JOIV : Int. J. Inform. Visualization, 5(2) - June 2021 162-169

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Data Mining Techniques for Pandemic Outbreak in Healthcare

Nur Izyan Suraya Abdul Satar[a], Azlinah Mohamed[b,*], Azliza Mohd Ali[c]

[a]*Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM), Shah Alam, Malaysia*
[b]*Institute for Big Data Analytics and Artificial Intelligence, Universiti Teknologi MARA (UiTM), Shah Alam, Malaysia*
[c]*Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM), Shah Alam, Malaysia*
*Corresponding author: *azlinah@uitm.edu.my*

***Abstract*— Pandemic outbreaks such as SARS-CoV, MERS-CoV and Covid-19 have attracted worldwide attention since these viruses have affected many countries and become a global public health issue. In 2019, Covid-19 was announced as a pandemic disease and categorized as a public health emergency globally. It is ranked as the sixth most serious pandemic internationally. This pandemic tracking and analysis require an appropriate method that gives better performance in terms of accuracy, precision and recall that defines its pattern since it involves huge and complicated datasets from the pandemic. Pattern identification is currently applied in many instances due to the rapid growth of data besides having the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions and identify the relationships between data items. Therefore, there is a need to review the techniques in data mining on the pandemic outbreak that focuses on healthcare. The goal of this study was to analyze the algorithms from the data mining method that had been implemented for pandemic outbreaks in past research such as SARS-CoV, MERS-CoV and Covid-19. The result shows that 2 main algorithms, namely Naïve Bayes and Decision Tree, from the classification method, are appropriate algorithms and give more than 90% accuracy in both the pandemic and healthcare. This will be further considered and investigated for future analysis on large datasets of Covid-19 which can help researchers and healthcare practitioners in controlling the infection of the coronavirus using the data mining technique discussed.**

***Keywords*— Data mining techniques; pandemic outbreak; healthcare; classification; algorithms.**

Manuscript received 15 Sep. 2020; revised 31 Jan. 2021; accepted 14 Mar. 2021. Date of publication 30 Jun. 2021.

I. INTRODUCTION

Currently data has become a commodity in many businesses across different sectors. Data becomes the source to understand business and plan for a better future. The large amount of data about its competitive environment requires methods that can extract pertinent information and provide insight on business strategy. Data mining is a concept that was found in the 1990's which is part of computer science and has been implemented in many sectors for analysis of data and knowledge discovery [1],[2]. Besides, several pandemic outbreaks such as SARS-CoV, MERS-CoV and Covid-19 have applied data mining for data analysis. Several works have been explored in various aspects of the outbreak since the widespread and the increasing prevalence rate of infection in infected countries such as the identification of the virus source, analysis of the gene sequences, analysis of patient information and analysis of the first cases in the countries involved.

However, pandemic databases consist of a huge amount of data, but hidden dataset cannot be discovered due to lack of effective tools. The conventional methods are not suitable to process and analyze the massive volume of data from the healthcare sector since the data is too complicated and voluminous [3]. This is true even in healthcare where there is a need for appropriate computer-based information or a decision support system for managing pandemic situations. Therefore ,the data mining method is essential to be implemented in the medical area and it has been widely used for the past few years mainly in patient outcome prediction, evaluation of treatment effectiveness, control of infection and diagnosis of diseases [4],[5]. Moreover, successful data mining applications have provided the opportunity for the various parties to make full use of them, as they have recognized that data mining is important for all sectors related to the healthcare industry to collect valuable information [5]. Data mining implementation enables health insurers to identify fraud cases and violence, health managers are able to make good decisions, especially when engaging with their customers, and healthcare practitioners can improve the treatment and care of their patients [6],[3]. Insights gained from data mining will affect expenditures,

sales and operational efficiency while preserving a high quality of service. Data mining performed by healthcare organizations are ideally prepared to address their long-term needs in the healthcare industry [3].

Classification is a commonly used data mining method, which relies on assigning each record to a predefined category. Predicting a class of new entrants is used in the prediction stage by predictive data mining techniques. An efficient and accurate implementation of suitable data mining techniques is seen as crucial, especially in healthcare. Thus, this review was to identify the relevant mining techniques of pandemic outbreaks in relation to healthcare. Furthermore, a review on the algorithms is included in the discussion to identify the algorithms for pandemic outbreaks and rank them. Further knowledge on this area can be found in the sections below. The sections below describe certain important understanding dealing in the areas of pandemic, data mining and its techniques.

A. Pandemic Outbreak

According to Merriam Webster, pandemic is an epidemic of diseases that befalls over a large geographical area and population which can be described as a pandemic outbreak of disease infection. WHO which has declared the pandemic outbreak including the coronavirus disease since 2002, started with SARS-CoV, then MERS-CoV in 2012 and the latest one in 2019, Covid-19 [7],[8]. In addition, SARS-CoV is a respiratory disease that is caused by an unknown contagious virus which is transmitted from a person to another [9]. WHO announced that the virus originated from an animal in an unknown animal reservoir, most likely from bats. Then it was transmitted to other animals such as civet cats which infected southern China in 2002. Besides, this disease infected 8,096 people and caused 774 deaths in 26 countries around the world. MERS-CoV is an infectious respiratory disease that is a pathogen of the MERS Coronavirus known as the RNA virus from the Coronaviridae family [10],[11]. The outbreak of this disease started in Arab Saudi in September 2012 and was identified as one of the large families of the coronaviruses that can cause mild and moderate fever [10]. This virus infected 191 people and caused 82 deaths in 27 countries around the world. The majority of the cases with about 156 positive cases and 63 deaths were from United Emirates [12]. Furthermore, Covid-19 is a virus undergoing the mutation process with spike glycoprotein, which has a high possibility to infect human beings [13]. This virus has a single-strand RNA belonging to the Betacoronavirus family and is called the coronavirus-2 syndrome as reported by the International Committee on Virus Taxonomy [14]. It has spread out actively from humans to humans and has been declared as a concern to public health and emergency [15]. This disease has caused the largest pandemic outbreak with more than 10 million positive cases and more than 200,000 deaths in 211 countries all over the world.

B. Data Mining

Data mining is a concept founded in the 1990s. It is one part of computer science that is bound up with disciplines such as statistics, probability, artificial intelligence, and machine learning, which has become a good research field that has gained a lot of attention due to its approach to data processing and information discovery [1],[2],[16]. Besides, it is defined as a set of rules, procedures, and algorithms to produce valuable insights, derive patterns, and draw connections from huge datasets [16]. It also incorporates advanced data retrieval, processing and simulation using several methods and techniques. It is a modern field recognized as one of the top ten sciences affecting technology, with different applications [17]. In order to use data mining algorithms for health data, the understanding of the researchers about the nature and roles of data mining techniques should be clear. Descriptive which is unsupervised learning and predictive which is supervised learning are both data mining algorithms [3]. Descriptive is analysed by detecting unknown trends or correlations in the data whereby the users may recognize an intensive data pool by the similarity of the items or records. It is known as investigative data mining for clustering, association, summarization, and sequence discovery.

According to [18] a predictive model aims to make it possible for the data miner to predict a specific of an uncertain variable, the future value of the targeted variable. Thus, the predictive model, probably including the previous target variable values, is created from the known variable values given [19]. Several major data mining techniques have been developed and applied in pandemic outbreaks such as classification, clustering, and association rules. Classification is a method that allocates objects in a series to target groups whereby the same set of features is put into a class by this approach [16]. Naïve Bayes, Decision Tree, Artificial Neural Network, Support Vector Machine, Associative Classification and K-Nearest Neighbours are some of the methods of classification [2]. Moreover, the method of clustering identifies clusters of data objects that are identical to each other in certain data [20]. The data points are clustered to maximize intra-class similarities and minimize inter-class similarities based on the characteristics of the data points [19]. This method identifies classes and sets objects in each category, while objects are listed in predefined categories in the classification method. Some of the clustering approaches are K-Means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Fuzzy Clustering, and Expected Maximization (EM) [20].

Besides, healthcare data will be categorized into manageable data with useful information and may help to recognize patterns by enabling many effective data extraction processes by using this clustering method [18]. The association rule method discovers new relationships in a database of variables. Based on the input data set, it is designed to identify well-built rules from the database using various important procedures. In addition, it comprises data mining process for identifying rules, finding recurrent patterns, correlations, similarities, or combinatorial logic between sets of items that can regulate associations between sets of items and causal artifacts [5]. Basket data analysis, cross-marketing and catalogue design are the core applications of association rule mining.

C. Data Mining in Pandemic Outbreak

Traditional methods are very limited to be used in healthcare because the data produced are vast and complex

to be interpreted and analysed using the system; thus, data mining has been introduced in this area to predict outcomes, evaluate the effectiveness of treatment, control of infection and diagnosis of diseases [4]. A study by [10] applied two types of classification algorithms; Naive Bayes and J48 from Decision Tree to analyse the status of cases of MERS-CoV disease by building predictive models for a dataset of 1082 cases between 2013 and 2015. The result of the accuracy, precision and recall for all the models was between 53.6% and 71.58%. [9] compared SARS-CoV and MERS-CoV datasets by using different techniques of data mining. Various types of algorithms were repeated through normal, polynomial, and sigmoid to show that MERS-CoV and SARS-CoV are distinct viruses. [14] employed three classification algorithms including Random Forest, Support-Vector Machine (SVM) and Naïve Bayes to diagnose patients with early symptoms of MERS-CoV using 322 datasets which consisted of 92 infected cases and 230 uninfected cases where Random Forest showed a good performance with a ROC measurement of 0.942%.

AlMoammar [4] performed two classification types for the model performance using the 2013 to 2017 MERS-CoV dataset from Saudi Arabia. Three types of algorithms, namely k-NN, SVM and Decision Tree were used to measure the accuracy of the dataset. Despite the emerging technology, researchers have recently used other data mining techniques, called trajectory data mining, for scientific studies as well as for infection prevention and control approaches. Trajectory data means a trace that uses a sequence of geographical locations that consists of a geospatial coordinate set and a timestamp [21]. Research by [22] on epidemic prevention and control has given 90% accuracy on the sample of 20% of the infected population. Moreover, the nature of moving objects in pandemic introduces biasness and complexity in data processing. Therefore, the study introduced the spatio-temporal variability method. This allows the research to consider spatial information within temporal dimensions in which the data is sampled based on different frequencies [21]. In addition, a Covid-19 research was undertaken by [23] to examine the situation of the pandemic worldwide in terms of its cases, deaths, and recovery. Most studies focus on predicting the situation by commonly using forecasting techniques ranging from the naïve method to ARIMA in improving the root mean square error score for each technique. Another study by [24] predicted Covid-19 cases for 187 countries using 13 different time series forecasting models consisting of statistical and machine learning models for long- short- term memory (LSTM). Mainly, the results of these studies show that ARIMA offers a better projected result for the exponential growing pattern, while Holt's Linear Trend outperforms the other models for both the exponential and linear patterns.

D. Data Mining in Healthcare

In healthcare, traditional methods have limited use because the dataset is massive and complex to be processed and analysed within the method. Hence, the implementation of data mining is essential in that field. Nowadays, data mining is widely used in this area to predict outcomes, evaluate the effectiveness of treatment, control infection and diagnose diseases [4]. Some studies have also shown a positive outcome of the use of data mining in the medical field, which aims to enhance treatment efficacy and the detection of fraud in health insurance, which lightens the workload burden, helps doctors make diagnosis in making their clinical decision, as well as reduces healthcare sector costs [4],[25]. In data mining, disease or virus prediction plays a crucial role and a lot of study in fields such as breast cancer, heart disease, neonatal jaundice, and coronaviruses, uses data mining techniques [26].

Besides, more information and new perspective approaches to certain diseases can be identified which can gain knowledge to increase research in the field of medicine. [27] stated that a good example of contribution of data mining to medicine is through the high degree of accuracy of the models developed in the field. [28] applied data mining techniques in their research in the medical area by using different types of techniques in the prognosis and diagnosis of heart disease. This study helps physicians by giving an accurate result of the diagnosis to differentiate between benign breast tumour and an invasive one without performing any surgical biopsy besides reducing the treatment cost. The result shows the ability of data mining in the medical area to support clinical decision- making based on the diseases.

Furthermore, [29] applied a few of classification methods such as KNN, Naïve Bayes and Random Forest for diabetes analysis and its prediction using two datasets; PIDD (Pima Indian Diabetes Dataset) and the 130_US hospital diabetes dataset. The result showed that the ensemble approach facilities outperformed other algorithms. The accuracy of the proposed ensemble approach was 93.6 2%for PIDD and 88.56% for the 130_US hospital dataset. In addition, RBF Network, Naïve Bayes and J48 was applied by [23] for developing a benign and malignant breast cancer prediction model using the breast cancer dataset from the UCI Machine Learning repository. Besides, the 10-fold cross-validation method was used to compare the performance between three algorithms. The findings showed that with 97.36% accuracy on the holdout study, the Naive Bayes was the best predictor. RBF Network was second with 96.77% accuracy and J48 was third with 93.41% accuracy based on the average accuracy Breast Cancer dataset.

The structure of the paper consists of, next section powers the paper towards the understanding of reviews, themes, and its relevant keywords. Section III presents the findings and elaborates further its algorithm with respect to the objective of the paper and section IV concludes the discussion and suggests further work.

II. Material and Method

In this section, two different gaps of review are presented which consist of the artificial intelligence (AI) techniques in data mining used in healthcare and pandemic outbreaks. A literature search was performed to classify recent reviews on these latest techniques performed in healthcare and pandemic outbreak research. The research utilized a semi structured qualitative method for gathering and analysing information. The information was arranged based on patterns or topics and composed to analyse important data and keywords. In addition, this study also focused on various

types of information sources that consisted of research papers, electronic databases, computer journals and books. Some of the keyword terms used were "data mining techniques", "data mining algorithms", "data mining in healthcare", "pandemic outbreak", "Covid-19", "MERS-CoV" and "SARS-CoV". The sources for the articles were Google Scholar, Science Direct, Springer, IEEE, Scopus and Medical Website. 38 papers consisting of 28 papers on data-mining techniques in healthcare and 10 papers on the pandemic outbreak were studied and analysed to gather all the information on data mining. The review included articles published within the last 10 years mainly from 2011 until 2020. All the sources of the review were analysed to come out with the appropriate review about data mining techniques in healthcare and pandemic outbreaks.

The description of the classification methods that mostly used in the healthcare sector are described below.

1) Naïve Bayes:

The Bayesian classifier is a type of statistical probabilistic model that essentially uses the theorem of Bayes and treats all characteristics as independent. Class conditional independence is also included, where correlations between class attributes are ignored. [30],[31] mentioned that many researchers found that Naïve Bayes has excellence performance in terms of computational efficiency since it has better performance than other algorithms; Decision Tree and neural networks which can handle missing data and produce high prediction accuracy.

Furthermore, this model is easy to build because it uses the probabilistic conditional approach to analyse datasets by multiplying the individual probabilities of each pair of value attributes, which will help the user derive more functionality from the data without being over-fitted even for large datasets [30],[25]. This classifier is very stable; it can be trained with small datasets to create accurate parameter predictions since it only needs the calculation from the frequencies and the outcomes of the pairs of attributes in the training datasets [32], [27].

2) Decision Tree:

Decision tree algorithm was invented by Ross Quinlan in 1979. Known as Iterative Dichotomiser (ID)3 it has been widely used for practical methods of classifications [33]. It represents a method for assessing or grouping interesting objects by graphical means consisting of a flowchart that is like a tree structure which is simple and easy to be implemented and interpreted by human experts [34]. It exhibits functions by mapping the constituent of a specific domain to a corresponding set of constituents with characters or figures that represent a class [11]. According to [35], a decision tree is part of the classification algorithm in which a test on an input case attribute is shown by each non-leaf node; each branch corresponds to a test result; and a class prediction attribute is shown by each leaf node. The outgoing node is an internal node while the other nodes are referred to as leaves or terminal nodes or decision nodes [32]. The dependent variable is estimated on the basis or selected by the evaluated values of all the other attributes, while the other attributes are known as independent variables in the dataset [32] Through this method, learned trees can be constructed as a set of rules of IF-THEN which help readability. The precision of classification and the size of a decision tree are applied to estimate its performance.

3) Neural Network:

Neural network (NN) or artificial neural network (ANN) is a type of biological neuron system that recognizes patterns and generates predictions in data mining [34],[19]. This algorithm forms structures which are made up of multiple nodes and is modelled on the function of the human brain [36]. Knowledge is expressed in ANN as a complex set of linked processors called neurons. The neurons will interact with each other, connected by many connections. Input data is approved by the nodes and basic operations on the data are carried out. The product of these processes is transferred to another neuron. Each output node is named or enabled as a node value [36]. Various kinds of neural network models are applied to solve market problems and play a critical role as a scientific form of research for operations [2]. In recent years, the biggest breakthroughs in neural networks have been their implementation to real-world problems such as the prediction of consumer response, detection of fraud, etc. Besides, ANN is one of the newest tools for processing signals. It is an adaptive, non-linear system that learns to perform a data function, and that adaptive step is typically a stage of training where during operations the system parameter changes [36], [2].

A. Model Evaluation

Through this study it was found that the most common evaluation models used were mainly three performance measures: accuracy, precision, and recall [1],[4],[33],[22]. Since the pandemic and healthcare involve voluminous and large datasets, this performance was significant to identify the best result besides developing a predictive model for the datasets [10],[1]. Accuracy is the calculation of the correctly classified record in percentage and referred to as summation of the correctly classified positive records and the correctly classified negative records over the summation of positive and negative records. Meanwhile, precision is the amount of data that the model accurately categorized as positive for all positive predictions. Recall classifies the correct number of records as positives while classifying incorrect number of records as negatives for the overall summation of these records. These measures are calculated as follows:

$$\text{Accuracy} = (TP+TN)/(P+N) \quad (1)$$

$$\text{Precision} = TP/(TP+FP) \quad (2)$$

$$\text{Recall} = TP/(TP+FN) \quad (3)$$

In these three equations, the number of positive records is P while the number of negative records is N. The number of records that were correctly classified as positive is TP while the number of records that were correctly classified as negative is TN. The number of records that were misclassified as negative is represented as FN.

III. Result and Discussion

This section describes the data-mining techniques from the reviews and their implementation on pandemic outbreaks

which focused on SARS-CoV, MERS-CoV and Covid-19 in healthcare.

A. Data Mining Techniques in Healthcare

The current global mobility exposed healthcare to many new and complex diseases. Table 1 shows several methods of data mining implemented by past researchers in the domain of healthcare.

TABLE I
REVIEW PAPER WITH DIFFERENT TYPES OF DATA- MINING METHODS IN HEALTHCARE

Data mining methods	Number of papers
Classification	20
Clustering	2
Association	0
Classification & clustering	1
Classification & association	2
Classification & clustering & association	3

It is common in many countries to monitor and practice good control of healthcare by investigating data collected about people, movement, and environment for health prediction. Some research has been conducted to determine the patterns and predict useful information and threat from medical datasets [37],[38]. Based on this aim and through the reviews, three methods were found to be widely used by past researchers in healthcare. These three methods were classification, clustering, and association rule. Generally, classification is used by researchers to extract valuable information besides determining pattern, predict future outcomes and trends in healthcare [3] which act as a discovery of data items in predefined classes [21]. Besides classification, there is also a combination of methods and this is to increase accuracy and performance of the algorithms.

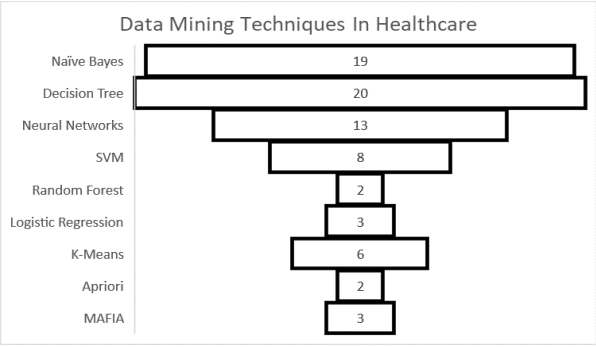

Fig. 1 Types of algorithm that have been Implemented in healthcare.

Figure 1 shows the different types of algorithms used in healthcare. Algorithms for classification methods, namely Naïve Bayes and Decision Tree, and K-Means for clustering are commonly used in healthcare. 20 papers implemented Decision Tree while 19 papers used Naïve Bayes to predict various types of datasets in healthcare such as heart disease, cancer, diabetes, liver disease and neonatal jaundice. 6 papers used the clustering method with K-Means while 5 papers used the association method where 2 papers used the Apriori algorithm while 3 papers used the MAFIA algorithm. These methods and their algorithms present better accuracy and performance compared to other methods like clustering and association rule.

B. Data Mining Techniques in Pandemic Outbreaks

Referring to the reviewed papers identified, the analysis depicted various algorithms used by many researchers on pandemic outbreaks. Table 2 presents the most common algorithms and their datasets as well as the tools used by the researchers to build predictive models in the pandemic outbreak.

Naïve Bayes, Decision Tree and SVM are the most common algorithms used in diagnosing pandemic outbreaks. These algorithms have the highest accuracy and better performance, so they are commonly used by the researchers. Based on 10 papers, Decision Tree is the highest algorithm used for pandemic outbreak datasets. Besides, [27] stated that the automatic selection of features and the reduction of complexity can be carried out by Decision Tree classifiers, where the tree structure provides easy-to-understand and comprehensible information on the predictive or generation capability of classification. [31] mentioned that results from many researchers show that Naïve Bayes has excellent performance in terms of computational efficiency since it has better performance than other algorithms as it can handle missing data and produce high accuracy in prediction.

Furthermore, this model is easy to build because it uses the probabilistic conditional approach to analyse dataset [34]. It also helps the user to derive more functionality from the data without being over fitted even for large datasets [10],[21]. SVM is also selected by researchers since this algorithm is akin to solving a linear, quadratic programming problem and is commonly used in disease diagnostics because of the high accuracy in prediction [21].

In addition, [22] used a high-resolution spatio-temporal model with a fine and complex spatial resolution for disease risk assessment. High precision is given by the spatio-temporal pattern of Covid-19 using various algorithms. The regional risk measured using modelling maps shows a strong correlation between the total number of reported cases in the area [39]. Besides, type and structure of datasets will affect the performance of algorithms. The finding also found that some researchers are trying to explore more techniques such as using ensemble approach between the algorithms.

TABLE III
TYPES OF ALGORITHMS USED IN PANDEMIC OUTBREAK DATASETS

Authors	Datasets	Details of datasets	Data mining techniques	Tool	Outcomes
[10]	MERS-CoV datasets between 2013-2015	The datasets are divided into three categories, namely new cases with recorded data of 633 cases; recoveries with 231 records and death records of 1082.	Naïve Bayes, Decision Tree (J48)	WEKA	Naïve Bayes achieved the highest accuracy (71.58%) for the recovery model; meanwhile Decision Tree achieved 55.69% for the stability model.
[9]	DNA sequences of MERS-CoV	These NCBI datasets provide protein sequence of SARS and MERS Spike glycol protein data.	Decision Tree, SVM and Apriori algorithm	Mathematical model	SVM showed a better performance in distinguishing the Spike glycoprotein of MERS and SARS with more than 75% of accuracy.
[11]	DNA sequences of MERS-CoV	The datasets contain pandemic virus data in different regions of the world.	Decision Tree, and Apriori algorithm	Mathematical model	The results indicated countries that are vulnerable to MERS-CoV transmission, namely middle eastern regions.
[40]	Multiple health-related attributes that consist of personal attributes and MERS-CoV related attributes	The datasets contain data over a period of time on personal attributes and MER-CoV attributes	BBN Classification	WEKA, Amazon EC2 and R Studio	The result showed that 83.1% was the accuracy by using the classifier with synthetic data
[14]	MERS-CoV datasets from UCI and medical analytical papers	The datasets represent infected cases of 92 out of 322 records, meanwhile 230 are uninfected cases. Both contain 24 attributes.	Random Forest, Naïve Bayes, and Support Vector Machine	WEKA	Random Forest Classifier presents good performance with ROC measurement at 0.942%
[1]	MERS-CoV datasets between 2013 until second half of 2016	Datasets in the text-based form (infographic)	k-NN, Decision Tree, and Naïve Bayes	RapidMiner Studio (version 7.0)	The highest accuracy for binary classification is Decision Tree (90%), meanwhile multiclass classification is k-NN (51.60%) and Naïve Bayes for multilabel classification (77%)
[4]	MERS-CoV datasets between January 2013and October 2017	The datasets represent records on living and death that occurred from 2013 to 2017. There were 1,186 and 224 records respectively.	Decision Tree, Support Vector Machine, and k-NN classifiers	RapidMiner Studio version 7.6	The highest accuracy: 86.44%, was achieved for Binary classification such as Decision Tree and SVM while for Multiclass classification such as Decision Tree it was 42.80%
[41]	Covid-19 datasets	The dataset had 482 records from Johns Hopkins Github repository	Random Forests, Logistic Model Trees, Decision Tree, and Naive Bayes		Random Forest showed the highest accuracy (0.9083) with Kappa Statistics
[42]	COVID-19 data from South Korea	The datasets had3254 instances and 8 attributes that were obtained from KCDC which was made available on the Kaggle Website	Random Forest, Support Vector Machine, Naive Bayes, Logistic Regression, Decision Tree, and k-NN	Python	The Decision Tree model developed gave the highest accuracy of 99.85%.
[22]	COVID-19 trajectory datasets in China	Datasets of 2020 with 9000 diagnosed cases and 120 thousand normal cases in 25 regions of China.	Random Forest, Decision Tree, Support Vector Machine, and K-Means	Mathematical model	All algorithms showed the performance that were above 90% for various rates of the population infection that ranged from 1% to 20%

IV. Conclusion

This paper has reviewed and analyzed the techniques of data mining, its method, and algorithms in healthcare and pandemic healthcare outbreak. The findings obtained indicate that large number of studies have applied classification method in healthcare for medical diagnosis and disease prediction with Naïve Bayes and Decision Tree algorithms top the list. This is also recorded in pandemic outbreak where three algorithms that widely used under classification is Decision Tree, Bayesian Network, and SVM. Moreover, J48 and C4.5 are models frequently used in Decision Tree to evaluate and compare results besides to calculate prediction and accuracy. For future research, many possibilities and solutions provided in this paper for data mining in healthcare besides there are many other avenues to be explored in relation to different aspects of health data, such as privacy, quality, cost and so on. Next, these methods will be tested and evaluated on large dataset of Covid-19.

Acknowledgment

This work is supported under the Fundamental Research Grant Scheme (600-IRMI/FRGS 5/3 (370/2019). We thank the Ministry of Higher Education and IRMI (Institute of Research, Management, and Innovation), UiTM for their continuous support and to anonymous reviewers for their useful suggestion.

References

[1] N. Almansour and H. Kurdia, "Identifying accurate classifier models for a text-based MERS-CoV dataset," *2017 Intelligent Systems Conference, IntelliSys 2017,* 2018-Janua (September), pp. 430–435, https://doi.org/10.1109/IntelliSys.2017.8324330, 2018.

[2] M. K. Gupta and P. Chandra, "Original Research," *International Journal of Information Technology,* https://doi.org/10.1007/s41870-020-00427-7, 2020.

[3] S. A. Lashari, R. Ibrahim, N. Senan and N. S. A. M. Taujuddin, "Application of Data Mining Techniques for Medical Data Classification: A Review," vol. 06003, pp. 1–6, 2018.

[4] A. AlMoammar, L. AlHenaki and H. Kurdi, "Selecting accurate classifier models for a MERS-CoV dataset," *Advances in Intelligent Systems and Computing.* https://doi.org/10.1007/978-3-030-01054-6_74, 2018.

[5] P. Kauser Ahmed, "Analysis of data mining tools for disease prediction," *Journal of Pharmaceutical Sciences and Research,* 2017.

[6] K. M. M. N. K and S. R, "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks," *International Journal of Data Mining Techniques and Applications,* https://doi.org/10.20894/ijdmta.102.007.001.027, 2018.

[7] H. Harapan, N. Itoh, A. Yufika, W. Winardi, S. Keam, H. Te, D. Megawati, Z. Hayati, A. L. Wagner and M. Mudatsir, "Coronavirus disease 2019 (COVID-19): A literature review," *Journal of Infection and Public Health,* https://doi.org/10.1016/j.jiph.2020.03.019, 2020.

[8] C. Wang, P. W. Horby, F. G. Hayden and G. F. Gao, "A novel coronavirus outbreak of global health concern," *The Lancet,* vol. 395, no.10223, pp. 470–473. https://doi.org/10.1016/S0140-6736(20)30185-9, 2020.

[9] S. Jang, S. Lee, S. M. Choi J., Seo, H. Choi and T. Yoon, "Comparison between SARS CoV and MERS CoV Using apriori algorithm, decision tree, SVM," *MATEC Web of Conferences,* vol. 49, pp. 4–7. https://doi.org/10.1051/matecconf/20164908001, 2016.

[10] I. Al-Turaiki, M. Alshahrani and T. Almutairi, "Building predictive models for MERS-CoV infections using data mining techniques," *Journal of Infection and Public Health,* vol. 9, no. 6, pp. 744–748, https://doi.org/10.1016/j.jiph.2016.09.007, 2016.

[11] D. Kim, S. Hong, S. Choi and T. Yoon, "Analysis of transmission route of MERS coronavirus using decision tree and Apriori algorithm," *International Conference on Advanced Communication Technology,ICACT,* https://doi.org/10.1109/ICACT.2016.7423472, 2016.

[12] Z. A. Memish, M. Cotten, S. J. Watson, P. Kellam, A. Zumla, R. F. Alhakeem, A. Assiri, A. A. A. Rabeeah and J. A. Al-Tawfiq, "Community Case Clusters of Middle East Respiratory Syndrome Coronavirus in Hafr Al-Batin, Kingdom of Saudi Arabia: A Descriptive Genomic study," *International Journal of Infectious Diseases,* vol. 23, pp. 63–68. https://doi.org/10.1016/j.ijid.2014.03.1372, 2014.

[13] M. Giovanetti, D. Benvenuto, S. Angeletti and M. Ciccozzi, "The first two cases of 2019-nCoV in Italy: Where they come from?," *Journal of Medical Virology,* vol. 92, no. 5, pp. 518–521. https://doi.org/10.1002/jmv.25699, 2020.

[14] M. Abdullah, M. S. Altheyab, A. M. A. Lattas and W. F Algashmari, "MERS-CoV disease estimation (MDE) A study to estimate a MERS-CoV by classification algorithms," *Communication, Management, and Information Technology - Proceedings of the International Conference on Communication, Management, and Information Technology, ICCMIT 2016, December,* pp. 633–638, 2017.

[15] J. H. Yoo, "The Fight against the 2019-nCoV Outbreak: An Arduous March Has Just Begun," *Journal of Korean Medical Science,* vol. 35, no. 4, pp. 2019–2021, https://doi.org/10.3346/jkms.2020.35.e56, 2020.

[16] V. Plotnikova, M. Dumas and F. Milani, "Adaptations of data mining methodologies: A systematic literature review," *PeerJ Computer Science,* https://doi.org/10.7717/PEERJ-CS.267, 2020.

[17] R. Ghorbani and R. Ghousi, "International Journal of Data and Network Science," vol. 3, pp. 47–70. https://doi.org/10.5267/j.ijdns.2019.1.003, 2019.

[18] P. N. Mahalle, N. P. Sable, N. P. Mahalle and R. Gitanjali, "Predictive Analytics of Covid-19 using Information, Communication and Technologies," April 1–9. https://doi.org/10.20944/preprints202004.0257.v1, 2020.

[19] S. Mukherjee, R. Shaw, N. Haldar and S. Changdar, "A Survey of Data Mining Applications and Techniques," vol. 6, no. 5, pp. 4663–4666, 2015.

[20] A. Saxena, M. Prasad, A. Gupta, N. Bharill, O. P. Patel, A. Tiwari, M. J. Er, W. Ding and C. T. Lin, "A review of clustering techniques and developments," *Neurocomputing*, https://doi.org/10.1016/j.neucom.2017.06.053, 2017.

[21] Y. U. Zheng, "Trajectory data mining: An Overview," vol. 6, no. 3, pp. 1–41, 2015.

[22] C. Zhou, W. Yuan, J. Wang, H. Xu, Y. Jiang, Q. H. Wen and P. Zhang, "Detecting suspected epidemic cases using trajectory big data," pp. 1–19, 2020.

[23] V. Chaurasia, "Application of machine learning time series analysis for prediction COVID-19 pandemic," Cdc, 2020.

[24] L. Ismail, H. Materwala, T. Znati, S. Turaev and M. A. B. Khan, "Tailoring time series models for forecasting coronavirus spread: Case studies of 187 countries," *Computational and Structural Biotechnology Journal,* vol. 18, pp. 2972–3206, https://doi.org/10.1016/j.csbj.2020.09.015, 2020.

[25] N. Jothi, N. A. Rashid and W. Husain, "Data Mining in Healthcare - A Review," *Procedia Computer Science,* vol. 72, pp. 306–313. https://doi.org/10.1016/j.procs.2015.12.145, 2015.

[26] M. A. Nishara Banu and B. Gomathy, "Disease forecasting system using data mining methods," *Proceedings - 2014 International Conference on Intelligent Computing Applications, ICICA 2014,* pp. 130–133. https://doi.org/10.1109/ICICA.2014.36, 2014.

[27] S. Kaur and R. K. Bawa, "Review on data mining techniques in healthcare sector," *Proceedings of the International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud), I-SMAC 2018,* https://doi.org/10.1109/I-SMAC.2018.8653795, 2019.

[28] R. Singh and E. Rajesh, "Prediction of Heart Disease by Clustering and Classification Techniques," *International Journal of Computer Sciences and Engineering.* https://doi.org/10.26438/ijcse/v7i5.861866, 2019.

[29] M. Alehegn, R. R. Joshi and P. Mulay, "Diabetes Analysis and Prediction Using Random Forest, KNN, Naïve Bayes And J48: An Ensemble Approach," vol. 8, no. 09, 2019.

[30] I. Ahmed and A. Mousa, "Security and privacy issues in e-healthcare systems: Towards trusted services," *International Journal of Advanced Computer Science and Applications,* vol. 7, no. 9, pp. 229–236. https://doi.org/10.14569/ijacsa.2016.070933, 2016.

[31] M. A. Jabbar, B. L. Deekshatulu and P. Chandra, "Computational intelligence technique for early diagnosis of heart disease,"

ICETECH 2015 - 2015 IEEE International Conference on Engineering and Technology. https://doi.org/10.1109/ICETECH.2015.7275001, 2015.

[32] T. R. Baitharu and S. K. Pani, "Analysis of data mining techniques for healthcare decision support system using liver disorder dataset. *Procedia Computer Science,* vol. 85, pp. 862–870. https://doi.org/10.1016/j.procs.2016.05.276, 2016.

[33] J. Han, M. Kamber and J. Pei, "Data mining: Concepts and Techniques," In *Data mining: Concepts and techniques.* https://doi.org/10.1016/C2009-0-61819-5, 2012.

[34] V. Gayathri, and M. C. Mona, "A survey of data mining techniques on medical diagnosis and research," vol. 6, no. 6, pp. 301–310. 2014.

[35] H. M. Zolbanin, D. Delen and A. Hassan Zadeh, "Predicting overall survivability in comorbidity of cancers: A data mining approach," *Decision Support Systems,* vol. 74, pp. 150–161. https://doi.org/10.1016/j.dss.2015.04.003, 2015.

[36] B. V. Chowdary, "A survey on applications of data mining techniques," vol. 13, no. 7, pp. 5384–5392, 2018.

[37] Amin, M. S., Chiam, Y. K., & Varathan, K. D. (2019). Identification of significant features and data mining techniques in predicting heart disease. Telematics and Informatics, 36, 82–93. https://doi.org/10.1016/j.tele.2018.11.007

[38] S. Vijayarani and S. Sudha, "An efficient clustering algorithm for predicting diseases from hemogram blood test samples," pp. 1–8. https://doi.org/10.17485/ijst/2015/v8i, 2015.

[39] P. Radanliev, D. D. Roure and R. Walton, Diabetes & Metabolic Syndrome: Clinical Research & Reviews Data mining and analysis of scienti fi c research data records on Covid- 19 mortality, immunity, and vaccine development - In the first wave of the Covid-19 pandemic," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews,* vol. 14, no. 5, pp. 1121–1132. https://doi.org/10.1016/j.dsx.2020.06.063, 2020.

[40] R. Sandhu, S. K. Sood and G. Kaur, "An intelligent system for predicting and preventing MERS-CoV infection outbreak," *Journal of Supercomputing,* vol. 72, no. 8, pp. 3033–3056. https://doi.org/10.1007/s11227-015-1474-0, 2016.

[41] A. Keshavarzi, "Coronavirus Infectious Disease (COVID-19) Modeling: Evidence of Geographical Signals," *SSRN Electronic Journal,* https://doi.org/10.2139/ssrn.3568425, 2020.

[42] 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, pp. 1–7. https://doi.org/10.1007/s42979-020-00216-w, 2020.