Journal of Medical Systems (2020) 44: 122
https://doi.org/10.1007/s10916-020-01582-x

SYSTEMS-LEVEL QUALITY IMPROVEMENT

Role of biological Data Mining and Machine Learning Techniques in Detecting and Diagnosing the Novel Coronavirus (COVID-19): A Systematic Review

A. S. Albahri[2] · Rula A. Hamid[3] · Jwan k. Alwan[4] · Z.T. Al-qays[5] · A. A. Zaidan[1] · B. B. Zaidan[1] · A O. S. Albahri[1] · A. H. AlAmoodi[1] · Jamal Mawlood Khlaf[6] · E. M. Almahdi[7] · Eman Thabet[8] · Suha M. Hadi[2] · K I. Mohammed[1] · M. A. Alsalem[9] · Jameel R. Al-Obaidi[10] · H.T. Madhloom[11]

Received: 14 March 2020 / Accepted: 27 April 2020 / Published online: 25 May 2020

Abstract

Coronaviruses (CoVs) are a large family of viruses that are common in many animal species, including camels, cattle, cats and bats. Animal CoVs, such as Middle East respiratory syndrome-CoV, severe acute respiratory syndrome (SARS)-CoV, and the new virus named SARS-CoV-2, rarely infect and spread among humans. On January 30, 2020, the International Health Regulations Emergency Committee of the World Health Organisation declared the outbreak of the resulting disease from this new CoV called 'COVID-19', as a 'public health emergency of international concern'. This global pandemic has affected almost the whole planet and caused the death of more than 315,131 patients as of the date of this article. In this context, publishers, journals and researchers are urged to research different domains and stop the spread of this deadly virus. The increasing interest in developing artificial intelligence (AI) applications has addressed several medical problems. However, such applications remain insufficient given the high potential threat posed by this virus to global public health. This systematic review addresses automated AI applications based on data mining and machine learning (ML) algorithms for detecting and diagnosing COVID-19. We aimed to obtain an overview of this critical virus, address the limitations of utilising data mining and ML algorithms, and provide the health sector with the benefits of this technique. We used five databases, namely, IEEE Xplore, Web of Science, PubMed, ScienceDirect and Scopus and performed three sequences of search queries between 2010 and 2020. Accurate exclusion criteria

This article is part of the Topical Collection on *Systems-Level Quality Improvement*

✉ A. A. Zaidan
aws.alaa@gmail.com; aws.alaa@fskik.upsi.edu.my

A. S. Albahri
ahmed.bahri1978@gmail.com

Rula A. Hamid
rula@fskik.upsi.edu.my

Jwan k. Alwan
Jwan@fskik.upsi.edu.my

Z.T. Al-qays
qays@fskik.upsi.edu.my

B. B. Zaidan
bilalbahaa@fskik.upsi.edu.my

A O. S. Albahri
osamah@fskik.upsi.edu.my

A. H. AlAmoodi
Alamoodi@fskik.upsi.edu.my

Jamal Mawlood Khlaf
Jamal@fskik.upsi.edu.my

E. M. Almahdi
Almahdi@fskik.upsi.edu.my

Eman Thabet
Eman@fskik.upsi.edu.my

Suha M. Hadi
Suha@fskik.upsi.edu.my

K I. Mohammed
Mohammed@fskik.upsi.edu.my

M. A. Alsalem
Alsalem@fskik.upsi.edu.my

Jameel R. Al-Obaidi
jr_alobaidi@fskik.upsi.edu.my

H.T. Madhloom
Madhloom@fskik.upsi.edu.my

Extended author information available on the last page of the article

and selection strategy were applied to screen the obtained 1305 articles. Only eight articles were fully evaluated and included in this review, and this number only emphasised the insufficiency of research in this important area. After analysing all included studies, the results were distributed following the year of publication and the commonly used data mining and ML algorithms. The results found in all papers were discussed to find the gaps in all reviewed papers. Characteristics, such as motivations, challenges, limitations, recommendations, case studies, and features and classes used, were analysed in detail. This study reviewed the state-of-the-art techniques for CoV prediction algorithms based on data mining and ML assessment. The reliability and acceptability of extracted information and datasets from implemented technologies in the literature were considered. Findings showed that researchers must proceed with insights they gain, focus on identifying solutions for CoV problems, and introduce new improvements. The growing emphasis on data mining and ML techniques in medical fields can provide the right environment for change and improvement.

Keywords Coronaviruses · MERS-CoV · SARS-CoV-2 · COVID-19 · Artificial Intelligence · Biological Data Mining · Machine Learning

Introduction

The outbreak of new coronavirus (COVID-19) infections has caused worldwide concern because this disease has caused illness, including illness resulting in death and sustained person-to-person spread in many countries [1, 2]. CoVs are a large family of viruses, including the Middle East respiratory syndrome (MERS)-CoV, severe acute respiratory syndrome (SARS)-CoV [3, 4], and the new virus named SARS-CoV-2 [1]. In 2012, Saudi Arabia experienced the outbreak of MERS-CoV, which is responsible for causing mild to moderate colds. Infection with MERS-CoV can lead to fatal complications. MERS-CoV is responsible for causing severe acute respiratory illness that leads to death in many cases. According to Al-Turaiki and his group [4], MERS-CoV symptoms include cough, fever, nose congestion, breath shortness and sometimes diarrhoea. Unfortunately, information on how the virus spreads and how patients are affected is limited [4]. Fifteen years after the first highly pathogenic human CoV caused the SARS-CoV outbreak, another severe acute diarrhoea syndrome-CoV devastated livestock production by causing fatal diseases in pigs. The two outbreaks began in China and were caused by CoVs of bat origin [5, 6]. On February 11, 2020, the World Health Organisation named the ensuing disease 'COVID-19' [1]. Chinese health officials have reported tens of thousands of cases of COVID-19 in China, with the virus reportedly spreading from person-to-person in several parts of the country. COVID-19 illnesses, where most of them are associated with travel from Wuhan, have been reported in a growing number of international locations, including the United States [1].

Artificial intelligence (AI) is gradually changing medical practice. With the recent progress in digitised data acquisition, machine learning (ML) and computing infrastructure, AI applications are expanding into areas that are previously believed to be only the area of human experts [7]. Various types of data mining methods have been applied by a few researchers with real CoV datasets (e.g. MERS-CoV) based on several types of ML classifiers [8]. Providing prediction systems that can accurately anticipate and diagnose such virus remains challenging. The growth of AI-driven techniques to identify epidemiologic risks in advance will be the key to improving the prediction, prevention and detection of future global health risks [9]. The main contributions of this study are the exploration of the CoV family by reviewing articles on data mining and ML algorithms, the acquisition of a clear understanding of its enhancements, and how previous research has addressed prediction, regression, and classification methods. This study also aims to collect various information from the literature that are relevant to ML, such as application nature, the use of ML and data mining algorithms, and evaluation methods and accuracy. The datasets utilised in the literature are constructed and presented with URL sources. The motivations, challenges, and limitations of this approach are examined, and recommendations on improving the approach efficiency are provided. Other important information collected, especially on the types of case study used, features and classes for CoV, are explained in separate tables. The rest of this paper is organised as follows. The second section describes the research methods used in the selected literature. The third section presents the literature review. The fourth section discusses the results, motivations, challenges, recommendations and limitations, case study used, and features and classes of CoV. The fifth section provides a conclusion.

Methods

This study followed the literature review style recommended by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement [10]. Five digital databases, namely, ScienceDirect (SD), IEEE Xplore, Web of Science (WoS), PubMed and Scopus, were selected. SD

provides access to a highly reliable journal in the field of science and technology. IEEE Explore contains updated research papers in the field of computer science, electronic engineering, and the applications of engineering and computer technology in medical applications. WoS is an extremely reliable resource on social sciences, engineering, science, arts, humanities and cross-disciplinary studies. PubMed is considered the optimal database for medical and biomedical engineering research. Scopus is a dependable resource in different areas of research, such as medical, health, science, technology and engineering. The five databases cover all academic aspects of COVID-19. The outcome of this literature review can help save lives by providing deep insights into this disease, existing medical diagnosis systems for this virus, and recommended solutions for developing reliable medical systems.

Search strategy

A comprehensive literature search was conducted in the five mentioned databases for English language citations published from 2010 to 2020. The selection of these indices was because of their sufficient coverage of studies related to our research considering that identified novel CoV requires greater attention than other infections. This study presented and conducted a three Boolean search strategy using various keywords related to pervasive 'coronavirus' (e.g. 'CoV' OR 'coronaviridae' OR 'coronavirus') and keywords related to the detection, diagnosis and classification of CoV under the concept of AI and ML. We used these query techniques to strengthen our search of different AI and ML systems and application studies for CoV.

Inclusion criteria

1. The article is an English journal or conference paper.
2. The main focus is on the development of different artificial intelligence and ML applications, systems, algorithms, methods and techniques.
3. The development only focuses on the detection, diagnosis and classification of adaptive CoV.

Table 1 summarises the sequences of the Boolean search query used in this paper and the results.

Study selection

This process was initiated by removing duplicated articles and screening nonduplicated articles by their titles and abstract to check their compatibility with our inclusion and exclusion criteria. The relevant articles were subjected to a full reading process for collecting and extracting research data and constructing the review article. In all research articles, the entire research process was monitored and supervised by a senior author (corresponding author) to ensure the production of a highly reliable and beneficial research paper.

Data extraction and classification

Given the multidisciplinary topic of this systematic review, data extraction and classification of the selected studies, including data concerning CoV with AI applications (especially ML techniques), were conducted to evaluate the efficacy of this virus in terms of detection, diagnosis and classification throughout AI enhancements. Data elements were extracted from academic literature and included authors' nationalities, date of publication, number of articles per year and number of articles per database. For conferring a comprehensive viewpoint of CoV, this study discussed the CoV and analysed the growth scale of the worldwide epidemic in the context of AI using various data mining and ML algorithms, such as classification, regression and prediction. For each study in the literature, this study extracted the important feature names, evaluation methods used, and state of accuracy for each method. Brief motivation, challenges, limitation and recommendation were extracted from the reviewed papers to address the serious public health concern for CoV.

Results

The results of search queries conducted in this study are presented in Figure 1. Three search queries were accomplished to encompass all databases and their search engine mechanisms during data collection. The first result comprised 1305 articles from all five databases. The number of duplicated articles in all databases was 66, and the results were 1239. The next process was screening the articles based on the title and abstract followed by the mapping of inclusion and exclusion criteria, and the results were 249 articles. The final process was the full reading of all articles, and the outcome was only 8 articles that met the inclusion and exclusion criteria. Our understanding of the purpose/aim of these studies inspired us to analyse each study depending on two related sequences within the search query that was conducted in this systematic review. The first sequence is the article should identify the CoV, and the second is the utilisation of ML.

Distribution results

The findings of academic literature search showed that various algorithms are used by previous researchers. Figure 2 summarises these algorithms and methods. The decision tree (j48) algorithm was the most frequently used (five times). Naive Bayes and support vector machine (SVM) algorithms were each used four times. *k*-Nearest Neighbour (*k*-NN) was utilised two times, and the others were each used once. Figure 3 presents the number of papers included in the

Table 1 Three sequences of the Boolean search query

Seq.	Query Details Terms	Result of Databases	Final Results
1st query	('coronavirus' OR 'coronaviridae' OR 'CoV') AND ('detection system' OR 'diagnosis system' OR 'diagnostic system' OR 'diagnostic application' OR 'diagnosis application')	SD = 766 IEEE = 1 PubMed = 16 WOS = 15 Scopus = 51	849−49 (duplicate) = 800 Articles
2nd query	('coronavirus' OR 'coronaviridae') AND ('detection' OR 'diagnosis' OR classification) AND ('machine learning' OR 'artificial intelligence')	SD = 34 IEEE = 2 PubMed = 2 WOS = 2 Scopus = 185	225−10 (duplicate) = 215 Articles
3rd query	('coronavirus' OR 'coronaviridae') AND ('detection' OR 'diagnosis') AND ('machine learning' OR 'artificial intelligence')	SD = 29 IEEE = 1 PubMed = 1 WOS = 1 Scopus = 265	297−7 (duplicate) = 290 Articles
Final results for all queries			**1305 Articles**

academic literature review following the publication year. The distribution of scientific papers from 2010 to 2020 is shown below. Three papers were published in 2016. Two papers were published each in 2017 and 2018. Only one paper was published in 2019. No paper was published in the other years. Table 2 illustrates the state-of-the-art CoV prediction algorithms. Table 3 presents the CoV dataset descriptions with available sources.

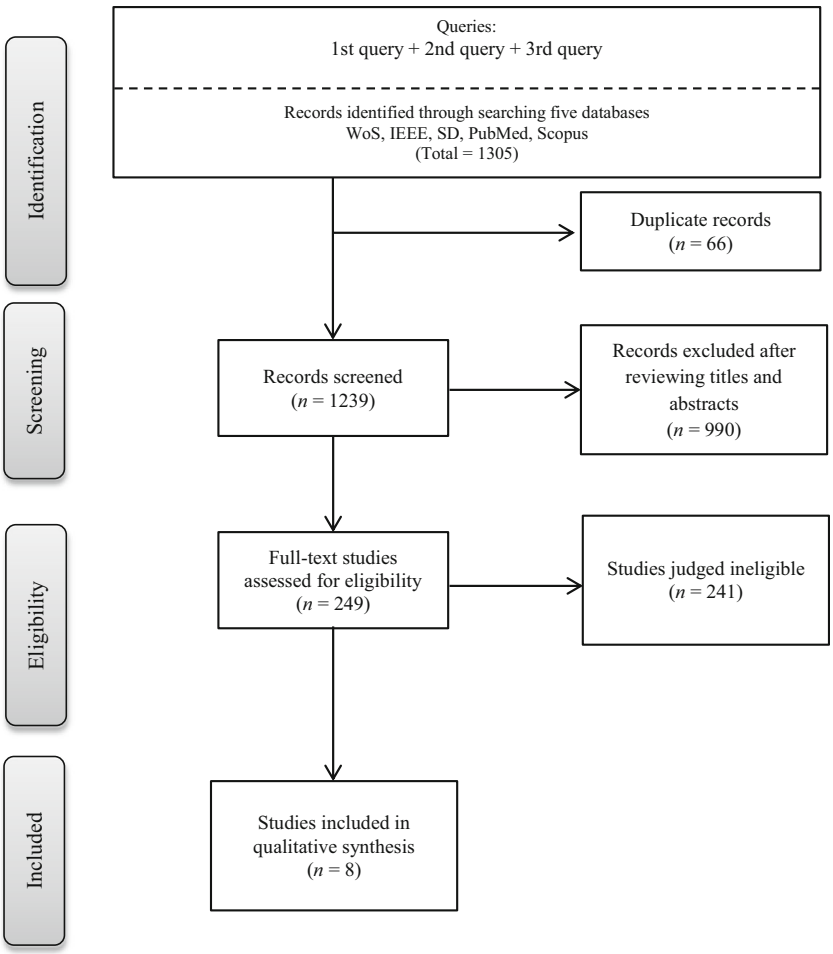

Figure 1 Schematic of the approach to identify, screen and include relevant studies.

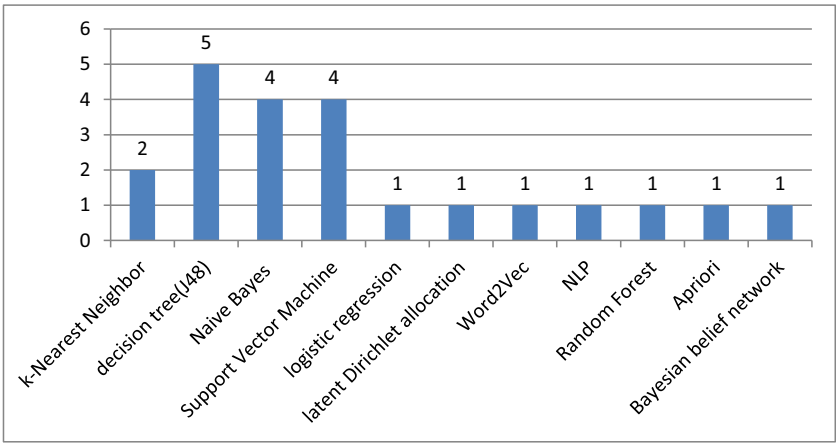

Figure 2 Summary of algorithms and methods used in the literature review

Discussion

The analysis indicated many important points that are discussed to identify the research gaps. In [11], three ML techniques were applied to the MERS-CoV dataset to identify the best classification model for binary class and multiclass labels. The results showed that the *k*-NN classifier is the best model for the two-class problems, and the decision tree and Naïve Bayes are the best models for multiclass problems. In [4], two experiments are applied to the MERS-CoV infection dataset, and the decision tree classifier shows higher prediction proficiency than the other models. The experimental results indicated that age and symptoms are the two dominant features for the prediction model and that healthcare staff are likely to survive.

In [12], a study was conducted in Saudi Arabia to identify the dominant factors that influence human infection using statistical methods, such as univariate and multivariate regression methods. The results indicated four dominant features, namely, disease severity, patient age, the patient job as a healthcare staff or not and history of chronic disease. Interaction with camels does not have a high impact on recovery. In [8], a study was conducted in Saudi Arabia between 2013 and 2017 to improve medical diagnosis systems for binary and multiclass problems in MERS-CoV datasets. The experimental results showed that the decision tree classifier achieves the best accuracy for the multiclass labels, and the SVM classifier obtains the highest accuracy for the two binary class labels in the MERS-CoV dataset.

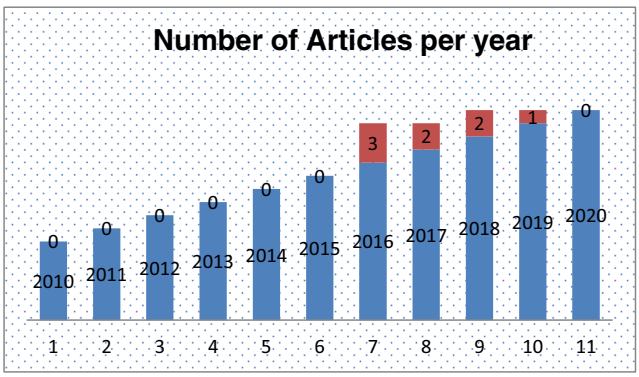

Figure 3 Statistics of included papers by publication year

In [13], an emotional recognition system based on ML technique was proposed to understand human reactions to a widespread epidemic of transferrable diseases, such as MERS. The study was conducted using a dataset collected in Korea in 2015, and the results indicated the impact of lightening excessive panic in reducing infection. In [14], the author investigated people over 50 years old using three ML methods to predict MERS CoV. The results showed that elderly people are more likely to be infected than others. In [15], an SVM classifier based on sigmoid, normal and polynomial iterations was used to analyse the MERS and SARS proteins. The results showed their behaviour similarity and approximate dissimilarities. In [16], data mining based on statistical methods was utilised to develop a cloud-based medical system with a high prediction accuracy to prevent MERS-CoV spread within different regions. The dataset comprises the following attributes: drug, patient and cloud-based user medical record.

The role of AI in healthcare for enhancing the detection and prediction of numerous viruses and diseases has been previously discussed [9]. In this review, we aimed to obtain a large number and extensive contributions of published articles regarding the utilisation of AI for the detection and clinical diagnosis of MERS-CoV and SARS-CoV. However, the lack of studies on the recent outbreak of COVID-19 indicates the need and opportunity to apply AI for predicting such outbreaks. ML technique based on supervised and unsupervised learning provides the opportunity to develop a medical diagnosis system. In supervised learning, the target class of each sample in the dataset, where MERS-CoV and SARS-CoV classes are previously identified and the developed system can be adapted to a new disease. Although MERS-CoV and SARS-CoV have similarity within the same cluster, they are dissimilar to the objects in other clusters. Thus, the clustering technique based on unsupervised learning is considered an

Table 2 State-of-the-art CoV prediction algorithms

Ref.	Application nature	ML and data mining classification algorithms	Evaluation	Accuracy
[11]	Improve infection prediction for MERS-CoV	(Classification) *k*-NN Decision tree Naïve Bayes	Accuracy using cross-validation model	90%
[4]	Build several prediction models for MERS-CoV	(Classification) Naïve Bayes classifier J48 decision tree	Accuracy, precision and recall	Between 53.6% and 71.58%.
[12]	Identify the important factors influence the recovery of MERS CoV	(logistic regression) Naïve Bayes SVM J48	Estimate p-value	-
[8]	Analysing, diagnosing and predicting MERS-CoV	SVM Decision Tree *k*-NN classifiers	Accuracy using cross-validation model	86.44%
[13]	Analysing a plausible explanation of the public overreaction to MERS-CoV	Latent Dirichlet allocation Word2Vec method Natural language processing	-	-
[14]	Diagnosing patients with MERS-CoV through early syndromes	Naïve Bayes Random forest, SVM	Receiver operating characteristic (ROC)	Random Forest (ROC) = 0.942 Naïve Bayes (ROC) = 0.907 SVM (ROC) = 0.68
[15]	Extracting difference and similarity between SARS-CoV and MERS-CoV	Apriori algorithm Decision tree SVM	10-fold validation test	Higher than 75%
[16]	Predicting and preventing for MERS-CoV	Bayesian belief network Global Positioning System (GPS)-based risk assessment	True-positive (TP) False-positive (FP) rates ROC area	more than 80% ROC= 0.970

efficient method to cluster the collected datasets, as presented in Table 3. Consequently, detection and diagnosis can be remarkably enhanced. All studies in this review reported the use of AI techniques, such as case-based reasoning and rule-based systems. However, none of the studies utilised other classification methods, such as neural networks, reinforcement learning and hybrid classification. None of the studies utilised and integrated optimisation techniques, such as genetic algorithms and particle swarm optimisation, to their systems. Data mining and ML algorithms used in diagnostic operations primarily rely on classification algorithms, including decision tree, SVM and Naive Bayes classifiers (Figure 2 and Table 2). Notably, no study in the literature exploited clustering algorithms for the detection and diagnosis of the CoV family. This technique can be used as a pre-processing step before feeding data into the classification model to gain valuable insights into the case study data by understanding which groups of disease symptoms fall into when applying these algorithms.

In this review, the sample sizes representing the observations in each dataset are displayed in Table 3. In [11], the sample size includes all MERS-CoV patients in Saudi Arabia in the second half of 2016, and in [8], the sample size includes the cases from 2013 to 2017. The sample size in [4] represents 1082 records of cases reported from 2013 to 2015 distributed as 633 new case records, 231 recovery records and 218 death records. The study in [12] includes 836 patient records, and 52 patients are reported as dead and only 784 cases are used. In [13], the sample size is represented by articles collected from the Internet and reported by 153 news media outlets in Korea and the comments associated with these articles from day 1 (first confirmed case on May 20, 2015) to the day 70. In [15], the dataset contains 322 records, 92 infected cases and 230 uninfected cases. In [16], synthetic data are generated for 0.2 million users. Most of the available datasets are related to MERS and SARS infection cases, but no dataset is found for COVID-19 because of its novelty.

Analysis of characteristics

Four aspects are presented in the following subsections. Data were collected through an academic literature review of CoV. These aspects are provided as follows: the motivations of using different ML and data mining algorithms to explore meaningful algorithms for the CoV family; the potential challenges in predicting CoV infections in humans that should be overcome to realise their potential; recommendations to alleviate challenges related to ML algorithms for recovery from CoV infection; and case study dataset types of CoVs to assess

Table 3 Descriptions of CoV datasets with available sources

Ref.	Datasets descriptions	Available sources
[11]	-Dataset of patients affected by MERS-CoV in Saudi Arabia consisted of all cases in the second half of 2016.	Available on the Ministry of Health Control and Command Centre website [https://www.moh.gov.sa/Ministry/OpenData/Pages/default.aspx].
[4]	-A total of 1082 records of cases reported from 2013 to 2015. -A total of 633 new case records, 231 recovery records, and 218 death records, for a total of 1082 records.	Collected from the website of the Control and Command Centre of Saudi Ministry of Health [https://www.moh.gov.sa/Ministry/OpenData/Pages/default.aspx].
[12]	- The analysed data were collected from the Control and Command Centre. -A total of 836 patient records were used for analysis. Fifty-two patients out of the 836 cases were initially reported as dead. Hence, those cases were removed from the dataset, and 784 cases were used in the study.	Ministry of Health website of the Kingdom of Saudi Arabia.
[8]	The MERS-CoV dataset consisted of all reported cases in Saudi Arabia from 2013 to 2017.	N/A
[13]	-Articles collected from the Internet reported by 153 news media outlets in Korea and comments associated with these articles from day 1 (the first confirmed case on May 20, 2015) to the day 70 (the de facto end declared by the government on July 28, 2015), in addition to short-text comments on news articles in Twitter and Facebook.	[http://www.naver.com]
[14]	- A dataset was collected from UCI. A dataset containing 322 records, 92 infected cases and 230 uninfected cases was obtained. -Each record contained 24 attributes.	[https://www.nejm.org/doi/full/10.1056/nejmoa1306742] [https://www.sciencedirect.com/science/article/pii/S1473309913702044]
[15]	-SARS and MERS spike glycol protein data from the National Centre for Biotechnology Information database	[www.ncbi.nlm.nih.gov]
[16]	-Synthetic data were generated for 0.2 million users. -Raw information from users was collected using body-worn sensors and manually recorded data using the mobile application.	N/A

and analyse the effects of features and classes of this virus on each case study with ML algorithms.

Motivations

To date, research areas in the field of AI, such as data mining and ML technique-based applications have been rapidly developed because of their large impact on human life in terms of social, scientific, medical and engineering-based applications. Accordingly, this section presents the motivation of studies conducted on the CoV problem to save lives. Data mining for medical diagnosis systems is efficient and can be utilised to control the spread of MERS-CoV and protect humans [8]. Data mining can also be used to estimate and predict the recovery rates from CoV infections [4]. ML technique can be used to identify and predict the dominant factors affecting the recovery from MERS-CoV [12, 13]. Thus, studies conducted on identifying the best model can help minimise the effects of epidemic diseases [11]. The distinction between SARS and MERS viruses can be considered a challenging task because of the similarities in their symptoms, such as breathing problems and high fever [15]. Medical diagnosis systems play a significant role in identifying patient health conditions with MERS-CoV symptoms [14]. Studies, such as [16], integrated the GPS with their medical diagnosis systems to cluster the population of infected people based on their geographical area.

Challenges and limitations

COVID-19 has spread worldwide and threatened human life. Accordingly, several studies have been conducted to develop an intelligent medical diagnosis system using AI technique to control the effects of this virus. However, numerous challenges and research limitations have been indicated in the academic literature and need to be addressed in the future [8]. Some of these challenges are related to MERS-CoV nature and behaviour because understanding how the virus spreads and how people can be infected caused by the complexity of this epidemic disease is extremely difficult. The lack of a large dataset in the academic literature for MERS-CoV is considered a challenging task for AI researchers because it hinders the understanding of viral patterns and features [4, 12]. The demand to construct a dataset that can be understood

by ML algorithms has increased because the current dataset involves infographic data [11]. Other challenges are correlated with people and government responses to MERS-CoV that requires more new monitoring approaches and additional efforts compared with the traditional approach for controlling epidemic diseases [13]. Another challenge with MERS-CoV is the large variation in symptoms that are mostly similar to common cold symptoms, with many other variations of diseases that may occur in cases but not in others. Some patients have unique symptoms, and others have no symptoms at all. Activists have generated huge and complex volumes of data that render its analysis impractical and difficult to predict using linear classifiers [14, 15]. The protection of citizens by the government and health agencies is a significant challenge because no specific vaccine exists for this virus to date and requesting people to undergo medical checkups is difficult [16].

Recommendations

This study aimed to mitigate some of the challenges that have been addressed in the academic literature with their recommended solution for future studies. Studies, such as [8], suggested a pre-processing method to solve the missing-value problem that directly impacts the classification accuracy for the MERS virus. They also suggested the use of an ensemble technique by combining the cosine method with *k*-NN ML algorithm to improve the classification accuracy to >50%. Another study [4] recommended increasing the number of samples for the CoV dataset and collecting data from patients within the same geographical area by directly communicating with dedicated hospitals and health agencies [13]. [11] proposed the use of SVM classifier for the binary class problem and conducted an empirical study for multiclass problems in MERS-CoV. [17] recommended the use of the R language because of its efficiency and high functionality in supporting AI algorithms that can enable the development of an effective intelligent medical diagnosis system for CoV. Another study [14] indicated the special medical issues related to the female status, such as whether she is pregnant, which need to be considered in the treatment of CoV-infected patients [15]. [16] suggested the utilisation of the Internet of Things (IoT) technology for developing a highly dependable medical diagnosis system for COVID-19.

Case study, features and classes used

Based on the discussion, analysis and details in Tables 2 and 3, a case study found in the literature review can be divided into two types, namely, real and analysis datasets (Table 4). Real datasets consist of a number of real cases of infected and healthy patients within a specific period. Four studies provided real datasets for patients affected by CoV [4, 8, 11, 12], and most datasets in other studies are published and redistributed by the Ministry of Health website of the Kingdom of Saudi Arabia. In analysis datasets, standardisation is intended to increase the consistency of review and assessment analysis for MERS-CoV and SARS-CoV. An organised collection of data is found in four studies for the two viruses. In [13], the massive media outlet data were collected during the nationwide outbreak of MERS-CoV in Korea in 2015. In [14], MERS-CoV cases were recorded from several medical analytical papers focused on the early symptoms of this virus. In [15], spike glycol protein sequence data of SARS (DQ412574.1) and MERS (KP236092.1) were obtained. [16] utilised the GPS to represent each MERS-CoV user on Google maps where 5000 users are adopted in R Studio through the 'bnlearn' package.

Two attributes, namely, (1) personal patient information attributes and (2) CoV attributes, were recognised in the collected studies to explore the effects of CoV features and classes on case study datasets used with ML algorithms. As shown in Table 5, only four studies focused on 'Personal patient information attributes', only two studies considered 'CoV attributes', and two studies considered the two attributes.

Table 5 shows that the details mostly encompass MERS-CoV and SARS-CoV attributes and classes. Age attribute is considered the most important and dangerous factor in the infected patients because people over the age of 50 are more likely to be at risk and to have this type of virus than others [4, 8, 12, 14, 16]. Gender attribute is an important predictor in four out of eight studies [4, 8, 11, 12]. The city was used in three studies [4, 8, 11]. Other related attributes are less frequently associated with personal patient information attributes, including address, sex and name. Although SOB is mentioned in only two studies, it is considered the most important concern with MERS-CoV and SARS-CoV attributes because the two studies focused on patients affected by the two viruses.

From a specialised medical perspective, the features and classes of the CoV family are similar to one another. In this context, the new epidemic of COVID-19 depends on the same

Table 4 Case study types in the literature review

Ref.	Real dataset case study	Analysis dataset case study
[11]	√	
[4]	√	
[12]	√	
[8]	√	
[13]		√
[14]		√
[15]		√
[16]		√

Table 5 Features and classes used in the literature review

Ref.	Features and classes	
	Personal patient information attributes	CoV attributes
[11]	- Gender - City - The Probable source of infection class	N/A
[4]	- Gender - Age - Nationality - City - The Patient isa healthcare personnel or not	N/A
[12]	- Gender - Age - Healthcare worker or not - Symptoms - Status at time of identification of disease - Presence of pre-existing disease or not - Patient in contact with animal or not - Hospital - Household or community-acquired - The patient died or recovered	N/A
[8]	- Gender - Age - Exposure to camels - Comorbidities - Exposure to MERS-CoV cases - City - The patient is employed in healthcare or not - The patient is alive or dead.	N/A
[13]	N/A	- MERS epidemic - Mass media - Public emotion
[14]	- Age - Sex	- Fever - Fasting blood sugar - Heart disease - Chronic kidney - Chills - Dry - Productive - Shortness of breath (SOB) - Sore throat - Runny nose - Abnormal pain - Nausea - Vomiting, diarrhoea - Myalgia - Headache - Hypertension - Chronic lung - Obesity - Smoking - Chest pain
[15]	N/A	- Spike glycoprotein of MERS and SARS - Amino acid isoleucine - Asparagine
[16]	- Name - Address - Telephone numbers - Age - Sex - Occupation - GPS geographic location of the house - Names of relatives - Mobile numbers	- SOB - Cough - Fever integer body temperature in C - Acute respiratory distress syndrome - Consumptive coagulopathy - Consumptive coagulopathy - Symptom - Food exposure - Animal exposure - Infected human exposure - Risk area exposure

features and classes of MERS-CoV and SARS-CoV. Thus, extensive research has been conducted to prove a new AI pathway. As proof for the above discussion, several reliable reports and government news have mentioned that age is the most important feature of patients with COVID-19. Patients over the age of 50 are susceptible to contract the disease and be exposed to its risks and complications. The gender and city of an infected patient are the next concerns. The medical team has reported that SOB is the most important symptom attribute because it carries a high specificity for COVID-19.

Conclusion

This systematic review provided an exhaustive overview of integrated AI based on data mining and ML algorithms with the CoV family. State-of-the-art CoV prediction algorithms were presented. Distinct information, such as application nature, ML used, the evaluation conducted for each study, and extracted features and classes with accuracy percentages for the utilised ML algorithms were indicated. A set of propositions for the risk recovery of this virus was established to serve as a guide for future research in the context of data mining algorithms. Despite the increasing rates of death and the number of people affected with CoV, developments based on ML algorithms to improve CoV datasets remain at a redefinition stage, especially for COVID-19. The shortage of studies in the literature is a real concern and may have serious implications for detecting and minimising the spread of this virus. An emerging, rapidly evolving situation for the virus must be considered in the viewpoint of AI applications, and researchers should provide updated contributions because they are necessary. Supportive information, such as new datasets, must be provided, and many complex features and classes must be added. Close cooperation among researchers in the biomedical engineering field and the medical community is necessary to stop the growing public health threat posed by the 2019 CoV. The goal is to conduct new studies that can guide governments and communities in the early control of the impact of this virus by utilising the features and classes collected in the literature. The need for integrated sensor technologies specifically for outdoor scenarios is highly recommended. This process is only possible when the technique is interlinked with IoT technologies, focusing on evaluating and improving ML algorithms for CoV datasets with increased efficiency. In this context, AI software developers in healthcare can develop different software packages to remotely help analyse the extracted features and classes for patients with COVID-19.

Compliance with ethical standards

Conflict of interest The authors declare no conflict of interest.

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent Informed consent was obtained from all individual participants included in the study

References

1. Centers for Disease Control and Prevention, "Coronavirus Disease 2019 (COVID-19) Situation Summary," *Published 2020*. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/summary.html#background.
2. D. Wu, T. Wu, Q. Liu, and Z. Yang, "The SARS-CoV-2 outbreak: what we know," *Int. J. Infect. Dis.*, 2020.
3. Y. Fan, K. Zhao, Z. L. Shi, and P. Zhou, "Bat coronaviruses in China," *Viruses*, vol. 11, no. 3, p. 210, 2019.
4. I. Al-Turaiki, M. Alshahrani, and T. Almutairi, "Building predictive models for MERS-CoV infections using data mining techniques," *J. Infect. Public Health*, vol. 9, no. 6, pp. 744–748, 2016.
5. C. Drosten *et al.*, "Identification of a novel coronavirus in patients with severe acute respiratory syndrome," *N. Engl. J. Med.*, vol. 348, no. 20, pp. 1967–1976, 2003.
6. P. Zhou *et al.*, "Fatal swine acute diarrhoea syndrome caused by an HKU2-related coronavirus of bat origin," *Nature*, vol. 556, no. 7700, pp. 255–259, 2018.
7. K. H. Yu, A. L. Beam, and I. S. Kohane, "Artificial intelligence in healthcare," *Nat. Biomed. Eng.*, vol. 2, no. 10, pp. 719–731, 2018.
8. A. AlMoammar, L. AlHenaki, and H. Kurdi, "Selecting accurate classifier models for a MERS-CoV dataset," *Adv. Intell. Syst. Comput.*, vol. 868, pp. 1070–1084, 2018.
9. J. B. Long and J. M. Ehrenfeld, "The Role of Augmented Intelligence (AI) in Detecting and Preventing the Spread of Novel Coronavirus," *J. Med. Syst.*, vol. 44, no. 3, p. 59, 2020.
10. D. Moher, A. Liberati, J. Tetzlaff, and D. G. Altman, "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement," *Int. J. Surg.*, vol. 8, no. 5, pp. 336–341, 2010.
11. N. Almansour, H. Kurdia, N. Almansour, and H. Kurdia, "Identifying accurate classifier models for a text-based MERS-CoV dataset," in *2017 Intelligent Systems Conference, IntelliSys 2017*, 2018, vol. 2018-Janua, pp. 430–435.
12. M. John and H. Shaiba, "Main factors influencing recovery in MERS Co-V patients using machine learning," *J. Infect. Public Health*, vol. 12, no. 5, pp. 700–704, 2019.
13. S. Choi, J. Lee, M. G. Kang, H. Min, Y. S. Chang, and S. Yoon, "Large-scale machine learning of media outlets for understanding public reactions to nation-wide viral infection outbreaks," *Methods*, vol. 129, pp. 50–59, 2017.
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," in *Communication, Management and Information Technology - Proceedings of the International Conference on Communication, Management and Information Technology, ICCMIT 2016*, 2017, pp. 633–638.
15. 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," in *MATEC Web Conf.*, 2016, vol. 49, p. 8001.
16. R. Sandhu, S. K. Sood, and G. Kaur, "An intelligent system for predicting and preventing MERS-CoV infection outbreak," *J. Supercomput.*, vol. 72, no. 8, pp. 3033–3056, 2016.

17. M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, 2009.

Affiliations

A. S. Albahri[2] · Rula A. Hamid[3] · Jwan k. Alwan[4] · Z.T. Al-qays[5] · A. A. Zaidan[1] · B. B. Zaidan[1] · A O. S. Albahri[1] · A. H. AlAmoodi[1] · Jamal Mawlood Khlaf[6] · E. M. Almahdi[7] · Eman Thabet[8] · Suha M. Hadi[2] · K I. Mohammed[1] · M. A. Alsalem[9] · Jameel R. Al-Obaidi[10] · H.T. Madhloom[11]

[1] Department of Computing, FSKIK, Universiti Pendidikan Sultan Idris, Tanjong Malim, Malaysia

[2] Informatics Institute for Postgraduate Studies (IIPS), Iraqi Commission for Computers and Informatics (ICCI), Baghdad, Iraq

[3] College of Business Informatics, University of Information Technology and Communications (UOITC), Baghdad, Iraq

[4] Biomedical Informatics College/University of Information Technology and Communications (UOITC), Baghdad, Iraq

[5] Department of Computer Science, Computer Science and Mathematics College, Tikrit University, Tikrit, Iraq

[6] Respiratory Center, Army Force Hospital, Baghdad, Iraq

[7] General Secretariat for the Council of Ministers (GSCOM), Baghdad, Iraq

[8] Department of Computer Science, College of Education for Pure Sciences, University of Basra, Basra, Iraq

[9] Department of Management Information System, College of Administration and Economic, University of Mosul, Mosul, Iraq

[10] Department of Biology, Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris, Tanjong Malim, Iraq

[11] Information Technology Department, College of Applied Sciences, Ministry of Higher Education, Muscat, Iraq