JAZ INDIA

Journal of Advanced Zoology

ISSN: 0253-7214
Volume 44 Issue S-6 Year 2023 Page 1591:1597

Applications of Datamining Techniques for Predicting the Post - Covid 19 Symptoms in Saudi Arabia, Jazan

Dr. Rasitha Banu[1]*, Dr. N. Sasikala[2], Dr. Amal Ramadan[3], Thani Babikar[4], Dr.Maha Yousif Rizgalla[5], Ashraf Abdelmageid Ibrahim khattab[6]

[1,2]Assistant Professor, Department of Health Informatics, CPHTM, JazanUniversity, Jazan, Saudi Arabia
[3,5]Lecturer, Department of Health Education and Promotion, CPHTM, JazanUniversity,Jazan,Saudi Arabia
[4,6]Assistant Professor, Department of Health Education and Promotion, CPHTM, JazanUniversity, Jazan,Saudi Arabia

**Corresponding author's: Rasitha Banu*

Article History

Received: 06 June 2023
Revised: 05 Sept 2023
Accepted: 06 Dec 2023

Abstract

Background *The entire world was combating COVID-19; however, a significant proportion of patients demonstrate the persistence of some COVID-19 symptoms, new symptom development, or exaggeration of pre-existing disease after a negative viral load. They are referred to as a post-COVID-19 syndrome. According to various researches, COVID-19 has a wide range of long-term effects on virtually all systems, including the respiratory, cardiovascular, gastrointestinal, neurological, mental, and dermatological systems. Finding the various symptoms of post-acute and chronic is critical since they might have a significant impact on the patients' everyday functioning. As a result, we aimed to distinguish the symptoms immediately after the initial phase in which the symptoms affected them for more than three weeks using data mining techniques.* ***Methodology:*** *Post-COVID conditions do not affect everyone the same way. They can cause various types and combinations of symptoms in different people. The purpose of this research is to analyse the complications of post covid-19 syndrome. The purpose of Data mining is for discovering the knowledge from vast amount of database. To classify the symptoms of post covid-19, data mining techniques is used. In this study, ranking method was used in preprocessing to select subset of attributes for strengthening the rate of accuracy of classifiers. The data were collected through Google form of 384 household of students from Public Health College in Jazan University. The WEKA open-source software is used for this research work under Windows7 environment. An experimental study is carried out using data mining technique such as J48 and Random Forest tree. The data records are classified as six categories such as General symptoms, Nervous symptoms, Respiratory symptoms, Heart symptoms, Digestive symptoms and normal.* ***Result:*** *The performances of classifiers are evaluated through the confusion matrix in terms of accuracy, time taken to build the Model and error rate. It has been concluded that Random Forest Tree gives better accuracy, minimum time taken to build the model and less error rate than the J48 classifier.*

Keywords: *Post-Covid-19, Data Mining, Classification, Decision Tree, Confusion matrix.*

1. Introduction

A range of new, recurring, or persistent symptoms that people suffer longer than four weeks after contracting COVID-19 are collectively referred to as post-COVID-19 syndrome. Post-COVID-19 syndrome can cause incapacity or linger for months or years in certain individuals. According to research, 1 in 5 adults between the ages of 18 and 64 have at least one medical problem that may be related to COVID-19 between one month and one year after contracting the virus. One in four adults 65 years of age and older have at least one medical issue that may be related to COVID-19. These medical issues are categorized as Neurological symptoms or mental health conditions, including difficulty thinking or concentrating, headache, sleep problems, dizziness when you stand, pins-and-needles feeling, loss of smell or taste, and depression or anxiety and Joint or muscle pain, Heart symptoms or conditions, including chest pain and fast or pounding heartbeat, Respiratory symptoms , including difficulty in breathing, continuous cough, Digestive symptoms, including diarrhea and stomach pain, and General symptoms includes other symptoms, such as a rash and changes in the menstrual cycle.

In the past, several studies revealed the psychological symptoms, weariness, and restrictions in fitness, and a lower quality of life followed in SARS infection. Q-fever, Legionnaires' disease, mononucleosis, and pandemic polyarthritis are some other infectious disorders that cause post-infectious symptoms. [9] Six months following these infections, patients reported fatigue, cognition issues, musculoskeletal pain, and mood disturbances, according to prospective research. [9] Although these findings cannot be generalized to patients with COVID-19, it is believed that many of the patients would recover within 3 weeks many of the patients will experience physical, cognitive, or psychological complaints and more specifically persistent fatigue.

Consequently, our goal is to identify the symptoms as soon as possible following the initial phase, during which they persisted for longer than three weeks. Acute COVID-19 symptoms last for a few weeks after infection, but chronic COVID-19 symptoms last for more than three months in the infected person. We aimed to distinguish the symptoms immediately after the initial phase in which the symptoms affected them for more than three weeks using data mining techniques.

Post-covid syndrome:

Most patients with COVID-19 (coronavirus disease) recover in a few weeks. However, some people may experience symptoms for a very long period following the disease, even those who had milder forms of it. The terms post-COVID-19 syndrome, post-COVID conditions, long-term COVID-19, long-haul COVID-19, and post-acute sequelae of SARS COV-2 infection (PASC) are occasionally used to describe these persistent health issues. Anyone can now differentiate between post-acute COVID-19 and chronic COVID-19 (i.e. long-COVID). Post-acute COVID-19 symptoms are those that last more than three weeks after the initial Covid -19 symptoms, while long-COVID symptoms are those that last more than 12 weeks after the initial symptoms. [7] Although the difference between post-acute and long-term COVID is somewhat arbitrary, it is important to distinguish between the stages in order to better understand and research the short- and long-term effects of COVID-19.

The illness and other symptoms of covid-19 that persist beyond the acute illness are referred to as "Chronic " Covid - 19 by several people. If symptoms persist for more than 3–4 weeks, they are referred to as acute, when symptoms persist for more than 12 weeks; they are referred to as chronic. The existence of Covid-19 has yet to be properly confirmed. There is no requirement for a positive Covid-19 test for the diagnosis of post-acute or chronic illness.

Related Work:

‘Emma Ladds and Alex Rushforth’ suggested some quality principles for an extended Covid service for ensuring access to worry, reducing the burden of illness, taking clinical responsibility and providing continuity of care, multi-disciplinary rehabilitation, evidence-based investigation and management, and any development of the cognitive content and clinical services." [12]

Sophie A M van Kessel, Tim C Olde Hartman et.al stated that the frequency of patients with persistent symptoms ranges from 100% to 35%. Numerous studies have revealed that fatigue is the most common or second most common symptom of post-acute COVID-19 and long-COVID. Symptoms occur on a regular basis. Furthermore, most papers define cough, discomfort, and headache as square measure mental and psychological feature symptoms." [13].

They studied 442 and 353 individuals over four and seven months, respectively, when symptoms first appeared. Four months after SARS-CoV-2 infection, 86% (38/442) of patients were diagnosed with shortness of breath, 124% (55/442) with dysomia, 111% (49/442) with ageusia, and 97% (43/442) with fatigue. [14]

According to Bircan Kayaaslan, Fatma Eser and Ayse K. Kalem, the severity of acute illness presentation is one of the most important indicators of post COVID - 19 syndrome. If the intensity persists, it may be treated as a chronic covid symptom, which may result in the development of complications, physiological abnormalities, and deconditioning; this is an expected outcome.

2. Materials And Methods

Data Collection:

A simple pre-coded questionnaire will be developed, and data is collected from 384 household students from the Public Health College in Jazan University. The data included patient demographic data, vaccination status, and patient status during COVID-19, and post-COVID-19 syndrome. There are 384 instances in post covid syndrome dataset. In these instances, 9 instances have heart symptoms, 144 instances are No symptoms, 121 instances are General symptoms,56 instances are Neurosymptoms,34 instances are belonging to Respiratory symptoms, 20 instances are belongs to digestive symptoms. Totally there are 16 attributes in dataset. But in our research work we have taken 12 attributes which will be used to classify the data. The post covid syndrome data set is given below table 1.

Table 1: Post Covid dataset

SN	Attribute Name	Values
1	Age	Numeric
2	Gender	M,F
3	Blood group	Nominal
4	Vaccination status before covid positive	Numeric
5	Vaccination details	Nominal
6	Preexisting medical conditions	Nominal
7	Self-immune status	Nominal
8	Food habits	Nominal
9	Physical activity	Nominal
10	When did the post covid symptoms began to appear from the day of recovery	Numeric
11	Symptoms developed after covid recovery	Nominal
12	Class	Neurologic symptoms, Heart symptoms, Respiratory symptoms, Digestive symptoms, General symptoms

Proposed System:

a. Pre-processing

Data pre-processing is the main technique in data processing that involves transforming data into a clear format, and removing irrelevant attributes, filling the missing values and so on. When the Real-world data is collected then there is usually incomplete, inconsistent, and in some places, there will be a lack in certain behaviors or trends and is probably going to contain many errors. Data pre-processing may be a proven method of resolving such issues. There are many tasks in Data pre-processing to prepare raw data for further processing. These tasks include data cleaning, integration, transformation, and data reduction [16].

The data is collected from the 384 household of students from Public Health College in Jazan University. The collected data was checked for the presence of error in data entry including misspellings and missing data. In our work, we have used Replace with missing values filter to fill the missing values to make the data complete and the preprocessing technique was applied to select relevant attributes using the ranking method.

b. Classification

Classification is one among the data mining Technique. It is accustomed classify the data supported similarity of instances. There are two forms of learning. One is supervised and another is unsupervised learning. It is a supervised learning, during which predefined training data is out there. The most popular data processing classification techniques are decision trees and neural networks and so on.

c. Decision tree

One of the most important classification techniques in data mining is Decision Tree. It is a treelike graph. Testing each attribute represented as internal node and each branch represents an outcome of the test and the leaf node represents classes. It is a graphical representation of possible solutions, based on these solutions; optimum course of action is carried out. In this research, we have used two decision tree classifiers such as Random Forest Tree and J48 to classify the post covid data set. The Algorithm of J48 and random forest tree is given below.

J48 Algorithm

J48 is a tree-based learning approach. It is developed by Ross Quinlan which is predicated on iterative dichtomiser (ID3) algorithm. J48 uses divide-and-conquer algorithm to separate a root node into a subset of two partitions till leaf node (target node) occur in tree. Given a group T of total instances the subsequent steps are won't to construct the tree structure.

Step 1: If all the instances in T belong to the same group class or T has fewer instances, than the tree is leaf labelled with the most frequent class in T.

Step 2: If step 1 doesn't occur then select a test supporting one attribute with a minimum of two or greater possible outcomes. Then consider this test as a root node of the tree with one branch of every outcome of the test, partition T into corresponding T1, T2, T3 ...., according to the result for every respective case, and therefore the same could also be applied in recursive thanks to each sub node 13,14. Step 3: Information gain and default gain ratio are ranked using two heuristic criteria by algorithm J48. [17]

Random Forest Tree

Random forest Tree is a collective learning method for enhancing the uses of classification, regression, and other tasks. The algorithm operates by constructing a mess of decision trees at training time and it gives the output as the category with the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set [17].

Step 1: On the off chance that all the occurrences in T have a place to the same gather course or T is having less occurrences, than the tree is leaf labelled with the foremost visit lesson in T.

Step 2: On the off chance that step 1 doesn't happen at that point select a test upheld one property with a least of two or more noteworthy conceivable results. At that point consider this test as a root hub of the tree with one department of each result of the test, parcel T into comparing T1, T2, T3 ...., agreeing to the result for each particular case, and so the same seem moreover be connected in recursive much appreciated to each sub hub 13, 14.

Step 3: Data pick up and default pick up proportion are positioned utilizing two heuristic criteria by calculation J48.

3. Results and Discussion

In this research work, Weka 3.8.4 open-source software can be used to predict the symptoms of post covid -19 syndrome. In Post covid data set, there are 384 instances. In these instances, 9 instances have heart symptoms, 144 instances are No symptoms, 121 instances are General symptoms, 56 instances are Neurosymptoms, 34 instances are belongs to Respiratory symptoms, 20 instances are belongs to digestive symptoms. In pre-processing, we have used ranker method to remove irrelevant attributes and then classifier is applied to the data set. Finally the classifier is evaluated based on confusion matrix,accuracy and error rate.

The following table 2 shows results show the confusion matrix of Random Forest tree classifier before pre processing

Table 2: Confusion matrix of Random Forest tree classifier before pre processing

Target class	Heart Symptoms	Normal	General Symptoms	Neurology symptoms	Respiratory symptoms	Digestive symptoms
Heart Symptoms	9	0	0	0	0	0
Normal	1	136	4	1	1	1
General Symptoms	0	5	112	1	2	1
Neurology symptoms	0	0	1	54	1	0
Respiratory symptoms	0	1	0	2	31	0
Digestive symptoms	0	0	0	1	0	9

Before applying pre-processing technique to Random Forest tree classier, the correctly classified instances are 361 and incorrectly classified instances are 23.

The following table 3 shows results show the confusion matrix of J48 classifier before pre processing

Table 3: Confusion matrix of J48 classifier before pre processing

Target class	Heart Symptoms	Normal	General Symptoms	Neurology symptoms	Respiratory symptoms	Digestive symptoms
Heart Symptoms	9	0	0	0	0	0
Normal	1	135	4	1	1	1
General Symptoms	0	5	112	1	2	1
Neurology symptoms	1	0	1	54	1	0
Respiratory symptoms	0	1	0	2	31	0
Digestive symptoms	0	0	0	1	0	7

Before applying pre-processing technique to J48 classier, the correctly classified instances are 357 and incorrectly classified instances are 27.

The following Table 4 depicts the detailed accuracy, time, error rate for Random Forest tree and J48 classifier.

Table 4: accuracy, time, error rate for Random Forest tree and J48 classifier before preprocessing

Classifier	Accuracy	Error Rate	Time
Random Forest tree	94.01%	5.99%	0.1 seconds
J48	92.96%	7.04%	0.3 seconds

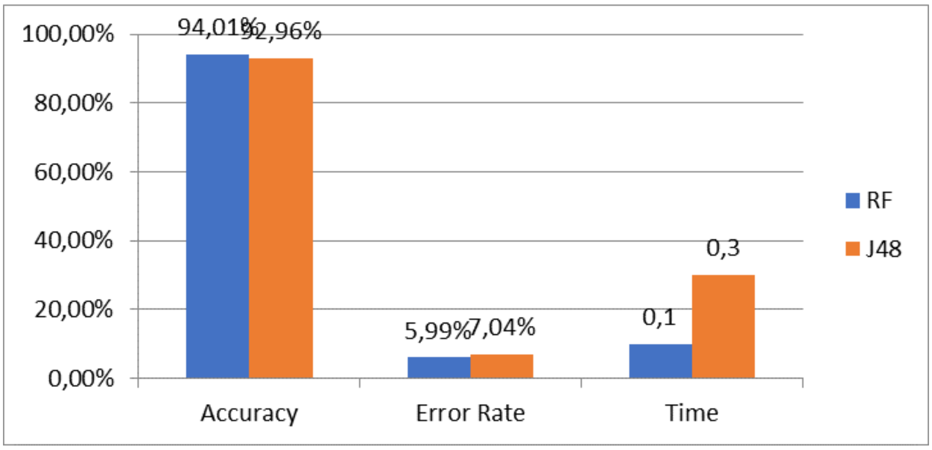

Chart 1: Accuracy, Error rate and Time taken to build model of classifiers before pre-processing

Chart 1 show that the classifier Random Forest Tree is giving highest accuracy, minimum error rate and less time taken to build the model than J48 classifiers.

The following table 5 shows results of confusion matrix of Random Forest tree classifier after pre processing

Table 5: Confusion matrix of Random Forest tree classifier after pre processing

Target class	Heart Symptoms	Normal	General Symptoms	Neurology symptoms	Respiratory symptoms	Digestive symptoms
Heart Symptoms	9	0	0	0	0	0
Normal	1	136	4	1	1	1
General Symptoms	0	5	112	1	2	1
Neurology symptoms	0	0	1	54	1	0

Respiratory symptoms	0	1	0	2	31	0
Digestive symptoms	0	0	0	1	0	9

The following table 6 shows the confusion matrix of Random Forest tree classifier after pre processing

Table 6: Confusion matrix of Random Forest tree classifier after pre processing

Target class	Heart Symptoms	Normal	General Symptoms	Neurology symptoms	Respiratory symptoms	Digestive symptoms
Heart Symptoms	9	0	0	0	0	0
Normal	0	135	4	1	1	0
General Symptoms	0	5	112	1	2	0
Neurology symptoms	1	0	1	54	1	0
Respiratory symptoms	0	0	0	2	31	0
Digestive symptoms	0	0	0	1	0	9

After applying pre-processing technique, the dataset is applied to J48 classier, the correctly classified instances are 363 and incorrectly classified instances are 21.

The following Table 7 depicts the detailed accuracy, time, error rate for Random Forest tree and J48 classifier.

Table 7: Accuracy, time, error rate for Random Forest tree and J48 classifier after preprocessing

classifier	Accuracy	Error Rate	Time
Random Forest tree	**96.88%**	**3.12%**	**0 seconds**
J48	94.53%	5.47%	0.1 seconds

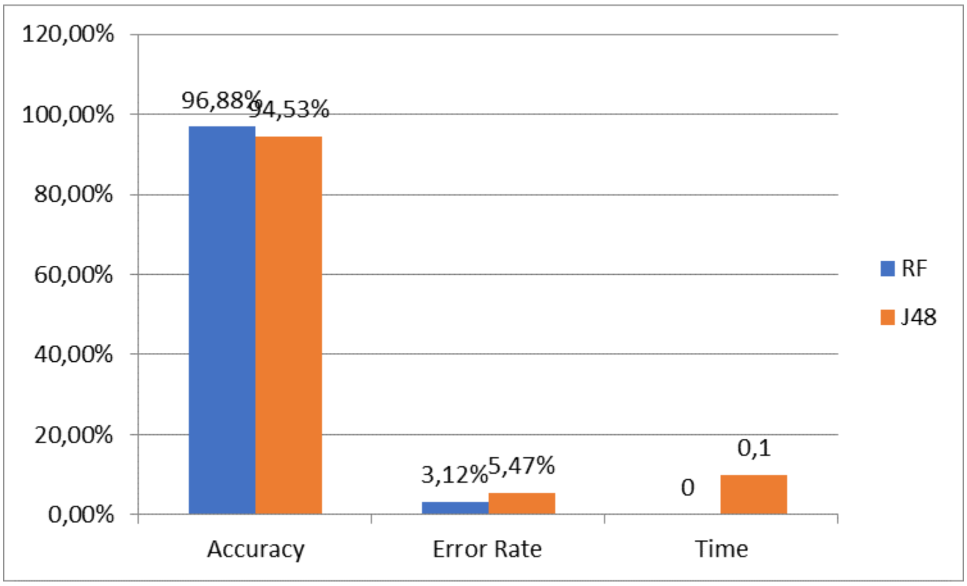

Chart 2: Accuracy, Error rate and Time taken to build model of classifiers After preprocessing

Chart2 shows that the classifier Random Forest Tree is giving highest accuracy, minimum error rate and less time taken to build the model than J48 classifiers. The chart shows that after applying preprocessing the results of classifiers performance has been improved.

4. Conclusion

The diagnosis of disease is one of the most difficult and major tasks in the medical industry. Various data mining techniques have proven to be extremely helpful in decision making. In our work, we have used preprocessing technique to fill up the missing values and we have used ranker method to select the subset of attributes to improve the accuracy of classifier and then dataset is applied to Random Forest Tree and J48 classifiers of data mining classification techniques which are used to predict post covid symptoms. The performances of classifiers are evaluated through the confusion matrix in terms of accuracy, time taken to build the model and error rate. We have compared the results of classifiers before preprocessing with after processing. Before preprocessing the accuracy of Random Forest tree was (94.01%), error rate (5.99%), Time taken to build the model was 0.1 seconds. Before preprocessing the accuracy of J48 was (92.96%), error rate (7.04%), Time taken to build the model was 0.3 seconds. After preprocessing, the J48 gives accuracy (94.53%), gives very minimum error rate (5.47%) and less time taken to build the model (0.1 Seconds). After preprocessing, the Random Forest Tree Algorithm gives high accuracy (96.88%), gives very minimum error rate (3.12%) and less time taken to build the model (0 Seconds) than J48 Algorithm. In Future the Preprocessing techniques will be implemented on different datasets breast cancer, weather, hypothyroid and lung cancer and so on.

References:

[1]. Coronavirus disease 2019 (COVID-19): situation report, 51. World Health Organisation; 2020, (accessed on 11 March 2020).

[2]. Zaim S, Chong JH, Sankaranarayanan V, Harky A. COVID-19 and multiorgan response. Curr Probl Cardiol 2020; 45(8): 100618.

[3].Buitrago-Garcia D, Egli-Gany D, Counotte MJ et al. Occurrence and transmission potential of asymptomatic and presymptomatic SARS-CoV-2 infections: a living systematic review and meta-analysis. PLoS Med 2020; 17(9): e1003346.

[4]. Gavriatopoulou M, Korompoki E, Fotiou D et al. Organ-specific manifestations of COVID-19 infection. Clin Exp Med 2020; 20(4): 493–506.

[5].Machhi J, Herskovitz J, Senan AM et al. The natural history, pathobiology, and clinical manifestations of SARS-CoV-2 infections. J Neuroimmune Pharmacol 2020; 15(3): 359–86.

[6] Carfì A, Bernabei R, Landi F; Gemelli Against COVID-19 Post-Acute Care Study Group. Persistent symptoms in patients after acute COVID-19. JAMA 2020; 324(6): 603–5.

[7].Greenhalgh T, Knight M, A'Court C, Buxton M, Husain L. Management of post-acute covid-19 in primary care. BMJ 2020; 370: m3026.

[8].Tenforde MW, Kim SS, Lindsell CJ et al. ; IVY Network Investigators; CDC COVID-19 Response Team; IVY Network Investigators. Symptom duration and risk factors for delayed return to usual health among outpatients with COVID-19 in a multistate health care systems network - United States, March-June 2020. MMWR Morb Mortal Wkly Rep 2020; 69(30): 993–8.

[9]. Hickie I, Davenport T, Wakefield D et al. ; Dubbo Infection Outcomes Study Group. Post-infective and chronic fatigue syndromes precipitated by viral and non-viral pathogens: prospective cohort study. BMJ 2006; 333(7568): 575.

[10]. Bannister BA. Post-infectious disease syndrome. Postgrad Med J 1988; 64(753): 559–67.

[11]. Chan KS, Zheng JP, Mok YW et al. SARS: prognosis, outcome and sequelae. Respirology 2003; 8 Suppl: S36–40.

[12] Emma Ladds, Alex Rushforth, Sietse Wieringa, Sharon Taylor, Clare Rayner, Laiba Husain & Trisha Greenhalgh, "Persistent symptoms after Covid-19: qualitative study of 114 "long Covid" patients and draft quality principles for services", BMC Health Services Research volume 20, Article number: 1144 (2020)

[13] Lau HM, Lee EW, Wong CN et al. The impact of severe acute respiratory syndrome on the physical profile and quality of life. Arch Phys Med Rehabil 2005; 86(6): 1134–40.

[14] Lee AM, Wong JG, McAlonan GM et al. Stress and psychological distress among SARS survivors 1 year after the outbreak. Can J Psychiatry 2007; 52(4): 233–40.

[15] Sophie A M van Kessel, Tim C Olde Hartman, Peter L B J Lucassen, Cornelia H M van Jaarsveld Family Practice, "Post-acute and long-COVID-19 symptoms in patients with mild diseases: a systematic review"Volume 39, Issue 1, February 2022, Pages 159–167.

[16]. https://en.wikipedia.org/wiki/Data_pre-processing

[17] https://en.wikipedia.org/wiki/Classification