, pp. 1842-1845, 2016.