Effect of Corona Virus on Multi-Disease Patients using Association Rule Mining

Vijai Dev
Department of Physics and Computer Science
Faculty of Science
Dayalbagh Educational Institute
Agra, India
vijay.dev.dei@gmail.com

Preetvanti Singh
Department of Physics and Computer Science
Faculty of Science
Dayalbagh Educational Institute,
Agra, India
preetvantisingh@gmail.com

Abstract— **Data mining is used by the industries and scientists for extracting useful information form the raw data. This raw data has been collected from a lot of resources like internet, population survey, market survey, questionnaire etc. In this paper data mining has been applied to the data collected for the pandemic Novel Corona Virus. The two techniques which have been used here are Market Basket Analysis and Frequent Pattern Growth Analysis. Both the techniques analyse support, confidence and lift. The patterns obtained from these two techniques give us information about the patients who are most likely and less likely to catch the virus. This knowledge is useful for formulating policies and rule in order overcome the pandemic situation.**

Keywords— Data Mining, Corona Virus, Market Basket Analysis, Frequent Pattern Growth Analysis.

I. Introduction

A family of viruses known as the corona virus can cause conditions including the common cold, Middle East respiratory syndrome (MERS), and severe acute respiratory syndrome (SARS). In 2019, a new corona virus was identified as the severe acute respiratory syndrome corona virus 2 (SARS-CoV-2) and the disease it induces is now known as corona virus disease 2019 (COVID-19). COVID-19 has become a pandemic due to human-to-human transmission from sick persons.

As of April 12, 2022, 6:59 GMT, the COVID-19 pandemic affected 210 countries worldwide with 500,055,888 confirmed cases, 449,919,848 recovered cases, and 6,206,099 fatalities [1].

Every day, an enormous amount of data connected to the COVID-19 pandemic is generated globally, making it a valuable resource to be mined and evaluated for unique, relevant, and valid knowledge or patterns extraction for better healthcare decision-making.

Data Mining (DM) is a technique used to extract new, practical, and accurate hidden patterns or knowledge from datasets [2]. The technique shows associations, knowledge, or patterns between a single or multiple datasets. It is frequently employed in the diagnosis and prognosis of numerous diseases. Data mining has been widely used in the healthcare sector for a variety of purposes, including patient outcome prediction, health outcome modeling, hospital ranking, and evaluation of treatment efficacy and infection control, stability, and recovery [3]. This study focuses on the application of data mining in analyzing the effect of COVID-19 on patients with multiple diseases.

II. Literature Review

In literature, many authors have applied data mining techniques to analyze the effect of COVID-19. For the purpose of identifying and diagnosing COVID-19, Albahri et al. [4] presented a review on automated artificial intelligence solutions based on data mining and machine learning methods. In order to estimate the number of positive COVID-19 cases, Ayyoubzadeh et al. [5] used long short-term memory models and linear regression to forecast the prevalence of COVID-19 in Iran. According to Buscema et al. [6], the Topological Weighted Centroid technique has the capacity to extract new, pertinent data from small, unreliable datasets. Huang et al. [7] used data mining on Sina Weibo to extract the data of 485 patients with clinical symptoms who had suspected or laboratory-confirmed cases of COVID-19. In order to optimize monitoring techniques in affected areas of India, Kumar [5] used cluster analysis to classify real groups of COVID-19 data sets from various states and union territories. This information can be extremely helpful to the government, physicians, police, and other parties involved in understanding the severity of the COVID-19 outbreak.

To identify COVID-19 related stress symptoms at a spatiotemporal scale in the United States, Li et al. [9] introduced the CorExQ9 method, which combines a correlation explanation learning algorithm with clinical patient health questionnaire vocabulary. During the early phases of the COVID-19 epidemic, Li et al. [10] carried out a quantitative and qualitative analysis of Chinese social media posts on the Chinese micro blogging website Weibo that evolved in Wuhan City. The physiological causes of this clinical finding connecting diabetes with the severity and unfavorable prognosis of COVID-19 were examined by Marhl et al. [11]. To identify similar physiological situations where diabetes and COVID-19 have been studied together, publication mining was used. Maram and Satish [12] employed machine learning classification methods to provide a framework for analyzing the COVID-19 dataset. Using an epidemiological dataset of COVID-19 patients from South Korea, Muhammad et al. [13] created data mining models to predict the recovery of COVID-19 infected patients. The models were created by using the decision tree, support vector machine, naïve Bayes, logistic regression, random forest, and k-nearest neighbor algorithms directly to the dataset using the Python programming language.

Using the keywords Covid-19, mortality, immunity, and vaccination, Radanliev et al. [14] performed data mining on records from the web of science core collection. Corona-related risk issues were highlighted by Stephany et al. [15] along with their assessed importance for various businesses. The CoRisk-Index keeps track of industry-specific risk evaluations as the crisis grows throughout the economy. To expedite the search for current SARS-CoV-2 Mpro inhibitors, Ghosh et al. [16] developed a chemical-informatics based data mining technique using historical activity data of SARS-CoV main protease (Mpro) inhibitors.

The literature review reveals that use of data mining in analyzing Covid symptoms or in predicting its effect is very helpful in making healthcare decisions. This paper uses data mining techniques in analyzing the effect of COVID-19 on patient with multiple diseases.

III. Materials and Methods

Data mining techniques help in making effective decisions in various real-life situations [10]. Market Basket Analysis and Frequent Pattern Growth Analysis are two techniques of data mining used widely for extracting data.

A. The Market Basket Analysis

Market Basket Analysis (MBA) determines the associations' rules between data items. It is a tool of knowledge discovery about co-occurrence of nominal or categorical items. MBA allows decision makers to uncover unobvious and counter intuitive associations between items co-occurring on a frequent basis. Recently, the use of MBAs has expanded outside the discipline of marketing, for example finance, telecommunications and web analysis, geophysics, legal aid services, ocean, land and atmospheric process [17]. The analysis used support, confidence and lift to find patterns.

1) Support

An expression of association rule can be represented in the form of X → Y (X implies Y), where X and Y are two distinct item sets (having no items in common), X is the antecedent and Y is a consequent.

The applicability of a rule to the specified dataset is measured by the support. The support can be derived by using Eq. (1).

$$Support(XY) = \frac{Support\ Count\ of XY}{Total\ Number\ of\ Transaction\ in\ a\ Dataset} \quad (1)$$

The support is referred to as P(A∩B) and is often expressed as a percentage measure with a range of 0 to 100. It is the likelihood that A and B will occur together.

2) Confidence

How frequently items from item set Y appear in transactions that contain item set X is determined by confidence (a metric for the strength of association rules). The confidence can be derived by using Eq. (2).

$$Confidence(X/Y) = \frac{Support\ (XY)}{Support\ (X)} \quad (2)$$

Similar to support, confidence is usually measured in percentage (ranging from 0 to 100) and is defined as P(A∩B)/P(A). The likelihood that a set of items will be selected after another set has already been selected is known as confidence. Confidence is the likelihood that a set of items will be selected given that another set of items has already been selected.

3) Lift

A rule's significance is measured by lift. The ratio between the confidence and the predicted confidence of a rule is used to express the lift value. Any value between zero and infinity can be taken over by the lift. If the lift value is larger than 1, it signifies that the rule body (antecedent) and the rule head (consequence) occur together more frequently than anticipated. The occurrence of the rule body influences the occurrence of the rule head in a positive manner. The occurrence of the rule body has a negative impact on the occurrence of the rule head if the lift value is less than 1. The rule body and rule head appear together as frequently as anticipated when the lift value is close to 1. The lift can be derived by using Eq. (3).

$$L(X \rightarrow Y) = \frac{P\ (X,Y)}{P(X) \cdot P(Y)} \quad (3)$$

B. Frequent Pattern Growth

By compressing the database into a frequent pattern tree and keeping track of the relationships between item sets, the frequent pattern growth approach discovers frequently used patterns. The compressed database is separated into several condition databases where each condition database associates with a frequent item.

A tree-like structure, called a frequent pattern tree, is created using the database's initial item sets. The frequent pattern tree's goal is to extract the most common pattern. Each node of the frequent pattern tree corresponds to an individual item in the item set.

IV. Case Study

This section describes the applicability of both the algorithms for mining the effect of COVID on a multi disease patient. The data to achieve the goal was collected by visiting various hospitals of Agra. The data set comprises of Covid positive patients with multiple diseases as shown in Table I.

TABLE I. Diseases with their Frequencies in the Considered Patient Database

Disease	Short Name	Frequencies
Asthma	As	3
Blue lips or face	BL	1
Blood Pressure	Bp	2
Cholera	Ch	1
Cough	Co	7
Diabetes	Di	8
Fever	Fe	5
Heart	He	6
Kidney	Ki	6
Lungs Cancer	Lc	1
Loss of taste or smell	Lo	1
Sickle cell disease(Anaemia)	Sc	1
Syphilis	Sy	1
Tuberculosis	Tb	1
Tiredness	Ti	1
Tumour	Tu	1

The database was examined to obtain a list of recurrent items and their respective support counts. The collection of frequently occurring items is arranged in descending order of support count, with minimum count of support = 2.

Thus, the combination determined is [Di:8, Co:7, He:6, Ki:6, Fe:5, As:3, Bp:2]. The database is examined second time. Each item in a transaction is ranked according to the sequence of the supplied dataset. A branch is constructed for each transaction comprising four items (As, Co, Di, and He) according to the sequence, resulting in the first branch ⟨(Di:1), (Co:1), (He:1), and (As:1)⟩ for creating an frequent pattern tree. The branch has four nodes. Here, Di is the child link of root, Co links to Di, He links to Co, As links to He, and so on, as shown in Figure 1.

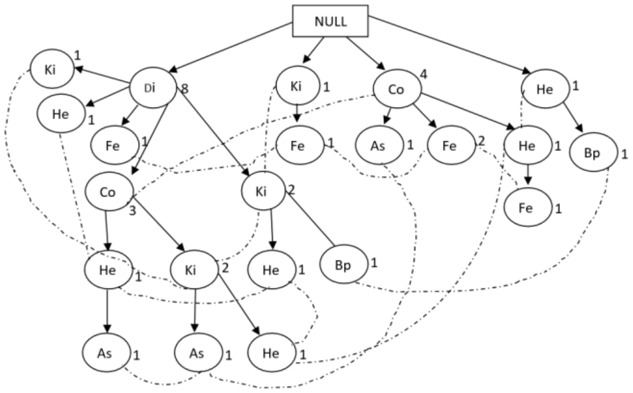

Fig. 1. The FP-Tree for the Problem

After removing the frequencies which are lower than minimal support, the remaining frequencies were organized in decreasing order. A column "Node Link" was added to the table (Table II). This column contains a reference to a linked collection of nodes with identical items.

TABLE II. ORDER DISEASE SET

Source	Node Link		
Di	Co	He	As
Di	Co	Ki	As
Di	He	Ki	
Di	Ki	Bp	
Di	Co	He	Ki
Di	Fe		
Di	He		
Di	Ki		
Co	As		
Co	Fe		
Co	Fe		
Co	He	Fe	
He	Bp		
Ki	Fe		

Now, all the ordered sets of diseases are used to create the FP-Tree (Figure 1).

From the FP-Tree (as shown in Figure 1), the Conditional Pattern Base was determined as

Bp {{Di, Ki : 1} {He : 1}}

As {{Di, Co, He : 1}, {Di, Co, Ki : 1} {Co : 1}}

Fe {{Di : 1} {Ki : 1} {Co : 2}, {Co, He : 1}}

Ki {{Di : 1}, {Di, Co : 2}, {Di : 2}}

He {{Di : 1}, {Di, Co : 1}, { Di, Ki : 1}, {Di, Co, Ki : 1} {Co : 1}}

Co {Di : 3}

Di --

This was used to determine Conditional FP–Tree as:

As {Di : 2, Co : 2}

Fe {Co : 3}

Ki {Di : 5, Co :2}

He {Di : 4, Co : 2, Ki : 2}

Co {Di : 3}

Finally, the frequent pattern rules are generated by matching each item of the conditional frequent pattern tree set with its corresponding item as shown below:

As {Di, As : 2} {Co, As : 2} {Di, Co, As : 2}

Fe {Co, Fe : 3}

Ki {Di, Ki : 5} {Co, Ki : 2} {Di, Co, Ki : 2}

He {Di, He : 4} {Co, He : 2} {Ki, He : 2} {Di, Co, Ki, He : 2}

Co {Di, Co : 3}

A. Market Basket Analysis

Using the frequencies determined in Table II above, different rules (combination) were generated. The selected combinations are shown in Table II. (Items with support < 2 are removed).

TABLE III. COMBINATION OF DISEASES

Combination		Support
As	Co	3
As	Di	2
Co	Di	3
Co	Fe	3
Co	He	3
Co	Ki	2
Di	He	4
Di	Ki	5
He	Ki	2

Next, Support, Confidence and Lift were computed as explained below.

1) Support

Let S be the set of all possible effected and n be the number of patients. Each patient record is a subset of S. Now consider the rules of the form "(x_1, x_2, ..., x_j) implies (y_1, y_2, ..., y_k)" where x_1, x_2, ..., y_1, y_2… are elements of S. The x_1, x_2, ...x_j is known as item set.So, Support Rule is defined as: Supp (x_1, x_2, ...) implies (y_1, y_2, ...) = $\frac{No.ofDisease\ x_1,x_2,...and\ y_1,y_2,...}{n}$

Supp (x_1, x_2, ...) = $\frac{No.\ of\ Patient\ contaning\ x_1,x_2,...}{n}$

In this paper, support is viewed from probability perspective, i.e. probability of transaction having a disease. Referring to Table II, the support is calculated as:

Support (A) = % of Patient Transaction with Disease – Kidney.

$$= \frac{Count\ of\ Transaction\ which\ has\ Kidney}{Total\ Number\ of\ Transaction}$$

P(A) = Probability of a Transaction having Kidney Disease – Kidney = 6/15. Table IV presents the computation of support.

TABLE IV. SUPPORT COMPUTATION

Shot Name	Frequency of Patient	Support	%Support
As	3	0.20	20.00
BL	1	0.07	6.67
Bp	2	0.13	13.33
Ch	1	0.07	6.67
Co	7	0.47	46.67
Di	8	0.53	53.33
Fe	5	0.33	33.33
He	6	0.40	40.00
Ki	6	0.40	40.00
Lc	1	0.07	6.67
Lo	1	0.07	6.67
Sc	1	0.07	6.67
Sy	1	0.07	6.67
Tb	1	0.07	6.67
Ti	1	0.07	6.67
Tu	1	0.07	6.67

In case of Corona, we are interested in finding combinations of disease rather than single ones. Table V lists support for combination of two diseases and Table VI lists support for combination of three diseases.

TABLE V. SUPPORT FOR COMBINATIONS OF TWO DISEASES

Disease	Disease	Support	Support%
As	Co	3	20.00
As	Di	2	13.33
Co	Di	3	20.00
Co	Fe	3	20.00
Co	He	3	20.00
Co	Ki	2	13.33
Di	He	4	26.67
Di	Ki	5	33.33
He	Ki	2	13.33

TABLE VI. SUPPORT FOR COMBINATIONS OF THREE DISEASES

Disease	Disease	Disease	Support	Support%
As	Co	Di	2	13.33
Co	Di	He	2	13.33
Co	Di	Ki	2	13.33
Di	He	Ki	2	13.33

2) Confidence

The Confidence for all rules is computed as: Conf $(x_1,x_2,...)$ implies $(y_1,y_2,...)$ = $\frac{Supp\ ((x_1,x_2,...)implies\ (y_1,y_2,...)}{Supp(x_1,x_2,...)}$. For example, $Confidence(Co \rightarrow Fe) = P(Co/Fe)$

$$= \frac{Support(Co\ and\ Fe)}{Support(Co)} = \frac{20}{46.67} \times 100 = 42.86$$

3) Lift

Lift rule is computed as: Lift$(x_1, x_2, ...)$ implies $(y_1, y_2, ...)$ = $\frac{Supp\ (x_1,x_2,...)and\ (y_1,y_2,...)}{Supp(x_1,x_{\ 2},...)\ Supp(y_1,y_2,...)}$. For example, $Lift(Co \rightarrow Fe) = \frac{P(Co/Fe)}{P(Fe)}$

Table VII shows the computed values of support, confidence and lift.

TABLE VII. MEASURES OF STRENGTH

Diseases	Rules	Support	Confidence	Lift
As, Co	As->Co	20.00	100.00	214.29
As, Di	As->Di	13.33	66.67	125.00
Co, Di	Co->Di	20.00	42.86	80.36
Co, Fe	Co->Fe	20.00	42.86	128.57
Co, He	Co->He	20.00	42.86	107.14
Co, Ki	Co->Ki	13.33	28.57	71.43
Di, He	Di->He	26.67	50.00	125.00
Di, Ki	Di->Ki	33.33	62.50	156.25
He, Ki	He->Ki	13.33	33.33	83.33
As, Co,Di	As, Co->Di	13.33	66.67	125.00
	As, Di->Co	13.33	100.00	214.29
	Co, Di->As	13.33	66.67	333.33
Co, Di, He	Co, Di->He	13.33	66.67	166.67
	Co, He->Di	13.33	66.67	125.00
	Di, He->Co	13.33	50.00	107.14
Co, Di, Ki	Co, Di->Ki	13.33	66.67	166.67
	Co, Ki->Di	13.33	100.00	187.50
	Di, Ki->Co	13.33	40.00	85.71
Di, He, Ki	Di, He->Ki	13.33	50.00	125.00
	Di, Ki->He	13.33	40.00	100.00
	He, Ki->Di	13.33	100.00	187.50

All of this can be expressed using the conditional probability and counting measure. Let Ω represent the set of records, and for each record ω, P(ω) is determined using the Eq. (4).

$$P(\omega) = \frac{1}{\Omega} = 1/n \qquad (4)$$

Define event E_A as the collection of records containing item set A, then Supp(A) = P(E_A) and Conf(A→B) = P(E_B| E_A). Therefore, Lift(Co→Fe) is > 1 if we know that It is more likely that a record will have a cough if a record also has fever.

The lift value is greater than 1 for the following rules based on market basket analysis.

Co (As → Co) (214.29)

Co, Di, As (Co, Di → As) (333.33)

Co, Ki, Di (Co, Ki→ Di) (187.50)

He, Ki, Di (He, Ki→ Di) (187.50)

Next the Chi-Square test was performed to test the significance of the rules. The hypotheses are listed below:

H_0: Generated pattern is accepted.

H_1: Generated pattern is not accepted.

Chi-Square calculated value is 13.83088 which is less than the tabulated value, and hence the null hypothesis H_0 accepted which states that generated pattern is accepted.

V. RESULTS

On the basis of the confidence and support, the following rules were generated:

- If a patient with diabetes and a heart condition has a covid problem, chances of his/her death is 75%, and chances of getting cured is 25%.

- If a patient has cough and fever with covid and previously have had additional diseases such as lung cancer, chances of his/her death is 67% and 34% chance is of being cured.
- If the patient is facing the asthma problem in addition to diabetes and cough then the chances of his/her death 100%.
- Patients with diabetes and kidney problems have 80% chance of dying and20% chance of surviving.
- If the patients have a cough and diabetes, as well as any of the other ailments such as kidney, heart, or asthma, they have a 100% chance of dying.
- If a patient has diabetes, cough, kidney, or heart problems, there is a 100% chance that he or she will die.

VI. Conclusion

Data mining plays very important role for Corona Patient data analysis. These interesting patterns gives important information about most likely patients and less likely corona patients. This information is of sheer importance for making policy for determining the diseases to cure the patient. Based on the analysis, it can be concluded that data mining tools can be utilized effectively for optimizing the patterns associated with the dynamic behaviors of the transactions made by customers when purchasing specific products. The market basket analysis algorithm is applied for association rule in data mining. Frequent customer transactions have been analyzed using the algorithm. Also, using support and confidence, we identified the consumers who purchase related products.

References

[1] Covid-19 Coronavirus Pandemic. [Online]. Available: https://www.worldometers.info/coronavirus/. [Accessed: 12 April, 2022].

[2] What is data mining?. [Online]. Available: https://www.ibm.com/topics/data-mining. [Accessed: 30 May, 2022].

[3] 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, 2020.

[4] A. S. Albahri, R. A. Hamid, J. K. Alwan, Z. T. Al-Qays, A. A. Zaidan, B. B. Zaidan, and H. T. Madhloom, “Role of biological data mining and machine learning techniques in detecting and diagnosing the novel coronavirus (COVID-19): a systematic review,” Journal of Medical Systems, vol. 44, pp. 1-11, 2020.

[5] 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, 2020.

[6] P. M. Buscema, F. D. Torre, M. Breda, G. Massini, and E. Grossi, “COVID-19 in Italy and extreme data mining,” Physica A, 2020, doi: 10.1016/j.physa.2020.124991.

[7] C. Huang, X. Xu, Y. Cai, ..., and L. Yang, “Mining the characteristics of COVID-19 patients in China: analysis of social media posts,” Journal of Medical Internet Research, vol. 22, no. 5, 2020.

[8] S. Kumar, “Monitoring novel corona virus (COVID-19) infections in India by cluster analysis,” Annals of Data Science, vol. 7, pp. 417-425, 2020.

[9] D. Li, H. Chaudhary, and Z. Zhang, “Modeling spatiotemporal pattern of depressive symptoms caused by COVID-19 using social media data mining,” International Journal of Environmental Research and Public Health, vol. 17, no. 14, 2020.

[10] J. Li, Q. Xu, R. Cuomo, V. Purushothaman, and T. Mackey, “Data mining and content analysis of the Chinese social media platform Weibo during the early COVID-19 outbreak: retrospective observational infoveillance study,” JMIR Public Health and Surveillance, vol. 6, no. 2, 2020.

[11] M. Marhl, V. Grubelnik, M. Magdič, and R. Markovič, “Diabetes and metabolic syndrome as risk factors for COVID-19,” Diabetes & Metabolic Syndrome, vol. 14, no. 4, pp. 671-677, 2020.

[12] B. Maram and A. R. Satish, “A framework for performance analysis on machine learning algorithms using Covid-19 dataset,” Advances in Mathematics: Scientific Journal, vol. 9, no. 10, pp. 8207-8215, 2020.

[13] 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(4), 2020.

[14] P. Radanliev, D. D. Roure, and R. Walton, “Data mining and analysis of scientific research data records on Covid-19 mortality, immunity, and vaccine development - In the first wave of the Covid-19 pandemic,” Diabetes & Metabolic Syndrome, vol. 14, no. 5, pp. 1121-1132, 2020.

[15] F. Stephany, N. Stoehr, P. Darius, L. Neuhäuser, O. Teutloff, and F. Braesemann, “The CoRisk-Index: A data-mining approach to identify industry-specific risk assessments related to COVID-19 in real-time,” arXiv preprint arXiv:2003.12432.

[16] K. Ghosh, S. A. Amin, S. Gayen, and T. Jha, “Chemical-informatics approach to COVID-19 drug discovery: Exploration of important fragments and data mining based prediction of some hits from natural origins as main protease (Mpro) inhibitors,” Journal of Molecular Structure, 2021.

[17] P. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining. Paperback – International Edition, Pearson Publication, 2005.