Developing a Prediction Model Using J48 Algorithm to Predict Symptoms of COVID-19 Causing Death

Dr. Mutasim Al Sadig[1], Dr. Khalid Nazim Abdul Sattar[2]

Assistant Professor, Department of CSI, College of Science, Majmaah University, Majmaah 11952, Saudi Arabia,
ORCID: 0000-0002-5614-4527 1, 0000-0002-0759-0512 2

Summary
COVID-19 has swept the world since the end of the last year, with a cumulative total of more than 20,053,818 infected people worldwide, has caused over 734,556 deaths so far. The symptoms of the disease differ from one person to another, as well as the cause of death. A predictive model using J48 algorithm has been devised using data mining technique to determine the most symptoms of disease that may cause death. COVID-19 research is based on the dataset as provided by the dimensions site Developed by Digital Science in collaboration with over 100 leading research organizations around the world. The model predicts the most common symptoms causing death are Acute kidney injury and Coronary Heart Disease.

Key words:
COVID-19, Coronavirus Symptoms, Data Mining, Decision Tree, Classification, WEKA

1. Introduction

Novel coronavirus disease (COVID-19) is spreading worldwide and has caused over 734,556 deaths so far [1]. COVID-19 symptoms are reported to be ranging from mild to severe illness and may include Cough, Fever, difficulty breathing, Repeated shaking with chills, Muscle pain, Chills, Headache, Sore throat and New loss of taste or smell[1].

The diagnosis of COVID-19 surveillance and the death reason can be optimized, using one of the artificial intelligence methods. In this work, a modest attempt made to identify the most common symptoms of the COVID-19 that causes morbidity based on the use of Data Mining techniques and machine learning algorithms. Data mining is the process that uses machine learning technique to extract and identify interesting patterns in dataset and subsequently gain knowledge that can then be used in decision making[2][3].

A critical factor inhibiting effective Data Mining techniques in the medical fields is a lack of relevant data on patient outcomes. Due to the importance of data related to COVID-19 in research and decision-making, many agencies and institutions have provided data. For example, COVID analytics site sharing dataset, which aggregates data from over 160 published clinical studies and preprints released between December 2019 and April 2020[4].

Manuscript received August 5, 2020
Manuscript revised August 20, 2020

The World Health Organization is also working with international experts, governments, and partners to accelerate the expansion of scientific knowledge on this new virus[5]. The intent of this proposed system is to develop a predictive model, based on Machine Learning techniques, for the prediction of the death causing symptoms of COVID-19.

2. Related Work

Research on clinical record examination to decide examples of illness bunches utilizing the C4.5 calculation and its consequences of the C4.5 calculation estimation can break down the patterns of ailments experienced by general society[6].

In [7], the authors used the C4.5 algorithm for the diagnosis of COVID-19 and it was successfully modeled into a decision tree with PDP, ODP, and OTG classification.

Brinati, Davide Campagner et.al, developed two machine learning models where accuracy ranges between 82% and 86%, sensitivity is between 92% and 95%, to differentiate between positive or negative patients for SARS-CoV-2.

The authors in [9], have compared EJ48, REPTRee, and User Classifier using performance measures like Accuracy, Error Rate, Consumption Time using WEKA tool. Additionally, they have also compared these classifiers based various accuracy measures. The outcome of their research shows that EJ48 classifier has better classification accuracy in comparison with other classifiers.

3. J48 Algorithm

For classification and reduced error pruning this algorithm uses a greedy technique to induce decision trees [6]. It is an open source Java implementation of the simple C4, it is the ID3 extension. J48 calculates the result value of the new sample using predictive models and is based on various attribute values of available data. The different properties are represented by the internal nodes of the

decision tree, the branches between nodes indicate the possible values, where attributes are included in the observed sample, and the terminal node indicates the final value (classification) of the dependent variable[10].

4. Research Methodology

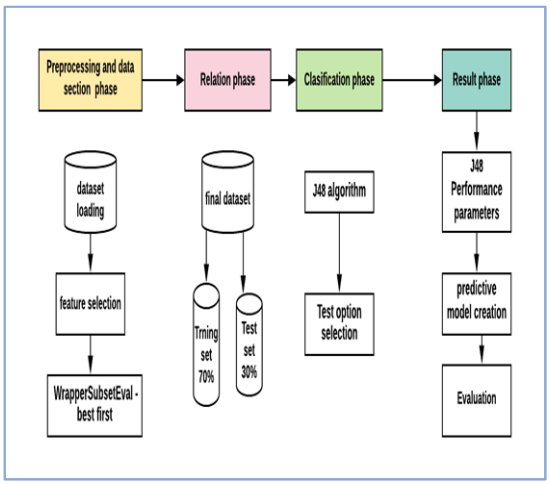

Fig. 1 Model Framework

The J48 decision tree algorithm has been used for the purpose of classification. The proposed classification using WEKA tools is as follows.

(1) datasets collected from Dimensions site that links many research knowledge systems.
(2) The dataset was preprocessed and converted to csv file.
(3) csv file loaded on WEKA system.
(4) classifier algorithm (J48) was used for selected attributes.
(5) Then, split 70.0% training dataset test option were used.
(6) Then, instance feature selected.
(7) After choosing the test option, the results were obtained.

Dataset taken from dimensions site. from study that posted on 02.07.2020 by Jin Hui about COVID-19 patients in-Jinhua Municipal Central Hospital [11].
The data set contained 183 patients, with 18 attributes and 1 class attributes. Some attributes are numeric, while others are nominal. The dataset is in .cvs format. Data has been collected on the following cases: fever, dry cough, Difficulty breathing, chest pain, Respiratory rate R, SPO2, Fatigue, headache, diarrhea, Heart rate, Acute kidney injury, Gender, Acute liver dysfunction, hypertension, Coronary Heart Disease, diabetes and Severe/Critical, Class attributes is about death case, as shown in Table 1.

Table 1: Attributes Description

Patient age	Numeric
Fever	Nominal
Dry cough	Nominal
Difficulty breathing	Nominal
Chest pain	Nominal
Respiratory rate R	Numeric
Spo2	Numeric
Fatigue	Nominal
Headache	Nominal
Diarrhea	Nominal
Heart rate	Nominal
Acute kidney injury	Nominal
Gender	Nominal
Acute liver dysfunction	Nominal
Hypertension	Nominal
Coronary heart disease	Nominal
Diabetes	Nominal
Severe/critical	Nominal
Class \ death	Nominal

Data preprocessing has the area of making data appropriate for analysis. It also improves the quality of data mining technique or tool better [2]. In this study, split 70.0% training dataset was used for attribute selection mode. Wrapper Subset Eval used as an attribute evaluator. It evaluates attribute sets by using a learning scheme.

One of the salient features in the process of data mining is the evaluation of the algorithms. for accessing the performance accuracy of the proposed classifier, confusion matrix, learning curves and receiver operating curves (ROC) have been measured

The various evaluation parameters were measured in terms of measured in terms of precision (1), recall (2), F-measure (3), accuracy (4), time taken to build a model, correctly classified instances percent, and incorrectly classified instances percent.

$$\text{Precision} = TP/(TP+FP) \quad (1)$$

$$\text{Recall} = TP/(TP+FN) \quad (2)$$

$$\text{F-measure} = (2 * \text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (3)$$

$$\text{Accuracy} = (TP+TN) / (TP+TN+FP+FN) \quad (4)$$

Where:
TP = the number of positive cases is classified as positive

TN = the number of negative cases is classified as positive

FP = the number of negative cases is classified as negative

FN = the number of positive cases is classified as negative

5. Results and Discussion

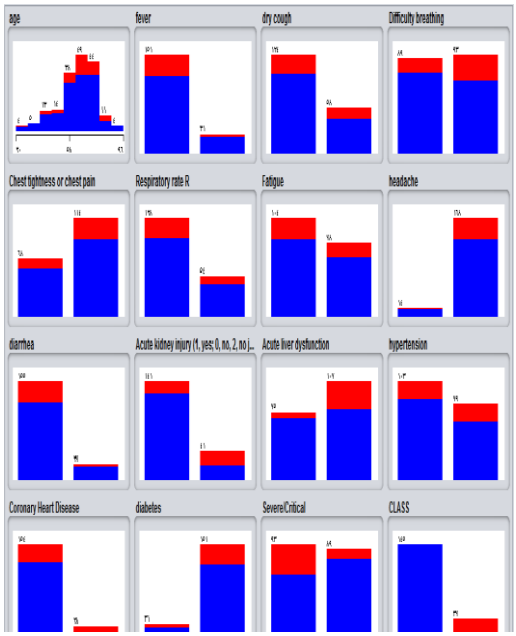

Fig. 2 Attributes Visualizer

Unsupervised classifier has been utilized for experimentation on WEKA information mining instrument, Scheme: WEKA.classifiers.trees.J48 -R -N 3 -Q 1 -M 2. CVS file with 182 instances and 16 attributes witch include : age, fever, dry cough, Difficulty breathing, Chest tightness or chest pain, Respiratory rate R, Fatigue, headache, diarrhea, Acute kidney injury (1, yes; 0, no, 2, no judgment), Acute liver dysfunction, hypertension, Coronary Heart Disease, diabetes, Severe/Critical, CLASS. Test mode: split 70.0% train, remainder test

```
Acute kidney injury (1, yes; 0, no, 2, no judgment) = F: N (95.0/12.0)
Acute kidney injury (1, yes; 0, no, 2, no judgment) = T
|  diarrhea = F
|  |  Chest tightness or chest pain = T: N (10.0/2.0)
|  |  Chest tightness or chest pain = F
|  |  |  Coronary Heart Disease = F
|  |  |  |  Severe/Critical  = T: Y (10.0/2.0)
|  |  |  |  Severe/Critical  = F: N (2.0/1.0)
|  |  |  Coronary Heart Disease = T: Y (2.0)
|  diarrhea = T: N (3.0)

Number of Leaves  :          6

Size of the tree :     11
```

Fig. 3 J48 Pruned Tree

Table 2 :Performance parameters and their values for J48 algorithm

S. No.	*Parameters*	*J48*
1	Correctly Classified Instances	61.818
2	Incorrectly Classified Instances	38.181
3	Kappa statistic	0.0204
4	Mean absolute error	0.357
5	Root mean squared error	0.461
6	Prediction Accuracy	62%
7	TP Rate Average	0.643
8	Precision Average	0.021
9	Recall Average	0.631
10	Time taken to build model (in Sec)	0.01
11	Time taken to test model on test split (in Sec)	0.02
12	Total Number of Instances	182
13	Total Number of attributes	16

Table 2 shows the Performance parameters and their values for J48 classification algorithm. It shows Prediction Accuracy (62%) , precision average (0.021), recall average (0.631) and Time required to Build Model (0.02 sec).

As illustrated in Table 3, the performance of a classification model according to the calculation of testing data shows the confusion matrix for the training dataset as used in the model.

Table 3: Confusion Matrix

a	*b*	*Classified as*
12	30	a = N
4	9	b=Y

J48 pruned tree for model test, shows that, root node is " Acute kidney injury ", this means that this result is the most important result and closely related to COVID-19 patient death.

The following algorithm as illustrated below explains the most common COVID-19 symptoms that may result in cause of death.

a. Algorithm_COVID-19_Symptoms:

```
IF (Acute kidney injury = T) THEN
{
IF ( Coronary Heart Disease= T) THEN
DEATH = YES
ELSE
IF (  Severe/Critical = T) THEN
 DEATH = YES
}
ELSE
DEATH = FALSE
```

6. Conclusion

The predictive model was based on dataset that posted on 02.07.2020, by Jin Huin, in dimensions site. J48 classification algorithm applied using WEKA tools to generate model, result obtained in the form of a decision tree.

The model generated correct predictions accuracy of 62%, with Precision Average 0.021 and Recall Average 0.631.

The J48 predictive model for COVID-19 symptoms that cause death has been successfully developed by using percentage split 70% for training and 30% for testing.

In this study, 14 of the symptoms of coronavirus were used to determine the most Severe symptoms that cause death.

References

[1] COVID-19 Coronavirus Outbreak [Internet]. Dadax. 2020 [cited March 14, 2020]. [Online].Available: https://www.worldometers.info/coronavirus/

[2] Turban, E., et al. Decision Support and Intelligent Systems. Upper Saddle River, NJ: Prentice Hall, 2005.

[3] F. Alam and S. Pachauri, “Comparative Study of J48 , Naive Bayes and One-R Classification Technique for Credit Card Fraud Detection using WEKA,” Adv. Comput. Sci. Technol., vol. 10, no. 6, pp. 1731–1743, 2017.

[4] [Online] : https://www.COVIDanalytics.io/dataset

[5] [Online]:https://www.who.int/ar/emergencies/

[6] diseases/novel-coronavirus-2019. Rafiska, R., Defit, S., & Nurcahyo, G. W. (2018). Analisis Rekam Medis untuk Menentukan Pola Kelompok Penyakit Menggunakan Algoritma C4. 5. Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi), 2(1), 391–396.

[7] Wildan Wiguna ; Dwiza Riana ; DIAGNOSIS OF CORONAVIRUS DISEASE 2019 (COVID-19) SURVEILLANCE USING C4.5 ALGORITHM; Jurnal PILAR Nusa Mandiri Vol. 16, No. 1 March 2020.

[8] Brinati D, Campagner A, Ferrari D, Locatelli M, Banfi G, Cabitza F. Detection of COVID-19 Infection from Routine Blood Exams with Machine Learning: A Feasibility Study. J Med Syst. 2020;44(8):135. Published 2020 Jul 1. doi:10.1007/s10916-020-01597-4.

[9] A. Kaur, E. Roop, and L. Sharma, “Classification and Prediction based Enhanced J48 and REPTree Algorithms to Predict Corona Virus Pandemic,” no. July 2020.

[10] F. Alam and S. Pachauri, “Comparative Study of J48 , Naive Bayes and One-R Classification Technique for Credit Card Fraud Detection using WEKA,” Adv. Comput. Sci. Technol., vol. 10, no. 6, pp. 1731–1743, 2017.

[11] [Online] https://app.dimensions.ai/details/data_set/12423800

Dr. Mutasim Mohamed Al Sadig, Assistant Professor, Department of CSI, College of Science, Majmaah University, Az Zulfi, Kingdom of Saudi Arabia, received his Bs in Science and Technology & M. Ec(CS) from GAZERA University, Sudan. PhD in Information Technology, from Academic Sudan University. He has 20 + years of research and teaching experience and his research expertise include but are not limited to and Data Mining, Biometric Identification System , Software Engineering, Artificial Intelligence, and Computer Networks.

Dr. Khalid Nazim Abdul Sattar, Assistant Professor, Department of CSI, College of Science, Majmaah University, Az Zulfi, Kingdom of Saudi Arabia, received his B. E. (CSE) Degree from Bangalore University, M. Tech(CSE) from VTU Belagavi, and Ph.D. in Computer Science & Engineering from Singhania University, Rajasthan, India. He has 20 years of research and teaching experience and his research expertise include Image processing, Data Mining, Digital Signal Processing, Artificial Intelligence, Internet of Things, and Data & Network Security.