*Journal of Computer Science*

Original Research Paper

# Big Data Framework for Predicting Infectious Diseases to Improve Healthcare by Discovering New Symptom Patterns

**Amal Mohamed Mounir, Mohamed Ibrahim Marie and Laila Abd-Elhamid**

*Department of Information Systems, Faculty of Computers and Artificial Intelligence, Helwan University, Egypt*

*Article history*
Received: 20-02-2024
Revised: 16-03-2024
Accepted: 06-04-2024

Corresponding Author:
Amal Mohamed Mounir
Department of Information Systems, Faculty of Computers and Artificial Intelligence, Helwan University, Egypt
Email: amalmounir00@gmail.com

**Abstract:** The utilization of big data in infectious disease control represents a captivating opportunity, as these novel data streams offer the potential to enhance the timeliness of preventive measures. Various healthcare providers in both the public and private sectors generate, store, and analyse extensive datasets to enhance the quality of services they deliver. Recently, the outbreak of the new coronavirus, COVID-19, has posed significant threats to human health, life, production, social connections, and international relations, placing them in substantial peril. Consequently, the adoption of big data technologies has played a pivotal role in the response to the pandemic. Infectious diseases manifest when a person contracts a disease from a pathogen transmitted by another person, posing challenges that affect both individual and macroscales. Furthermore, the unknown patterns of infectious illnesses add complexity to the prediction process. This study aims to establish a big data framework for predicting infectious diseases by uncovering new patterns of symptoms, ultimately enhancing healthcare infection prevention and control. To achieve this objective, machine-learning algorithms such as K-Nearest Neighbors and Random Forest were employed for cleaning and maintaining extensive datasets collected from December 2019 to June 2020. Additionally, FP-growth and the Park, Chen, and Yu algorithms were applied to identify new patterns. The results demonstrated the superior performance of the Support Vector Machines (SVM) classifier, which achieved the highest accuracy of 98.2%. The Random Forest (RF) classifier had the highest precision (92.80%), and the SVM classifier had the highest F1 score (94.80%). Similarly, the Park, Chen, and Yu algorithm outperformed FP growth, achieving an accuracy rate of 98.5%. These findings underscore the potential of big data and machine learning in pattern recognition and predicting infectious diseases, ultimately contributing to improved public health outcomes.

**Keywords:** Big Data, Healthcare, Association Rule Mining, Random Forest, Infection Diseases, PCY Algorithm

## Introduction

The emergence of big data platforms and advancements in machine learning algorithms have enabled researchers to extract insights from massive amounts of data. This could potentially lead to the development of more effective approaches for the surveillance and control of infectious illnesses Anwar and Khan (2020).

The simplicity of storing, manipulating, and analyzing diverse data formats at large scale May be beneficial for health institutions and public health officials to promptly respond to and manage infectious disease outbreaks. Recently researchers have recognized the promises of big data in enhancing infectious disease monitoring and control.

Lee *et al*. (2022) examined the difficulties in using big data to comprehend the spatial distribution and transmission of infectious illnesses from a technological, practical, and ethical standpoint. Integration of multiple data streams gathered at various spatial scales is technically difficult due to the heterogeneity of data sets and data types in this sector; a larger usage of multilevel Bayesian statistical methodologies would help to solve this problem and the conceptual gap between traditional epidemiology and the world of big data.

It is possible to find new patterns and relationships by mining the passive data created by the Internet, mobile phones, satellites, and radio-frequency sensors, which are becoming more widely available. Since around 2001, the number of papers at the intersection of big data, digital epidemiology, and infectious diseases has increased almost exponentially, attesting to the fact that the area of infectious diseases research is not immune to the big data revolution.

Since around 2001, the number of researchers at the nexus of big data and infectious diseases has increased almost exponentially, attesting to the fact that the area of infectious diseases research is not immune to the big data revolution. For this special issue, this study introduces a smart Big Data framework to predict infectious diseases by finding novel symptom patterns to enhance healthcare's infection prevention and control.

To achieve this objective, the effectiveness of the performance of the machine learning algorithms like K-Nearest Neighbors (K-NN) and the Random Forest (RF) models was used to clean and maintain big data in addition to the mining model FP-growth and Park, Chen, and Yu of China (PCY) to discover new symptoms rules. This study focused on Covid-19 symptoms as a case study for a type of infectious disease.

## *Related Works*

Big data offers great potential in the field of communicable disease monitoring, as it can improve the timeliness and accuracy of information by leveraging new data sources. These sources also provide access to previously inaccessible populations. By utilizing these data sources, we can gather information about the effectiveness of vaccines and medications, as well as enhance disease surveillance efforts. The promises of these big-data flows must be weighed against caution, though.

Singh *et al*. (2020) provided a succinct history of disease monitoring, pointed out the flaws in active systems, and argued that symptomatic surveillance has to be strengthened and deepened using big data. Influenza is used as a case study, where a high volume of medical claims data gathered by large private sector data warehouses sheds light on the spread of pandemics with good spatial detail. The high volume of medical claims data is partly caused by privacy issues and restrictions on access to electronic health data in government and academia.

They also showed how cutting-edge digital monitoring technologies, like Google Flu Trends (Grein *et al*., 2020), could malfunction due to overfitting and quickly become obsolete. This issue becomes especially important after significant changes in disease dynamics, such as the advent of a new pandemic virus. Continuous testing against conventional surveillance systems acts as a buffer against these problems. To enhance the timeliness, accuracy, and depth of current surveillance indicators, "hybrid" systems that combine digital big data with traditional laboratory-based surveillance and electronic health data are most likely the way forward considering the rise and fall of systems based solely on digital search engine data.

Koppeschaar *et al*. (2017) discussed participatory surveillance, utilizing the European-influenza Reporting-System-Influenza net as an example. In an attempt to assist Sentinel physician-based systems in Europe that have become established, this monitoring system depends on volunteers signing up online to report their health on a weekly basis. The researchers recognize how a system like this can be utilized to track any urgent medical issues instantly.

Liu *et al*. (2018) Discovered that smartwatches might be used to detect COVID-19 pre-symptoms. They examined the physiological and activity data gathered from the COVID-19 infection cases' smartwatches. They concluded that by using a two-level warning system based on significant increases in resting heart rate relative to individual baseline, 63% of COVID-19 cases might be identified before symptoms manifest. Additionally, they discovered that employing wearable technology for activity tracking and health monitoring could aid in the early recognition of respiratory infections. Some investigations have concentrated on determining the medical traits and symptoms connected with positive COVID-19 cases since the symptoms of COVID-19 have not yet been thoroughly defined and because COVID-19 is a dynamic disease.

Buchy *et al*. (2021) based on identifying the symptoms associated with the positive results of the COVID-19 test and it was focused on a set of Healthcare Workers (HCWs). Initial examination was by phone and a COVID-19 PCR test was conducted on each HCW to record symptoms associated with each case. The study discovered that the most general symptoms of positive COVID-19 cases were fever, myalgia, and anosmia, while the negative cases essentially have no symptoms, or the symptoms are restricted to nasal congestion and sore throat. Because of the timely scanning, collection, analysis, and distribution of regional and local online resistance index reports, they offer Resistance Open, a novel online platform for monitoring bacterial resistance to antibiotics. This strategy is a natural progression from earlier attempts to monitor epidemics of infectious diseases on a worldwide scale through the curation and analysis of various online data sources.

Salathé *et al*. (2023) used data from individuals conducting online research or reporting their symptoms of post-exposure in health forums, on Twitter, or on Facebook to improve the identification of drug adverse events in the future. These data streams can be used to generate statistical mining of unstructured texts, which can considerably increase the timeliness of monitoring in this field and reveal correlations among adverse events and certain medications. The same justification might be given for tracking vaccine-related adverse events, which now depend on physicians' passive reporting. To track vaccine hesitancy and medicine uptake, it is crucial to extract publicly generated digital data for data on behavior and sentiment (Salathé *et al*., 2023).

Finding a balance between the cost of baseless notifications and the timeliness of social media information is a significant difficulty in this situation. Moreover, allegations of side effects can damage quickly and permanently the public's perception of life-saving medications or vaccines. Using hybrid systems that combine big-data streams with passive physician reports of adverse events will help protect the specificity and accuracy of the alerts, just like with disease monitoring.

Wu *et al*. (2020) consider the status of epidemiological modeling and determine whether it is comparable to particle physics in the 1970s. They contend that one way to meet the problem of catching up with this Postponed development is to use epidemiological modeling of non-health data, such as online search queries. They draw an Intriguing comparison between illness predictions. The authors also point out that it can be difficult to convey forecast uncertainty, although it is usually done in meteorology rather than in disease forecasts because everyone can grasp a 20% likelihood of rain but not a 20% risk of an epidemic. Finally, although the fundamental laws of physics govern weather forecasts, human behavior alterations can also affect an outbreak's dynamics and skew its associated digital footprints, potentially complicating disease forecasts.

Aiello *et al*. (2020) presented that systems of disease surveillance are essential to monitoring and preventing public health issues. The use, potential, risks, and ethics of Internet- and social media-based data collection for public health surveillance are discussed in this study. By incorporating digital surveillance into public health as well as existing applications, that could be enhanced with greater integration, validation, and clarification of the regulations pertaining to ethical considerations. Hybrid systems that combine traditional surveillance data with information from social media posts, crowdsourcing, and search queries are promising advances.

## Materials and Methods

### *Proposed Big Data Framework for Enhancing Health Care Infection and Control*

This study presents a big data framework to control infectious diseases. The main purpose is to manage and control the healthcare system by preventing the spread of infectious diseases. The proposed framework consists of four main phases, which are the data acquisition phase, data pre-processing phase, Association Rule Generation (ARG) Phase, and classification Phase. Depicted in (Fig. 1.)

### *Data Set*

In this study, the data set comprises various patients, both male and female, from various age groups. Different characteristics or symptoms are employed for analysis, forecasting, and pandemic detection were taken from https://github.com/beoutbreakprepared/nCoV2019. The selected seven attributes are chosen from among 31 attributes, as shown in (Table 1).

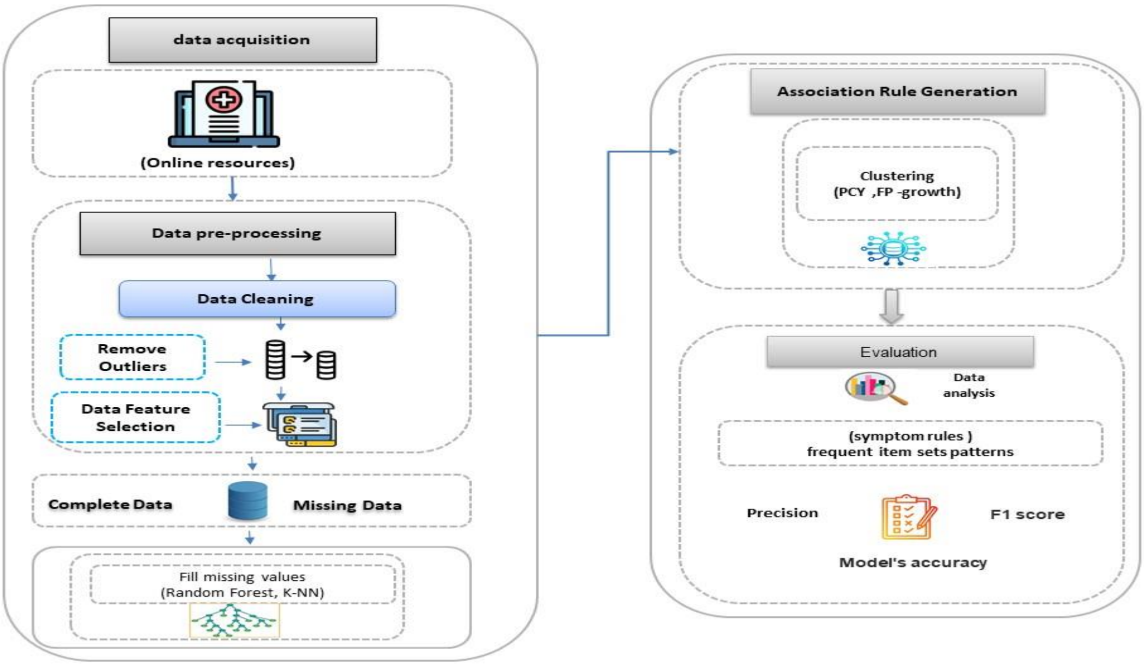


**Fig. 1:** Big data framework for enhancing healthcare infection and control

**Table 1:** Selected attribute

| S | Attribute | Description |
|---|---|---|
| 1 | ID | Identify the document for each patient |
| 2 | Age | Age of the patient |
| 3 | Gender | Male/female |
| 4 | Country | Name of the reported country |
| 5 | Date onset symptoms | the date of onset of the patient's infection |
| 6 | Symptoms | List of reported symptoms in the case description |
| 7 | Outcome | 0 the patient's discharge<br>1 the patient recovered<br>2 the patient died |

### Data Acquisition Phase (DAP)

Timely yet accurate Data Acquisition (DA) is required during health emergencies to report public health responses. Epidemiological data is necessary in case of emerging epidemics to observe and expect spread of infection. For example, novel coronavirus Cases were first recorded in Wuhan, Hubei province, China, in December 2019 and have since prevalence across the world. Epidemiological data is collected and organized individual-level data from national, data of province, and data of cities health reports, as well as additional information from online reports and official government sources (Xu, 2020). The structured data sets were collected from epidemiological data with a set size of 5.9 GB from the COVID-19 outbreak from December 2019 and 2020 and 143 countries' real-time case information.

### Data Pre-Processing Phase (DPP)

Data pre-processing is the main stage in the process of data mining by converting the raw data to accurate data. An important feature of DPP is the ability to enhance the accuracy of the proposed model. For that, this Phase starts by cleaning data, to ensure that the algorithm is only considering relevant data and not being influenced by any abnormal or incorrect data. In addition, the proposed model aims to minimize the time and resources desired to train the model by handling missing values or removing redundant data. In the proposed Phase, the major processes of DPP include data cleaning and handling missing values models. Data cleaning consists of two major steps: Remove outliers and feature selection. Outliers are data points that dramatically deviate from the other observations in a dataset. It could occur because of measurement inconsistency or incorrect data point filling. For example, the distribution like people's age 356 is not a valid age, while 45 is. Outliers can be discovered by using the Inter Quartile Range (IQR). *IQR* calculates the variation in the data set. Any value, that is above the range of -1.5-1.5× *IQR* is classified as outliers. Calculating the interquartile range takes the third quartile value and subtracts the first quartile value, as shown in the formula (Eq. 1):

$$IQR = Q3 - Q1 \tag{1}$$

The second step of data cleaning is feature selection, which is a selection of sample sets of relevant features. Where feature selection can be obtained by calculating the mutual information between all features as a score using $(f_i \in F)$ and the target class is $(c)$ then, the feature that will achieve the largest score is selected. The selection of the feature with the largest score can be calculated with Eqs. (2-3):

$$\underset{f_i \in F}{argmax}\left(I_{derived}\left(f_{i,c}\right)\right) \tag{2}$$

$$D(S, c) = \frac{1}{/S/}\sum_{f_i \in S} I\left(f_{i;}c\right) \tag{3}$$

The average value of all mutual information values between each individual feature $f_i$ and class $c$ determines the significance of a feature set $S$ for that class. Cleaning has two steps in handling missing values, where it is vital to fill in the missing data values in data sets. Classification tasks can be applied to fill the missing data values in the dataset. However, the simplest solution is to simply fill them with zero which will dramatically decrease the accuracy of the model. Using machine learning models, such as the Random Forest model or k-nearest neighbors model can support enhancing the accuracy of data.

### Handle Missing Values (HMV)

HMV is an imputation technique and it can be effectively scaled to handle substantial amounts of data. For instance, *RFM* can be employed for this purpose. Additionally, *RFM* can effectively manage the non-linearity of data and address outliers.

The Random Forest Model (RFM) contains contained mixed-variables of data (both numbers and categorical). An iterative imputation method supported by *RFM*. For instance, (Fig. 2) shows an example of how *RF* can handle missing values. *RFM* is arranged in three steps:

- Step 1(initialization): When the selected column has missing values, the missing values are filled with the variable's mean of the respective columns for continuous variables or its most frequent class (for categorical variables)
- Step 2 (imputation): The dataset is split into two groups: The training data, which is made up of the observed variables, and the prediction set which is made up of the missing observations. These training and prediction sets are fed to random forest and then the predicted value is imputed to the variable's missing portion. After imputing all the variables, one iteration is completed.
- Step 3 (stop): The second step is repeated until a halting condition is reached. The MF ran for three iterations and then stopped. The iterative stopping criterion was reached when the variance between the previously imputed values and the newly imputed data

increased for the first time with respect to categorical and numerical variable types. Multiple iterations enabled the algorithm to be trained on better quality data than it previously predicted (Su *et al*., 2020)

- When all the values with missing variables have been imputed, one imputation iteration is completed
- Step 4 (error evaluation): Calculation of the value of the Absolute Error using Eq. (4):

$$Absoluteerror = |actualvalue - measured\ value| \qquad (4)$$

### Association Rule Mining Phase (ARMP)

Association rule mining is a method for discovering interesting associations among items in big databases. It is a proposed scheme to establish strong rules discovered in datasets by some measures of interest (Shahin *et al*., 2021). In any specified transaction with an assortment of variables, association rules are intended to discover the rules that set how or why certain items are joint. To achieve this proposed goal, there are various types of algorithms available for generating association rules in data mining (Rasheed *et al*., 2020), including the FP-growth Algorithm and the PCY Algorithm.

Park-Chen-Yu (PCY) algorithm. Three researchers park, Chen, and Yu from China developed the PCY algorithm. This algorithm is used in the field of big data analysis for association rules mining when the dataset is very huge and numerous transactions (Zhang *et al*., 2023). The PCY algorithm contains two passes known as pass 1 and pass 2. By leveraging the idle memory from pass 1, a hash table is maintained to track the occurrences of all items in pass1, referred to as 1 Count. For each item set, consisting of items = [*i*_1…. *i*_*k*], hash all pairs to an item set of the hash table and increment the count of the item set by 1. The min sup is calculated by using Eq. (5):

$$Minsup(x) = (e\hat{}(ax - b)) + c \qquad (5)$$

where, *x* is the number of transactions in the dataset, *a*, *b*, and *c* are constants because it increases *minsup* when there is little data and decreases it when there is more data as shown in Fig. (3) (Zhang *et al*., 2023).

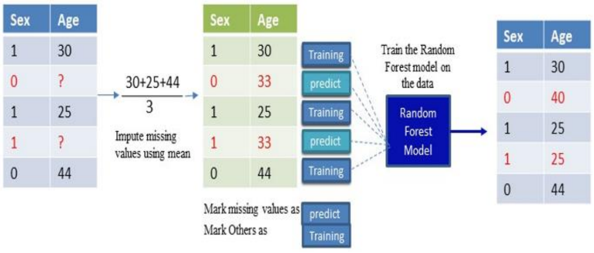


**Fig. 2:** Example of RF algorithm for handling missing values

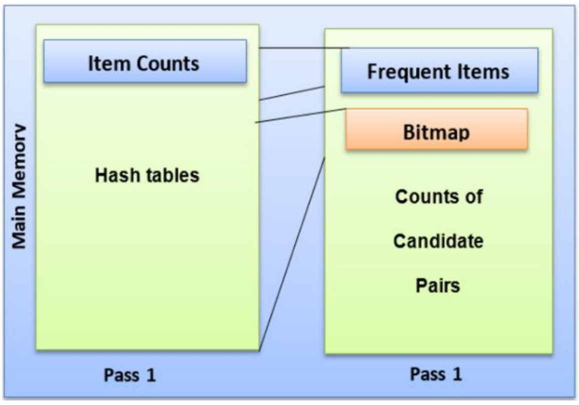


**Fig. 3:** Two passes of the PCY algorithm

For example, after the spread of COVID-19 disease, patients used search engines to search for symptoms of common infections via the Internet. This transaction is noted. If a patient is infected with COVID-19 from any other person with the symptoms, the main purpose of this algorithm is to discover frequent item sets, along with fever patients frequently infected coughs. The transactional dataset in (Table 2) contains eight transactions, with a Threshold value or minimization value equal to three, the letter *p* represents the patient and hash function.

So, from the above example Table 1, the Cough is the most frequent symptom with fever symptom, so, it is considered a frequent item. Suppose we assign cough = 1, fever = 2, malaise = 3, ARVI = 4 and dry mouth = 5. Then, Table 2 shows the data set after the change.

Using buckets and the technique of map-reduce for solving the problem:

- Candidates are mapped and the length of each pair is determined as shown in (Table 3)
- Use a hash function to locate the bucket number

Step 1: Map each element to determine its length, by calculating the frequency of each symptom as shown in (Table 4).

Step 2: Eliminate all elements with counts lower than One, but there is no count lower than one, so the candidate set is equal to {1, 2, 3, 4, 5}.

Step 3: Every candidate set has been mapped into pairs and the lengths of each pair have been calculated as shown in (Table 5).

Avoid using pairs that have previously been written. Listing all sets with lengths greater than the threshold value: {(1, 3) (2, 3) (1, 4) (1, 5) (4, 5) (2, 5) (3, 4) (3, 5)}.

Step 4: Use hashing functions. (The bucket number is provided). By calculating the hash function *I* multiply *J* mod ten as represented in (Table 6).

Step 5: In the last step, the candidate set has been prepared as shown in (Table 7).

**Table 2:** Example transaction dataset

| Transaction Id | Transaction symptom |
|---|---|
| P1 | Cough, Fever, Malaise |
| P2 | Cough, (ARVI), Dry mouth |
| P3 | Cough, Malaise |
| P4 | Fever, Dry mouth |
| P5 | Cough, (ARVI), Fever |
| P6 | Cough, Fever, Malaise, Dry mouth |
| P7 | Fever, Malaise, Dry mouth |
| P8 | Malaise, Fever |

**Table 3:** Transaction dataset after changing symptoms with numbers

| Transaction Id | Transaction symptom |
|---|---|
| P1 | 1, 2, 3 |
| P2 | 1, 4, 5 |
| P3 | 1, 3 |
| P4 | 2, 5 |
| p5 | 1, 3, 4 |
| P6 | 1, 2, 3, 5 |
| P7 | 2, 3, 5 |
| P8 | 2, 3 |

**Table 4:** The length of each element is mapped

| Symptom # | Key | Frequency |
|---|---|---|
| 1 | 1 | 5 |
| 2 | 2 | 5 |
| 3 | 3 | 5 |
| 4 | 4 | 2 |
| 5 | 5 | 4 |

**Table 5** Lengths of each pair

| Transaction Id | Transaction symptom | Lengths of each pair |
|---|---|---|
| P1 | {(1, 2) (1, 3) (2, 3)} | (2, 3, 3) |
| P2 | {(1, 4) (1, 5) (4, 5)} | (4, 5) |
| P3 | {(1, 3)} | 1 |
| P4 | {(2, 5)} | (2, 5) |
| p5 | {(1, 3) (1, 4) (3, 4)} | (3, 4) |
| P6 | {(1, 2) (1, 3) (1, 5) (2, 3) (2, 5) (3, 5)} | (3, 5) |
| P7 | {(2, 3) (2, 5)} | 1 |
| P8 | {(2, 3)} | 1 |

**Table 6:** The value of each pair after implementing the hash function

| Pairs | Hash function | Value |
|---|---|---|
| (1, 3) | (1*3) mod 10 | 3 |
| (2, 3) | (2*3) mod 10 | 6 |
| (1, 4) | (1*4) mod 10 | 4 |
| (1, 5) | (1*5) mod 10 | 5 |
| (4, 5) | (4*5) mod 10 | 0 |
| (2, 5) | (2*5) mod 10 | 0 |
| (3, 4) | (3*4) mod 10 | 2 |
| (3, 5) | (3*5) mod 10 | 5 |

**Table 7:** Bucket Number arranges the pairs

| Bucket no. | Pairs |
|---|---|
| 0 | (4, 5) |
| 0 | (2, 5) |
| 2 | (3, 4) |
| 3 | (1, 3) |
| 4 | (1, 4) |
| 5 | (1, 5) |
| 5 | (3, 5) |
| 6 | (2, 3) |

**Table 8:** The candidate set is arranged by HSF

| Candidate set | Pairs | Highest support frequency | Bucket no. | Bit vector |
|---|---|---|---|---|
| (4, 5) | (4, 5) | 3 | 0 | 1 |
| (2, 5) | (2, 5) | 4 | 0 | 1 |
| (3, 4) | (3, 4) | 3 | 2 | 1 |
| (1, 3) | (1, 3) | 3 | 3 | 1 |
| (1, 4) | (1, 4) | 5 | 4 | 1 |
| (1, 5) | (1, 5) | 3 | 5 | 1 |
| (3, 5) | (3, 5) | 3 | 5 | 1 |
| (2, 3) | (2, 3) | 5 | 6 | 1 |

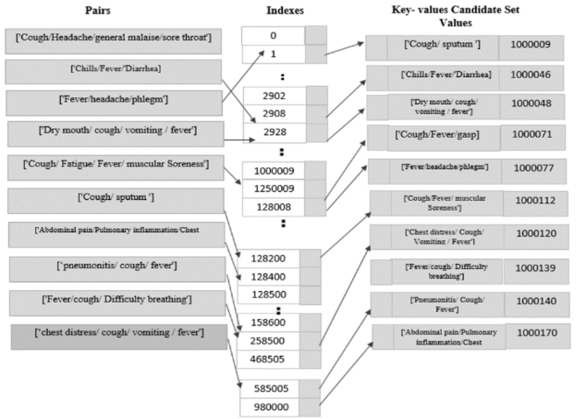

Pairs
Indexes
Key- values Candidate Set Values
['Cough/Headache/general malaise/sore throat']
['Chills/Fever/'Diarrhea]
['Fever/headache/phlegm']
['Dry mouth/ cough/ vomiting / fever']
['Cough/ Fatigue/ Fever/ muscular Soreness']
['Cough/ sputum ']
['Abdominal pain/Pulmonary inflammation/Chest
['pneumonitis/ cough/ fever']
['Fever/cough/ Difficulty breathing']
['chest distress/ cough/ vomiting / fever']
0
1
:
2902
2908
2928
:
1000009
1250009
128008
:
128200
128400
128500
:
158600
258500
468505
585005
980000
['Cough/ sputum ']
1000009
['Chills/Fever/'Diarrhea]
1000046
['Dry mouth/ cough/ vomiting / fever']
1000048
['Cough/Fever/gasp]
1000071
['Fever/headache/phlegm']
1000077
['Cough/Fever/ muscular Soreness']
1000112
['Chest distress/ Cough/ Vomiting / Fever']
1000120
['Fever/cough/ Difficulty breathing']
1000139
['Pneumonitis/ Cough/ Fever']
1000140
['Abdominal pain/Pulmonary inflammation/Chest
1000170

**Fig. 4:** The linked list represents a key-value pair-linked list using hashing function

The Highest Support Frequency (HSF) is the number of duplications of that vector. By reviewing pairs with a support frequency lower than three, the candidate set is rejected if its support is below three as represented in (Table 8).

Then, the pairs are arranged in ascending order of their acquired bucket number, as shown in (Fig. 4).

Calculating the confidence. Confidence of this association rule is the probability of $j$ given items $= [i_1; \dots . i_k]$, in the form *i j*, where *j* and *I* are a separate set of symptoms, i.e., *i j* =. *i* is known as the antecedent of the rule and *J* is known as the consequent. It is also referred to as "if-then," where "if" stands for the antecedent and "then" for the consequence (Kaushik *et al*., 2021). In general, Support, confidence, and lift are used to gauge how effective newly discovered rules are. Support can officially be described in Eq. (8):

$$support(I \rightarrow J) = support(\{I\} \cup \{J\}) \tag{6}$$

Therefore, the frequency (or generality) of a rule for a definition of confidence is calculated in Eq. (9):

$$Confidence(I \rightarrow J) = \frac{\Pr[I \cup j]}{\Pr[I]} = \frac{support(I \cup j)}{support(I)} \tag{7}$$

where, the percentage of patients who have $j$ to all illnesses is the fraction of patients who have $j$., Lift indicates the frequency of symptom $Y$ occurrence during symptom $i$ occurring while controlling the likelihood of symptom $j$ occurrence (Li and Sheu, 2021). The correlations among $j$ and $I$ are determined by the lift value, which can be independent (= 1), positively associated (>1), or negatively related (<1).

## Results

The proposed framework has been applied and tested. The performance of handling missing values was evaluated by comparing it with the RF model in the Machine Learning library (MLlib) and the K-NN model. MLlib is an Apache Spark machine-learning library that includes PF-growth and PCY algorithm implementation for data mining (Shahin *et al*., 2021) dataset.

### *Handle Missing Values Models*

### *The Support Vector Machine (SVM) Model*

The Support Vector Machine (SVM) uses a function $\phi$ to map the training data from the input space into a higher dimensional feature space and then build a separating hyperplane in the feature space with the maximum margin (Idri *et al*., 2018) (Idri A). The transformation should be chosen in a certain way so that their dot product leads to a kernel-style function by using Eq. (8):

$$K(x, x_i) = \phi(x).\,\phi(x_i) \tag{8}$$

Given a training set of data Eq. (9):

$$S = \{(x_1, y_1), (x_2, y_2) \ldots (x_n, y_n)\} \tag{9}$$

where, $x_i$ ϵ $Rd$ denotes an input vector and $y_i$ ϵ $R$ is its corresponding target value, the regression problem determines a function $f$ that can approximate targeted values accurately. $f(x)$ is given by Eq. (10):

$$f(X) = < w, \emptyset(X) > + b \tag{10}$$

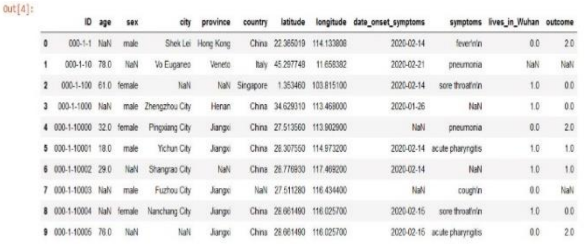
where, $w$ ϵ $R_d$, $b$ ϵ $R$, and $\phi$ denote a nonlinear transformation from $R_d$ to high-dimensional space.

ε-*SVR* aims to find a function $f(x)$ that has at most ε deviation from the actually obtained targets $y_i$ for the training data set and simultaneously is as flat as possible.

### *The Random Forest (RF) Model*

The Random Forest (RF) model is implemented in the Spark framework, which relies on Resilient Distributed Datasets (RDDs) for executing parallel tasks. NumPy and Pandas are imported to read in the mentioned COVID-19 dataset. To present the effectiveness of the imputers, a complete dataset without any missing values was taken and then the data at random was amputated and created missing values. Then the imputers are used to predict missing data and compare it to the original. The missing-PY library for Miss Forest Functions (MFF) and the mice forest library for mice forest were used for Implementation. The four techniques to impute data were used:

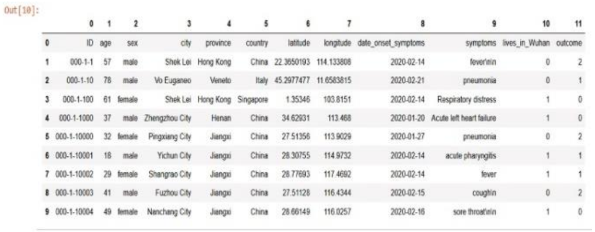
- Step 1: Six lists of unique random numbers ranging starting from number zero are made to the Covid-19 dataset's length. Using some pandas manipulation, the values of Age, Sex, date_onset_symptoms, lives_in_Wuhan, symptoms, and outcome are replaced with NaNs, based on the index positions generated at random, as shown in (Fig. 5)
- Step 2: Thus, the imputation is performed. The target variable is removed from the data and then missing values are now imputed
- Step 3: Imputed columns from the miss-forest algorithm are represented by creating a new data frame containing all columns in the original and imputed values as shown in (Fig. 6) that represent ten columns as a sample
- Step 4: The absolute errors also known as the approximation errors are calculated by the original's value (the actual value) and the imputed value (measured value), the absolute errors are small between the original's value and the imputed value as shown in (Fig. 7), as a sample of the absolute errors are calculated



Out[4]:

| | ID | age | sex | city | province | country | latitude | longitude | date_onset_symptoms | symptoms | lives_in_Wuhan | outcome |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 000-1-1 | NaN | male | Shek Lei | Hong Kong | China | 22.365019 | 114.133808 | 2020-02-14 | fever\n\n | 0.0 | 2.0 |
| 1 | 000-1-10 | 78.0 | NaN | Vo Euganeo | Veneto | Italy | 45.297748 | 11.658382 | 2020-02-21 | pneumonia | NaN | NaN |
| 2 | 000-1-100 | 61.0 | female | NaN | NaN | Singapore | 1.353460 | 103.815100 | 2020-02-14 | sore throat\n\n | 1.0 | 0.0 |
| 3 | 000-1-1000 | NaN | male | Zhengzhou City | Henan | China | 34.629310 | 113.468000 | 2020-01-26 | NaN | 1.0 | 0.0 |
| 4 | 000-1-10000 | 32.0 | female | Pingxiang City | Jiangxi | China | 27.513560 | 113.902900 | NaN | pneumonia | 0.0 | 2.0 |
| 5 | 000-1-10001 | 18.0 | male | Yichun City | Jiangxi | China | 28.307550 | 114.973200 | 2020-02-14 | acute pharyngitis | 1.0 | 1.0 |
| 6 | 000-1-10002 | 29.0 | NaN | Shangrao City | NaN | China | 28.776930 | 117.469200 | 2020-02-14 | NaN | 1.0 | 1.0 |
| 7 | 000-1-10003 | NaN | male | Fuzhou City | Jiangxi | NaN | 27.511280 | 116.434400 | NaN | cough\n | 0.0 | NaN |
| 8 | 000-1-10004 | NaN | female | Nanchang City | Jiangxi | China | 28.661490 | 116.025700 | 2020-02-15 | sore throat\n\n | 1.0 | 0.0 |
| 9 | 000-1-10005 | 76.0 | NaN | NaN | Jiangxi | China | 28.661490 | 116.025700 | 2020-02-15 | acute pharyngitis | 0.0 | 2.0 |

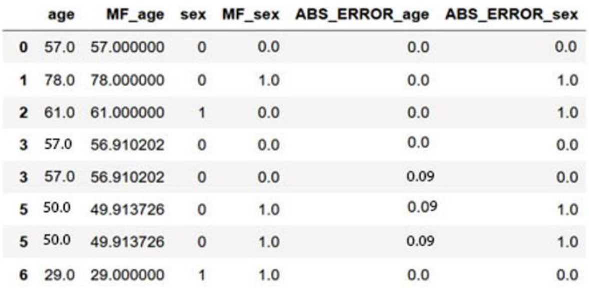
**Fig. 5:** The table represents the dataset with Nan's Values



Out[10]:

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | ID | age | sex | city | province | country | latitude | longitude | date_onset_symptoms | symptoms | lives_in_Wuhan | outcome |
| 1 | 000-1-1 | 57 | male | Shek Lei | Hong Kong | China | 22.3650193 | 114.133808 | 2020-02-14 | fever\n\n | 0 | 2 |
| 2 | 000-1-10 | 78 | male | Vo Euganeo | Veneto | Italy | 45.2977477 | 11.6583815 | 2020-02-21 | pneumonia | 0 | 1 |
| 3 | 000-1-100 | 61 | female | Shek Lei | Hong Kong | Singapore | 1.35346 | 103.8151 | 2020-02-14 | Respiratory distress | 1 | 0 |
| 4 | 000-1-1000 | 37 | male | Zhengzhou City | Henan | China | 34.62931 | 113.468 | 2020-01-20 | Acute left heart failure | 1 | 0 |
| 5 | 000-1-10000 | 32 | female | Pingxiang City | Jiangxi | China | 27.51356 | 113.9029 | 2020-01-27 | pneumonia | 0 | 2 |
| 6 | 000-1-10001 | 18 | male | Yichun City | Jiangxi | China | 28.30755 | 114.9732 | 2020-02-14 | acute pharyngitis | 1 | 1 |
| 7 | 000-1-10002 | 29 | female | Shangrao City | Jiangxi | China | 28.77693 | 117.4692 | 2020-02-14 | fever | 1 | 1 |
| 8 | 000-1-10003 | 41 | male | Fuzhou City | Jiangxi | China | 27.51128 | 116.4344 | 2020-02-15 | cough\n | 0 | 2 |
| 9 | 000-1-10004 | 49 | female | Nanchang City | Jiangxi | China | 28.66149 | 116.0257 | 2020-02-16 | sore throat\n\n | 1 | 0 |

**Fig. 6:** The table represents the dataset of the original addition to imputed values



| | age | MF_age | sex | MF_sex | ABS_ERROR_age | ABS_ERROR_sex |
|---|---|---|---|---|---|---|
| 0 | 57.0 | 57.000000 | 0 | 0.0 | 0.0 | 0.0 |
| 1 | 78.0 | 78.000000 | 0 | 1.0 | 0.0 | 1.0 |
| 2 | 61.0 | 61.000000 | 1 | 0.0 | 0.0 | 1.0 |
| 3 | 57.0 | 56.910202 | 0 | 0.0 | 0.0 | 0.0 |
| 3 | 57.0 | 56.910202 | 0 | 0.0 | 0.09 | 0.0 |
| 5 | 50.0 | 49.913726 | 0 | 1.0 | 0.09 | 1.0 |
| 5 | 50.0 | 49.913726 | 0 | 1.0 | 0.09 | 1.0 |
| 6 | 29.0 | 29.000000 | 1 | 1.0 | 0.0 | 0.0 |

**Fig. 7:** The table represents the dataset of the original and imputed values after absolute errors are calculated

### *Comparing Results between Support Vector Machine (SVM), Random Forest (RF), and (K-NN) Models*

The percentages of precision results of handling missing values of attributes are greater than by using the Random Forest (RF) model. The percentage accuracy result of handling missing values of attributes is greater than by using the Support Vector Machine (SVM) model as shown in Fig. 8. The F1-score result of handling missing values of attributes is greater than by using the Support Vector Machine model, The percentages accuracy result of handling missing values of attributes are smaller than by using the K-Nearest Neighbours (KNN) model as shown in (Tables 9-11). The accuracy of the backtesting is calculated using a Formula (11) that incorporates information from the current dataset:

$$Accuracy \ = \frac{\sum(True\ Positive)+\sum(True\ Negative)}{\sum(True\ Positive)+\sum(True\ Negative)\ \sum(F\ alse\ Positive)+\sum(False\ Negative)} \qquad (11)$$

$$Precision \ = \frac{\sum(True\ Positive)}{\sum(True\ Positive)\sum(False\ Positive)} \qquad (12)$$

The formula used to determine the *F1 score* is as follows:

$$F1 - score = \frac{2*\ Recall*\ Precision}{Recall + Precision} \qquad (13)$$

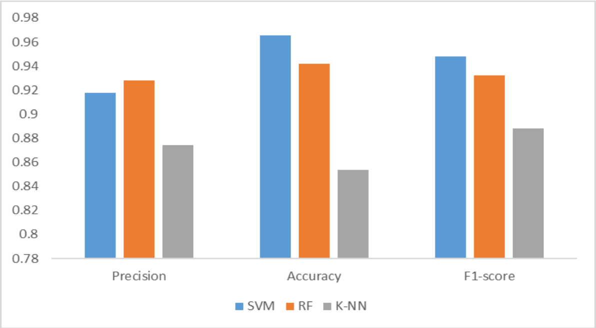


**Fig. 8:** The precision, accuracy, and F1-score comparison of ML models in the outcome of different attributes

**Table 9:** Outcome of five different attributes using the Support Vector Machine (SVM)

| Attribute | Precision | Accuracy | F1-score |
|---|---|---|---|
| Age | 0.92 | 0.97 | 0.951 |
| Gender | 0.93 | 0.976 | 0.942 |
| Country | 0.91 | 0.987 | 0.921 |
| Symptoms | 0.93 | 0.981 | 0.953 |
| Outcome | 0.9 | 0.985 | 0.932 |

**Table 10:** Outcome of five different attributes using random forest

| Attribute | Precision | Accuracy | F1-score |
|---|---|---|---|
| Age | 0.94 | 0.95 | 0.94 |
| Gender | 0.93 | 0.93 | 0.92 |
| Country | 0.91 | 0.92 | 0.91 |
| Symptoms | 0.92 | 0.97 | 0.96 |
| Outcome | 0.94 | 0.94 | 0.93 |

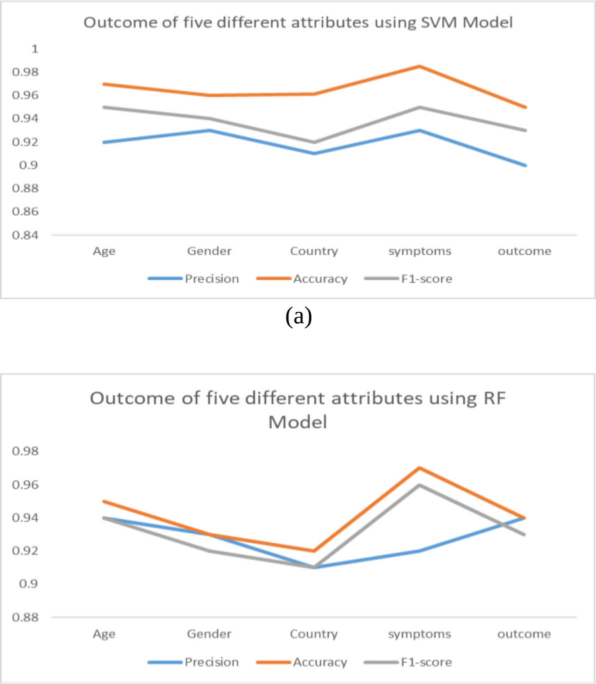
By comparing the performance of all the machine learning models when a different number of features are selected. We observed that all models produce maximum accuracy when the top 5 features are selected for the performance of SVM, RF, and K-NN classifiers, respectively as shown in Table 12. Support Vector Machines (SVM) have been reported to successfully handle missing values, particularly in datasets including different types of variables as shown in (Fig. 9). Overall, Support Vector Machines (SVM) had the smallest NRMSE (mean = 0.30) compared to Random Forest had NRMSE (mean = 0.35) and K-NN (mean = 0.39). In addition, the accuracy of Support Vector Machines (SVM) Random Forest (RF) is higher than the accuracy of K-NN.

**Table 11:** Outcome of five different attributes using K-NN

| Attribute | Precision | Accuracy | F1-score |
|---|---|---|---|
| Age | 0.90 | 0.88 | 0.89 |
| Gender | 0.89 | 0.87 | 0.88 |
| Country | 0.86 | 0.80 | 0.91 |
| Symptoms | 0.89 | 0.89 | 0.91 |
| Outcome | 0.83 | 0.83 | 0.85 |

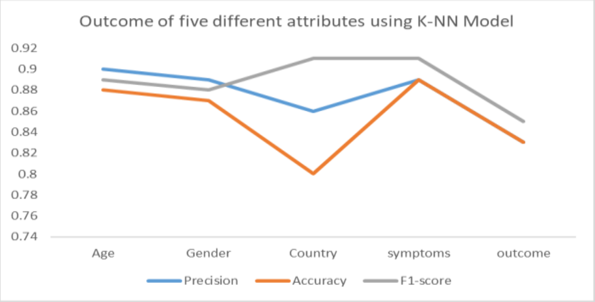
**Table 12:** Performance of ML models

| Model | Precision | Accuracy | F1-score |
|---|---|---|---|
| SVM | 0.918 | 0.9652 | 0.948 |
| RF | 0.928 | 0.942 | 0.932 |
| K-NN | 0.874 | 0.854 | 0.888 |



(a)

(b)



(c)

**Fig. 9:** Accuracy comparison of a different number of features are considered: (a) SVM, (b) RF, (c) K-NN

## *Data Association Rule Mining (ARM)*

This phase is based on Apache Spark data frame preprocessing. In the implementation of the PCY algorithm, Spark RDD is used. The algorithms were executed and run by the Anaconda navigator using spark. All 3,048,576 patients had their information retrieved and 1783 of them had frequency patterns of symptoms. The top 10 symptom rules are ordered by minimum support threshold ratings, so the case of symptoms in real life must be encoded.

## *Frequency Patterns According to Gender*

Important symptom rules broken down by males as shown in (Table 13) and important symptom rules broken down by females as shown (Table 14), it is clear that the number of produced frequent patterns in the gender of the male is larger than the gender of the female dataset. Therefore, the algorithms were executed on various values of minimum support thresholds from 10-100%.

Frequent patterns generated for females were 6001 rules and frequent patterns were produced for males 6060 rules, so a variance between male and female symptom rules was seen. The most common diagnoses in men were Smell and taste disturbance, Dry cough, Fever, and Headache. In women, the most common diagnoses were cough, rhinorrhoea, sore throat, pneumonia malaise/body aches, and weakness.

## *Frequency Patterns According to Age*

When reported symptoms. The median age of patients was 52 years (SD ±31.5 years; IQR 66 years), where 57% of the patients were male and 43% of the patients were female.

The age group between 20-45 years recorded 13319 frequent symptoms, where the greatest number of symptoms was a dry mouth and fever with minimum support thresholds equal to 10% as shown in (Table 15).

The age group between 45 and 65 years recorded 8056 frequent symptoms, where the greatest number of symptoms were breathing problems, coughs, and weakness with minimum support thresholds equal to 20%. As shown in (Table 16 ). The age group more than 65 years recorded 8023 frequent symptoms, where the greatest number of symptoms were breathing problems; nausea, and cough with minimum support thresholds equal to 20%.

The group of discharged recorded 6522 Frequent symptoms, where common symptoms were mild cough, fever, and weakness with minimum support thresholds equal to 20% as shown in (Table 17).

**Table 13:** Top 10 important symptom rules broken down by gender (male)

| Rules | Candidate Set | Mini support threshold % | Generated frequent rules |
|---|---|---|---|
| Rule 1 | {Malaise/Body Soreness /Dry mouth} | 10 | 1015 |
| Rule 2 | {Cough/ Headache/ General malaise} | 20 | 890 |
| Rule 3 | {Cough/Fever/Headache/ Malaise} | 30 | 860 |
| Rule 4 | {Anorexia/Fever} | 40 | 610 |
| Rule 5 | {Cough/General malaise/ Joint muscle pain} | 50 | 590 |
| Rule 6 | {/Dry Cough /Fever/ Breathing difficulty} | 60 | 560 |
| Rule 7 | {Headache/ Malaise/ body soreness,/(ARVI)} | 70<br>80 | 495<br>405 |
| Rule 8 | {Fever/Cough/ Headache} | 90 | 330 |
| Rule 9 | {(ARVI) / fever} | | |
| Rule 10 | {Chest distress/ Fever/ Weak/ Dyspnea} | 100 | 305 |

**Table 14:** Top 10 important symptom rules broken down by gender (female)

| Rules | Candidate set | Mini support threshold % | Generated frequent rules |
|---|---|---|---|
| Rule 1 | {Cough/Fever/(ARVI)} | 10 | 980 |
| Rule 2 | {Cough/ General malaise/ Joint muscle pain} | 20 | 910 |
| Rule 3 | {Sore throat/ Headache/ Tiredness} | 30<br>40 | 840<br>710 |
| Rule 4 | {Dry mouth/ Fever/myalgia} | 50 | 601 |
| Rule 5 | {Fever/ Cough/ Vomiting} | | |
| Rule 6 | {Aching muscles/ Fever/ pneumonia} | 60 | 540 |
| Rule 7 | {Fever/cough/ 'aggressive pulmonary} | 70 | 440 |
| Rule 8 | {Chills/Conjunctivitis/ Cough/ Fever} | 80 | 380 |
| Rule 9 | {Cough/Malaise/Body soreness, Sputum} | 90 | 310 |
| Rule 10 | {Chest distress/ Cough/ Fever/ Gasp} | 100 | 290 |

**Table 15:** Top 10 important symptom rules broken down by age (20–45 years)

| Rules | Candidate set | Minimum support threshold % | Generated frequent rules |
|---|---|---|---|
| Rule 1 | {Dry mouth, Fever} | 10 | 2654 |
| Rule 2 | {Dry mouth, Sore throat} | 20 | 2444 |
| Rule 3 | {Fever, Pneumonia, Sore throat} | 30 | 2115 |
| Rule 4 | {Cough, Malaise/body soreness, Sputum | 40 | 1982 |
| Rule 5 | {Fever/ Cough/ Vomiting} | 50 | 1230 |
| Rule 6 | {Diarrhea / Sore throat} | 60 | 924 |
| Rule 7 | {Fever/cough/ 'aggressive pulmonary} | 70 | 688 |
| Rule 8 | {Nausea/Nonrespiratory symptoms} | 80 | 561 |
| Rule 9 | {Cough/Malaise/Body soreness, Sputum} | 90 | 411 |
| Rule 10 | {Dry mouth/Weakness} | 100 | 310 |

**Table 16:** Top 10 important symptom rules broken down by age (45–65 years)

| Rules | Candidate set | Mini support threshold % | Generated frequent rules |
|---|---|---|---|
| Rule 1 | {Breathing problem/ Cough/Weakness} | 10 | 1502 |
| Rule 2 | {Weakness/ Nausea,} | 20 | 1244 |
| Rule 3 | {Myocardial infraction/ Cough} | 30 | 1015 |
| Rule 4 | {Cough/Malaise/body soreness/ Sputum | 40 | 982 |
| Rule 5 | {Dry mouth /Pneumonia} | 50 | 940 |
| Rule 6 | {Diarrhea / Sore throat} | 60 | 634 |
| Rule 7 | {Cough/ Sore throat/ Sputum} | 70 | 548 |
| Rule 8 | {Nausea/Nonrespiratory symptoms} | 80 | 480 |
| Rule 9 | {Cough/Malaise/Body Soreness} | 90 | 390 |
| Rule 10 | {Heart failure /Cough} | 100 | 321 |

**Table 17:** Top 10 important symptoms rules broken down by Status (Discharge)

| Rules | Candidate set | Mini support threshold % | Generated frequent rules |
|---|---|---|---|
| Rule 1 | {Mild Cough/ Fever} | 10 | 982 |
| Rule 2 | {Mild cough / Fever/ Weakness} | 20 | 892 |
| Rule 3 | {Cough/ Fever/ Runny nose} | 30 | 849 |
| Rule 4 | {Cough/ Fever/ Sputum} | 40 | 788 |
| Rule 5 | {Cough/Fever/ Gasp} | 50 | 686 |
| Rule 6 | {Cough/ Dizziness/ Fever} | 60 | 625 |
| Rule 7 | {Aching muscles/ Fever/ Pneumonia} | 70 | 570 |
| Rule 8 | {Fever / Aggressive pulmonary symptomatology} | 80 | 450 |
| Rule 9 | {Cough/Malaise/Body Soreness} | 90 | 370 |
| Rule 10 | {Fever /Weakness} | 100 | 310 |

Table 18 shows the group of recovered recorded 8397, where the common symptoms were malaise and body soreness in cases of recovered.

The group of deaths recorded 1158 frequent symptoms, where common symptoms were acute left heart failure and acute coronary syndrome, fever, and difficulty breathing. The results of all common tables represent fatigue by 0.713 (71.3%) followed by fever at 0.512 (51.2%), Smell and taste disturbance 0.424 (42.4%), mild cough at 0.341 (34.1%), Dry cough 0. 262 (26.2%), Muscle or joint pain 0.252 (25.2%) percentage, Headache 0. 232 (23.2%), Difficulty breathing 0.212 (21.2%), Sore throat 0. 192 (19.2%), malaise 0.184 (18.4%) of patients. Similar to this, the clinical characteristics and prognosis of the disease are strongly influenced by the patient's age. Similar trends were seen in our study's age-wise distribution of symptom patterns for individuals under the age of 45, as shown in Fig. 11. The other common symptoms in the pattern were body soreness, cough, dry mouth, Malaise, body soreness, sputum production, rhinorrhea and Aggressive pulmonary symptomatology which are included with the symptoms Among COVID-19 hospitalized patients who are young and middle-aged. Patients between the ages of 45 and 65 have problems with breathing symptoms, along with Sputum, cough, and fever in patients more than 65 years of age, the symptom patterns are more often breathing problems followed by Nausea and other symptoms such as Anorexia, fever, body soreness, and Sore throat. Overall, our study's findings are consistent with those reported in the literature; younger persons are more likely than older adults to experience symptoms linked to the ear, nose, and throat.

### *Comparison of the Rule Mining Algorithms in Real Time*

The performance of the most advanced rule-based algorithms across the COVID-19 symptoms data sets was compared as shown in (Fig. 11). To ensure an impartial evaluation, identical criteria are employed for assessing support and trust in both FP and PCY expansion techniques by conducting a comparison of running time on both techniques. To do that, we raised the number of transactions to ensure that the database was large enough. PCY and the FP-growth algorithms in the same setting, utilizing the MLlib package. Running the association rules using PCY is quicker than FP-growth, as demonstrated in Fig. 11. The investigation mentioned below showed that PCY offered the best setting for using the Covid-19 dataset. The algorithms were executed using the domain of minimum support with values starting at 10-20%. (Table 10) shows the number of produced frequent patterns for every Mini support value. (Fig. 10) represents the average running times by seconds for the PCY and FP-Growth algorithms by the thresholds. Each run was executed ten times to obtain a valid result for the running times and the average was calculated. For most thresholds. PCY performed better than FP-Growth. On average, PCY was 7.21% faster than FP-Growth. Modern information technologies are being used in the healthcare industry to address problems with global health such as unequal access to medical care, an increase in chronic illnesses, and rising medical expenses. Big data methods can be used to explain COVID-19 in terms of outbreak monitoring, viral development, disease prevention, and vaccine production. Health authorities can carefully manage and monitor the infectious disease.

**Table 18:** Top 10 important symptom rules broken down by Status (Recovered)

| Rules | Candidate set | Mini support threshold % | Generated frequent rules |
|---|---|---|---|
| Rule 1 | {Malaise/body soreness} | 10 | 1650 |
| Rule 2 | {Fever, Sore throat} | 20 | 1071 |
| Rule 3 | {Sputum/ Cough/Fever} | 30 | 987 |
| Rule 4 | {Fever, Weakness} | 40 | 901 |
| Rule 5 | {problem of Breathing/ Fever} | 50 | 887 |
| Rule 6 | {Aching muscles/ Fever/ Pneumonia} | 60 | 798 |
| Rule 7 | {Fever, Headache} | 70 | 653 |
| Rule 8 | {Chest pain/ Nasal congestion} | 80 | 550 |
| Rule 9 | {Dry mouth/ Dyspnea} | 90 | 475 |
| Rule10 | {Cough/ Dyspnea/ Fever} | 100 | 425 |

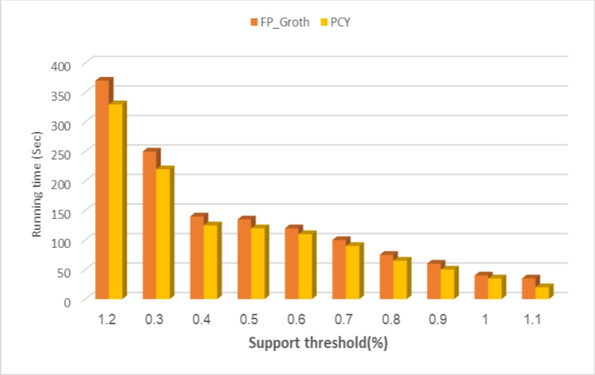


**Fig. 10:** The running times of PCY and FP-Growth algorithms

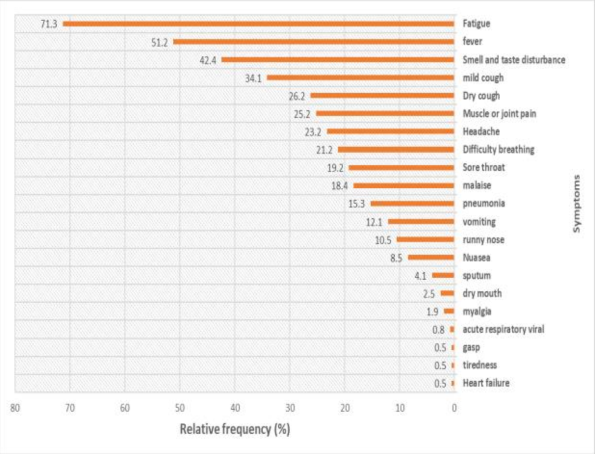


**Fig. 11:** Relative frequency of symptoms in COVID-19 patients

## Conclusion

Infectious disease, the single biggest hazard to life, was the primary cause of this mortality. Many of those who survived to adulthood died of infectious diseases, either directly or indirectly, but trauma took on greater importance. Each infectious disease has distinct symptoms that are unique to it. General warning signals and symptoms that several infectious illnesses share. Processing big data is therefore one of the biggest challenges facing users of this data. Therefore, a framework was created to address missing values in big data, where the data sets were manipulated using the K-NN model and random forest model. The results show that the Support Vector Machines (SVM) classifier achieved the highest accuracy of 98.2%. The Random Forest (RF) classifier had the highest precision (92.80%) and the SVM classifier had the highest F1-Score (94.80%). A new framework has been introduced to improve health care, like the use of big data for infectious disease control and prevention. This study offers a carefully optimistic view by using big data for infectious disease control and prevention, where new candidate sets of symptoms are discovered. The most common symptoms in our study encompassed malaise and body soreness, dry mouth, fever, chest distress, and breathing problems. Additionally, anorexia, sore throat, and aggressive pulmonary symptomatology are among COVID-19 patients' symptoms. For most thresholds. PCY performed better than FP-Growth. On average, PCY was 8.45% faster than FP-Growth. Moreover, PCY in environments of distribution processing, similar to most mining algorithms, needs a large amount of data to be transmitted over the network. Therefore, bandwidth limitation is one of the major problems for the PCY algorithm, especially in this epoch of big data. In the future, improvements could be applied to the PCY algorithm by enhancing the distribution of hashes, which will improve the efficiency of the algorithm and reduce the execution time.

## Acknowledgment

We would like to express our sincere gratitude to Helwan University and professors for their valuable guidance. The acknowledgment goes to the reviewers for their informative comments that helped us improve the quality of our manuscript. We also highly appreciate the efforts and the great support of the editorial team throughout the publication process.

## Funding Information

This research received no specific grant from any funding agency in the public, commercial, or not-for profit sectors.

## Author's Contributions

**Amal Mohamed Mounir:** Conceived ideas, collected papers from different sources, designed the outline of the manuscript and wrote the first drafted of the paper.

**Mohamed Ibrahim Marie:** Supervised and reviewed the manuscript and provided critical feedback.

**Laila Abd-Elhamid:** Supervising and reviewing designed the research plan and organized the study.

## Ethics

The authors declare no conflicts of interest. Any Potential ethical issues arising from this publication will be addressed promptly and transparently in accordance with the guidelines of the journal of computer science.

## References

Aiello, A. E., Renson, A., & Zivich, P. N. (2020). Social Media and Internet-Based Disease Surveillance for Public Health. *Annual Review of Public Health*, *41*(1), 101–118. https://doi.org/10.1146/annurev-publhealth-040119-094402

Anwar, H., & Khan, Q. U. (2020). Pathology and therapeutics of COVID-19: a review. *International Journal of Medical Students*, *8*(2), 113–120.

Bauchner, H., & Fontanarosa, P. B. (2020). Randomized Clinical Trials and COVID-19. *JAMA*, *323*(22), 2262–2263. https://doi.org/10.1001/jama.2020.8115

Buchy, P., Buisson, Y., Cintra, O., Dwyer, D. E., Nissen, M., Ortiz de Lejarazu, R., & Petersen, E. (2021). COVID-19 pandemic: lessons learned from more than a century of pandemics and current vaccine development for pandemic control. *International Journal of Infectious Diseases*, *112*, 300–317. https://doi.org/10.1016/j.ijid.2021.09.045

Idri, A., Abnane, I., & Abran, A. (2018). Support vector regression-based imputation in analogy-based software development effort estimation. *Journal of Software: Evolution and Process*, *30*(12), e2114. https://doi.org/10.1002/smr.2114

Grein, J., Ohmagari, N., Shin, D., Diaz, G., Asperges, E., Castagna, A., Feldt, T., Green, G., Green, M. L., Lescure, F.-X., Nicastri, E., Oda, R., Yo, K., Quiros-Roldan, E., Studemeister, A., Redinski, J., Ahmed, S., Bernett, J., Chelliah, D., … Flanigan, T. (2020). Compassionate Use of Remdesivir for Patients with Severe Covid-19. *New England Journal of Medicine*, *382*(24), 2327–2336. https://doi.org/10.1056/nejmoa2007016

Kaushik, M., Sharma, R., Peious, S. A., Shahin, M., Yahia, S. B., & Draheim, D. (2021). A Systematic Assessment of Numerical Association Rule Mining Methods. *SN Computer Science*, *2*(5), 348. https://doi.org/10.1007/s42979-021-00725-2

Koppeschaar, C. E., Colizza, V., Guerrisi, C., Turbelin, C., Duggan, J., Edmunds, W. J., Kjelsø, C., Mexia, R., Moreno, Y., Meloni, S., Paolotti, D., Perrotta, D., van Straten, E., & Franco, A. O. (2017). Influenzanet: Citizens Among 10 Countries Collaborating to Monitor Influenza in Europe. *JMIR public health and surveillance, 3(3), e66.* https://doi.org/10.2196/publichealth.7429

Lee, E. C., Arab, A., Colizza, V., & Bansal, S. (2022). Spatial aggregation choice in the era of digital and administrative surveillance data. *PLOS Digital Health*, *1*(6), e0000039. https://doi.org/10.1371/journal.pdig.0000039

Li, H., & Sheu, P. (2021). A scalable association rule learning heuristic for large datasets. *Journal of Big Data*, *8*(1), 86. https://doi.org/10.1186/s40537-021-00473-3

Liu, L., De Vel, O., Han, Q.-L., Zhang, J., & Xiang, Y. (2018). Detecting and Preventing Cyber Insider Threats: A Survey. *IEEE Communications Surveys & Tutorials*, *20*(2), 1397–1417. https://doi.org/10.1109/comst.2018.2800740

Salathé, Müller, M., M., & Kummervold, P. E. (2023). COVID-Twitter-BERT: A natural language processing model to analyse COVID-19 content on Twitter. *Frontiers in Artificial Intelligence*, *6*, 1023281. https://doi.org/10.3389/frai.2023.1023281

Rasheed, J., Jamil, A., Hameed, A. A., Aftab, U., Aftab, J., Shah, S. A., & Draheim, D. (2020). A survey on artificial intelligence approaches in supporting frontline workers and decision makers for the COVID-19 pandemic. *Chaos, Solitons & Fractals*, *141*, 110337. https://doi.org/10.1016/j.chaos.2020.110337

Shahin, M., Arakkal Peious, S., Sharma, R., Kaushik, M., Ben Yahia, S., Shah, S. A., & Draheim, D. (2021). Big Data Analytics in Association Rule Mining: A Systematic Literature Review. *2021 the 3rd International Conference on Big Data Engineering and Technology (BDET)*, 40–49. https://doi.org/10.1145/3474944.3474951

Singh, L., Bansal, S., Bode, L., Budak, C., Chi, G., Kawintiranon, K., & Wang, Y. (2020). A first look at COVID-19 information and misinformation sharing on Twitter. *ArXiv*, arXiv: 2003.13907.

Su, X., Xu, Y., Tan, Z., Wang, X., Yang, P., Su, Y., Jiang, Y., Qin, S., & Shang, L. (2020). Prediction for cardiovascular diseases based on laboratory data: An analysis of random forest model. *Journal of Clinical Laboratory Analysis*, *34*(9), e23421. https://doi.org/10.1002/jcla.23421

Wu, J., Li, J., Zhu, G., Zhang, Y., Bi, Z., Yu, Y., Huang, B., Fu, S., Tan, Y., Sun, J., & Li, X. (2020). Clinical Features of Maintenance Hemodialysis Patients with 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China. *Clinical Journal of the American Society of Nephrology*, *15*(8), 1139–1145. https://doi.org/10.2215/cjn.04160320

Xu, B. (2020). Epidemiological data from the COVID-19 outbreak, real-time case information. *Scientific Data*, *7*(1), 106.

Zhang, Y., Hong, J., & Chen, S. (2023). Medical Big Data and Artificial Intelligence for Healthcare. *Applied Sciences*, *13*(6), 3745. https://doi.org/10.3390/app13063745