Periodicals of Engineering and Natural Sciences ISSN 2303-4521
Vol.9, No.2, May 2021, pp.918-929

Automatic extraction of knowledge for diagnosing COVID-19 disease based on text mining techniques: A systematic review

Amir Yasseen Mahdi [1], Siti Sophiayati Yuhaniz[2]
[1]Computer sciences and mathematics college, University of Thi_Qar, Thi_Qar, Iraq
[1,2]Razak Faculty of Technology and Informatics, UniversitiTeknologi Malaysia, Kuala Lumpur, Malaysia

ABSTRACT

In late December 2019, an epidemic of the novel coronavirus (COVID-19) was informed, and because of the quick diffusion of the infection in various regins of the world, the World Health Organization proclaimed an emergency. In this context, researchers are urged and encouraged to research in various fields, to stop the spread of this deadly virus. To this end, we propose a systematic review that addresses the techniques and methods of artificial intelligence in diagnosing COVID-19 disease. The main aim of the current systematic review was to highlight the gaps and challenges within the academic literature of the disease COVID-19, which included the characteristics of the data, machine learning algorithms applied to the diagnosis of COVID-19, and using natural language processing (NLP)to reveal clinical data for COVID-19 disease.Seven reliable databases were used, namely Web of Science, ScienceDirect, IEEE Xplore, Scopus, PubMed, springer and google scholar, to obtain studies related to the specific topic many filtering and surveying stages were conducted consistent with the inclusion and exclusion criteria, to screen the acquired 1115 papers.We identified the bottleneck in explaining data as one of the major barriers to machine learning and NLP approaches. Supervised machine learning has been explored as an active method for diagnosing COVID-19 disease. Future studies in this area will benefit from alternatives like increasing the volume of data, using intelligence swarms to obtain accurate features, and using unsupervised learning that does not require explanatory data. Thus, this research supported us to get a more practical comprehension of the gaps and provide possible solutions for filling these gaps.

Keywords: COVID-19, Text mining, NLP, Swarms intelligence, Machine learning.

Corresponding Author:

Amir Yasseen Mahdi
Computer sciences and mathematics college, University of Thi_Qar
Thi_Qar, Iraq
E-mail: amiryasseen@utq.edu.iq, mahdi.amir@graduate.utm.my

1. Introduction

In late December 2019, group of patients identified with pneumonia by several local health facilities for obscure cause that were epidemiologically linked to wholesale market of wet animal and a seafood in Wuhan, Hubei Territory, China [1]. In January 2020, a pathogen known as the 2019 novel coronavirus (2019-nCoV) was successfully isolated [2]. 2019-nCoV is a different clade from the beta coronaviruses linked to human extreme severe respiratory syndrome (SARS) and Middle East respiratory syndrome (MERS), according to full-genome sequencing and phylogenic analysis [3]. In China, 51,857 patients with COVID-2019 had been registered as of February 16, 2020, with 1,121 deaths confirmed. The coronavirus disease (COVID-19) infection epidemic that began in Wuhan, Hubei regione, has distributed to other Chinese regions [4]. According to Z. Hu et al [4], COVID-19 symptoms include fever, nasal congestion, cough, fatigue, dizziness, arthralgia, etc. In the battle against COVID-19, artificial intelligence (AI) plays a key supporting role and can lead to solutions faster than we would otherwise achieve in many areas and applications [5]. Through its decision-making, artificial intelligence has displayed hopeful health care outcomes by analyzing the data. The detection of disease with the assisstance of different AI methods may be one of the solutions for handling the current havoc [6]. The White House issued a call to action on March 16, 2020, for global artificial intelligence (AI) researchers to collaborate

with research institutes and technology companies to develop new text and data mining techniques to help COVID-19-related research [7].

Computer and data mining techniques can aid in disease identification, diagnosis, and prediction, as well as virus infection monitoring [8]. SARS-CoV-2 has caused a coronavirus disease (COVID-19), which is an extremely pathogenic viral contagion. It is also currently causing global health concerns [9]. Thus, extra to the diagnostic accuracy and speed, the rapid development of advanced artificial intelligence and machine learning-based diagnostic systems can also protect health care workers by reducing their contact with COVID-19[7].To aid public health and make better healthier choices to aid in the diagnosis and prediction process, clinical records have proven to be a significant source for disease-related information [10]. According to [11], studies into clinical and paraclinical features of COVID-19 could indicate an essential approach for proper patient administration. Even so, A single textual data source base has been used by many previous clinical text mining applications, such as radiology reports, to classify or mine information. On the other contrary, increasing data linking in Hospital Information Systems, however, creates opportunities for stronger and more precise text mining techniques for the benefit of information from multiple data sources [10]. Consequently, the key purpose and contribution of this research is to stratify the principle of processing natural language (NLP), swarm's intelligence and ML techniques to the problem of extract information related to this disease automatically and classifying COVID-19 cases as positive or negative, even though the disease is still in its early stages. The major contributions of the present study are mentioned below:

- We show a systematic framework based on text mining approaches and NLP that is able to extracting meaningful information from COVID- 19–related clinical records.
- We suggest an automatic learning model based on swarm's intelligence and ML for classification of COVID-19–related data to multi cases, which generate better results compared to many other well-known methods.

2. Methods

This research adopted the literature review style on the bases of systematic literature survay approach, which has been recognized for its function in obtaining an adequate understanding regarding an interesting topic [12].

2.1. Data Sources

We systematically searched in several digital databases, such as Scopus, Web of Science (WOS), ScienceDirect (SD), PubMed, IEEE Xplore, Google Scholar, and springer from the last decades (from 2015 approximately) up to the present. These databases are considered adequate to include the most and latest reliable literature to understand the role of AI in extracting and mining clinical data for the diagnosis and detection of COVID-19.

2.2. Search strategy

Boolean operators have been used to extract a large number of related literatures:(coronavirus OR coronaviridae OR SARS-CoV-2 OR 2019-nCoV OR COVID-19) AND (detection system OR diagnostic system OR diagnosis system OR diagnosis implementation OR diagnostic implementation) AND (text OR unstructured OR natural language processing); (text OR unstructured) AND (clinical information OR information extraction OR information discovery OR automatic extraction OR knowledge discovery OR health information OR knowledge).There was a set of keywords that were adopted in extracting the relevant literature, which included words dedicated to automatically extracting information from medical texts, words to search for clinical texts of COVID-19 disease, words to diagnose and detect COVID-19 disease based on clinical data to ensure that all the literature related to the research questions is contained.

2.3. Eligibility criteria

For selection of the relevant literature, different criteria were imposed in the systematic literature review [13]. Several criteria were used to exclude articles in order to improve the scope, included the following: If they were not related to the extraction of information from the medical texts, the clinical data were not used in the diagnosis of human diseases, they were not written in the English language. Regarding the topic of interest, the SLR only selected publications related to the research questions that discuss the role of AI in diagnosing and detecting COVID-19 using patient clinical records. Moreover, this COVID-19 study discussed the role of AI in using algorithms to mine various textual clinical data by natural language processing and ML algorithms in the ranking

and prognosis process. This investigaton extracted the application, methodology, case study, type of data, and the state of accuracy for each research study in the literature. Extracts, limitations, challenges, and recommendations were made from papers reviewed to address and assist public health, in controlling on COVID-19 disease.

2.4. Study selection

This process began with were scanned of the article titles and abstracts it. Duplicate articles have been removed, and a number of articles excluded because they failed to meet our inclusion standards. The remaining articles were subjected to a further examination by reading the entire text in order to be included in the appropriate group of selected articles, extracting the research data and building the systematic review article. After these procedures, only 8 articles met all the criteria and were deemed appropriate to the research questions in this review. The research process was fully controlled and supervised in all research papers by a senior author to asure that a very useful and authoritative research paper was created.

2.5. Result

Figure 1 shows the findings of the search requists in the current investigtion that generated 1115 articles from all the databases mentioned in Section 2.2. The number of duplicate articles was 91 and the results were 1,024. We included 201 articles that have met the requirements for inclusion and exclusion. After reading and reviewing the articles, the result was only 8 studies that fulfilled the research questions and the inquiry that was conducted, which included the presence of basic elements that identify COVID-19 disease, and use of clinical data to diagnose the disease using the machine learning in this systematic review.

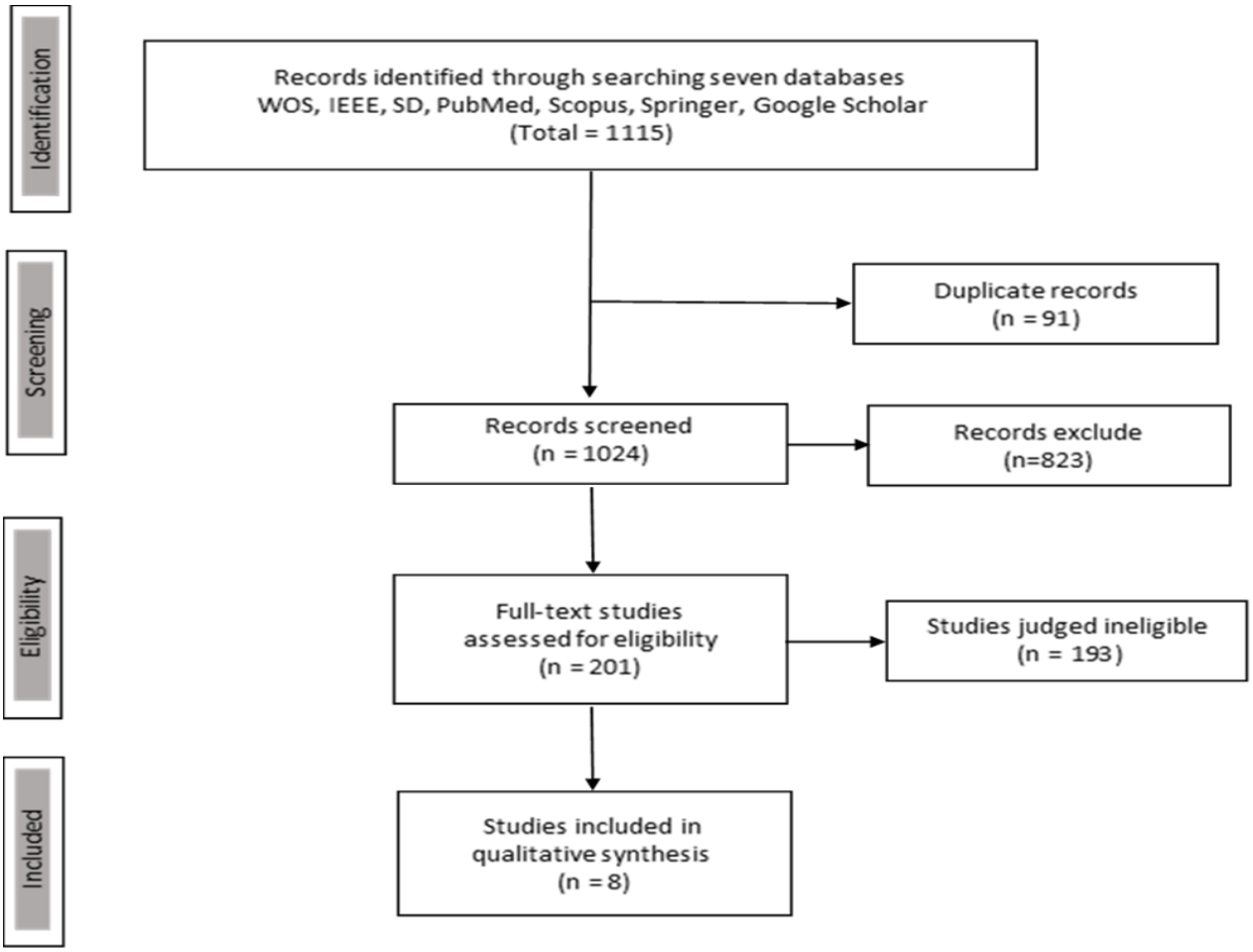

Figure1. Article selection flow chart

3. Distribution results

The research results obtained from the databases showed that previous researchers use different algorithms and methods in the process of obtaining features extraction and data classification. Figure 2 summarizes most of the machine learning algorithms used in classification. The figure 3 summarizes the methods and techniques that were used to obtain features extraction. Logistic Regression was the most used (six times). Support vector machine (SVM) algorithms were used five time. Each of the Random forest and Decision tree algorithms have been used four times. XGBoost and KNN algorithm were used three times. Multi-Layer Perceptron algorithm was used twice. Multinomial Naïve Bayesian, Bagging, Adaboost, Stochastic gradient boosting, neural networks, gradient boosting tree and extremely randomized trees one time. As for the methods and methods that were used to extract the features, Information gain, Gini index, Chi-Squared statistics, ontology learning, TF / IDF, Bag of words were used once. On the other hand, only two papers were used for the NLP method, as in the Table 1 Illustrates objective, framework and results for all studies. The papers included in this review were published in a year 2020. Table 2 and figure 4 shows the COVID-19 prediction algorithms and the accuracy obtained for each algorithm. Table 3 and 4 provides a full description of each set of COVID-19 data, with available sources. In addition, Table 5 that show the features and classes used in this systematic review.

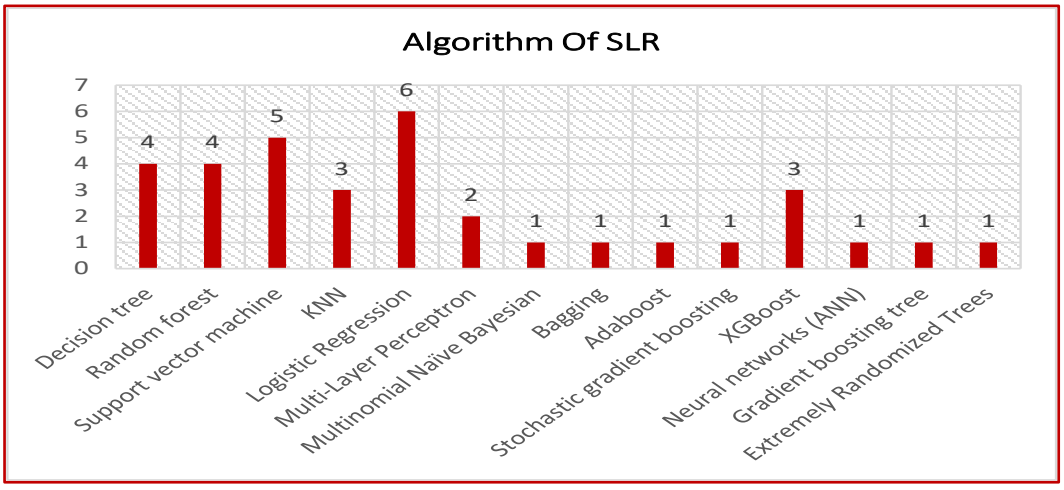

Figure 2. Summary of algorithms used in the systematic literature review

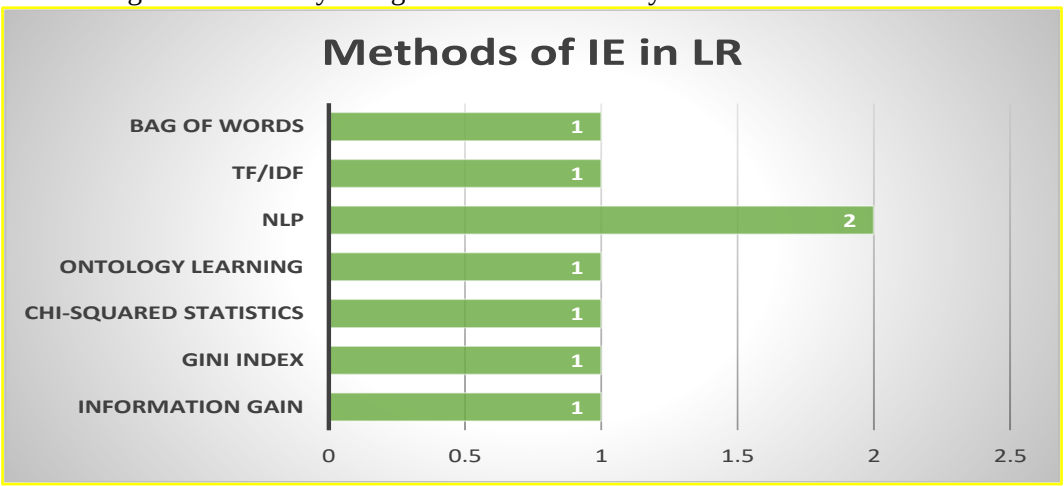

Figure 3. Summary of methods used in the literature review

4. Discussion

In [14], five ML algorithms were applied to the COVID-19 dataset to predict who will develop syndrome of acute respiratory distress based on the riskiness of the COVID-19 disease. The findings displayed that the support vector machines and KNN classifier is the best accuracy in prediction in than the other models, and the logistic regression are the less models for predictive. additionally, the results indicated that salient features for prediction were liver enzyme, red blood cells and body aches. In [15], text mining based on CBR and NLP, and

Semantic Web ontology was used for feature extraction to develop framework with a high accuracy to early diagnosis COVID-19 as positive or negative. In [16], three ML algorithms were applied to the COVID-19 dataset collected by an online questionnaire as an input to the predict patient's potential of COVID-19 depended on indications and symptoms. The results showed that the MLP has the best precision (91.62%) in prediction. Meanwhile, the SVM has shown the best precision (91.67%.). In [6], data mining based on natural language processing has been used. TF / IDF and Word Bag were used to extract features for use with machine learning algorithms. Nine ML algorithms were applied and Classifying the clinical data set into four groups named COVID-19, SARS, ARDS, and both (ARDS, COVID-19). The results showed that the Multinomial Naïve Bayes and Logistic regression has better findings than other ML algorithms which represents 96.2% accuracy. In [17], three ML algorithms were applied to diagnose COVID-19. The findings displayed that hat MLP, LR and XGBoost can surely classify COVID-19 patients found in a dataset. However, the author states that, the efficiency of the classifiers must be verified on a larger and more reliable dataset because this dataset contains anonymized samples. In [18], five ML algorithms were trained in order to identify COVID-19. The results showed that support vector machines algorithm has best performance (AUC: 0.85; Specificity: 0.85; Sensitivity: 0.68; Brier Score: 0.16) Compared to other algorithms. The experimental results indicated that lymphocytes, eosinophils and leukocytes are three most significant variables for algorithm of the predictive model implementation. In [19], XGBoost machine learning algorithm are applied to the COVID-19 infection electronic records to predict the survival or death based on criticality in patients and the classifier shows 90% accuracy. The experimental results indicated that lactic dehydrogenase, High-sensitivity C-reactive protein, and lymphocyte are the controlling characterstics for the prediction model and that patients are likely to survive or death.[20], two experiments are applied to the COVID-19 disease dataset, and the Random Forest classifier displays higher predictability than the other models. Furthermore, as a clear decision-making tool for clinicians interpreting blood tests for COVID-19 suspect cases, an interpretable Decision Tree model was developed. The findings suggested that hematochemical blood test values (namely: platelets and white blood cell counts, AST, ALT, CRP, GGT, plasma LDH, ALP) were used to for identifying COVID-19 positive patients.

Table 1. Objective, frameworks and outcome used in the included publications

Papers	Objective	Framework	Outcome
[14]	To Acute respiratory distress syndrome (ARDS) predict for patients with coronavirus disease infection (COVID-19)	ML+ Filter methods	P-ARDS/N- ARDS
[15]	To classifying cases of COVID-19	NLP+OWL	Positive/Negative
[16]	To predict potential patients of COVID-19	ML	Positive/Negative
[6]	To diagnosis of coronavirus disease	NLP+ML	Four classes/ SARS, ARDS and Both (ARDS, COVID-19)
[17]	To diagnose COVID-19	ML	Positive/Negative
[18]	Predicting the danger of a positive diagnosis of COVID-19 by ML	ML	Positive/Negative
[19]	To predict the survival of patients infected with COVID-19	ML	death/or Not
[20]	Discrimination between patients who are either negative or positive infection to COVID-19	ML	Positive/Negative

5. Critical analysis

On the basis of the review of previous literature, not many contributions were obtained from papers related to the use of AI techniques and approaches in the mining and detection of COVID-19 disease. This limited number of studies on COVID-19 disease, however, indicates the need and opportunity for artificial intelligence to be applied in this field.

The most significant gap in all previous studies is the lack of an adequate and comprehensive dataset to train supervised machine learning algorithms, as show Table 3. In addition, the clinical data sets used in the academic

studies were not from multiple centers and thus it will be likely that not all clinical attributes are declared in these data sets.

Another challenge is how to choose the best classification algorithm that will provide an accurate diagnosis of COVID-19 disease. where, in the (Table 2 and figure 4) displays the results of the general performance of all algorithms, we note that it is difficult to confirm one of the algorithms over the other, and the reason is the use of a small database and varies in the number of features (see Table 5), and this is the other research gap.On the other hand, not all of the patient's parameters or indicators were exploited in the diagnostic process, and according to [21], it requires data describing a wider clinical spectrum, especially when the viral load is low the detection rate is also low, as a result false negative findings occure. This is the third gap in previous studies. Previous studies, in particular the two studies [6] and [15], did not exploit the extraction of information in building a knowledge base from clinical facts that would be useful in clinical decision support systems and information management systems. Moreover, all research has focused on one result in the patient's diagnosis, which is either positive or negative (see Table 1), but this method is ineffective when the symptoms are insufficient and not very clear. Consequently, we need multiple diagnoses with the virus such as severe, moderate, low, and non-existent or previously infected and recovered, and this will be useful from two sides: first, it is not to exploit health resources for all cases, and secondly, the treatment protocol will be variable according to the infection case. In addition, all previous studies did not address the interaction between the treatments used in treating COVID-19 disease and its fatal effect on the patient. It is noteworthy that NLP has not been used greatly and effectively to process and extract information related to COVID-19 disease from clinical texts of hospitalized inpatients, that contain real and rich information. Moreover, none of the previous studies in this review used swarm's intelligence techniques, such as the ACO, BCO, PSO and CAO, to improve the feature problem and extract the most useful information especially when the data volume is large there will be features noisy and unrelated, that reduce the accuracy of the classification model. Therefore, more characteristic engineering is required to achieve better findings.

Table 2. Comparison performance of the ML algorithms

Techniques	Accuracy	Precision	Specificity	Sensitivity	F-score	Recall
Ref:[14]						
LR	50	-	-	-	-	-
KNN	80	-	-	-	-	-
DT	70	-	-	-	-	-
RF	70	-	-	-	-	-
SVM	80	-	-	-	-	-
Ref:[15]						
CBR	94.54	0.96	0.50	0.98	0.97	0.98
Ref:[16]						
LR	85.00	66.67	78.57	100.00	-	-
SVM	90.00	91.67	87.50	91.67	-	-
MLP	91.62	90.00	93.75	87.80	-	-
Ref:[6]						
LR	96.2	0.94	-	-	0.95	0.96
MNB	96.2	0.94	-	-	0.95	0.96
SVM	90.6	0.82	-	-	0.86	0.91
DT	92.5	0.92	-	-	0.92	0.92
RF	94.3	0.93	-	-	0.93	0.94
Ref:[17]						
MLP	93.13	93	-	-	93%	93
LR	92.12	92	-	-	92%	93

XGBoost	91.57	92	-	-	92%	92
KNN	88.91	89	-	-	89%	89
DT	86.71	87	-	-	87%	87
Ref:[18]						
SVM	-	0.778	0.850	0.677	0.724	-
RF	-	0.778	0.850	0.677	0.724	-
NN	-	0.742	0.800	0.742	0.742	-
LR	-	0.767	0.825	0.742	0.754	-
GBT	-	0.758	0.800	0.806	0.171	-
Ref:[19]						
XGBoost	0.93	1.00	-	-	0.91	0.83
Ref:[20]						
LR	82	83	65%,	92	-	-
RF	86	86	75	95	-	-

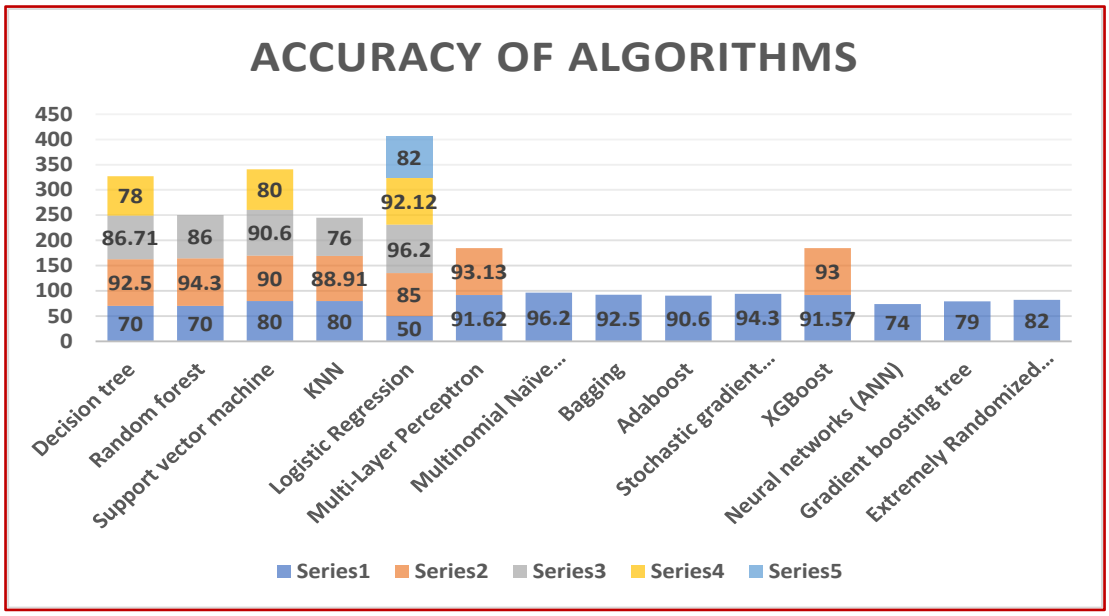

Figure 4. The accuracy algorithms of ML in SLR

6. Motivation

AI could be a device in the fight against COVID-19 and related epidemics [22]. Besides, automated diagnostic systems based on machine learning and artificial intelligence can increase the speed and accuracy of diagnosis for patients with COVID-19 and identify those most at risk of developing the disease [7][23]. Thus, quick and precise diagnosis of COVID-19 can save lives, limited disease spread, and produce data to train artificial intelligence models [22]. Studies, such as [24], In order to group the population of infected people depending on their geographical location, the GPS was integrated with their medical diagnosis systems. Moreover, due to the rapid spread of the infection, it makes the clinical data and texts for COVID-19 disease is diverse in terms of complexity and volume, largely unstructured and built in natural human languages. As a result, in a crisis like this, mobilization and the preservation of medical, logistical, and human capital are needed [25]. Therefore, in order to understand the data, automated information extraction techniques are needed, clinical notes analysis, clinical relation extraction and extract useful information for the end users in order to arrive at diagnostic reasoning. Additionally, building and discovering the knowledge resulting from the exploration of clinical texts

is also a major issue in order to infer additional facts from the patient's clinical data because of their great importance in clinical decision support systems and to retrieve useful information [26], [27].

7. Challenge and limitation

Some of these difficulties are linked to, limitations constraints of care services and hospital beds, in which physicians are forced to make challenging decisions to direct clinical acumen without past specific experience [14]. Furthermore, the size of the dataset, with some incomplete results, is a strong limitation of all studies in addition to the limited range of data indicating the severity of the patient, where most clinical outcomes have not yet been released [6] [14-20]. According to [15], not patient HER-based COVID-19 benchmarked datasets was accessible. Thus, the challenge was to detect the semantics / meaning and context of the use of patient data characteristics. Furthermore, choosing the best model for the computational similarity of cases is an issue that needs an optimal solution, given the sensitivity of medical cases. Another challenge worldwide is that patients cannot be examined quickly according to the restricted availability of PCR tests and other COVID-19 test, which can have dire implications in terms of proper clinical treatment [16][18]. On the other hand, their limitation it considers the cuuracy of the ground truth itself. According to a recent study, challenges like inappropriate procedures for collecting, handling, transporting, and storing swabs, the existence of intervening substances, and sample contamination, among others, may significantly affect the accuracy of this test (rRt-PCR) [28]. Also, more characteristics engineering is required for identify risk and better findings in order use deep learning approach in future [6][14]. On the other hand, it is difficult to manually identify patients affected by COVID-19 from the infectious crowd. Hence, assisting machine learning methods has become an urgent but challenging task to identify critical cases using clinical records. Unfortunately, no available prognostic biomarker is currently available so far to identify cases that are at imminent danger of mortality, in order immediate medical attention [19]. Other challenge, the similarity between the most popular symptoms of COVID-19 with previous coronaviruses and other infectious diseases, such as fever and cough. On the other hand, the ne rise in the use of modern fast diagnostic tests can increase the risk of wasteful health resource allocation, making immediate diagnosis difficult for health professionals who are vulnerable to certain accuracy issues[18-20]. Finally, the challange of distiguishing between COVID-19 positive and negative individuals since the COVID-19 symptomatology, which not show on large number of patients [20].

Table 3. Descriptions of COVID-19 datasets with available sources

Ref	Datasets descriptions	Available sources
[14]	• Of the 53 patients with COVID-19, all tested (RT-PCR). • range 13-67 years	**N/A**
[15]	71 (4 pediatrics and 67 adults) cases	Italian Society of Medical and Interventional Radiology
[16]	120 online survey questionnaire (15 rejected) accepted sample 105	**N/A**
[6]	212 clinical report	https://github.com/Akibkhanday/
[17]	• 5644 data set given by Israelita Albert Einstein Hospital of Brazil • only1091 records and 61 columns accept after remove records with mostly null values	**N/A**
[18]	• 235 adult patients from the Hospital Brazil • from 17 to 30 of March 2020 • The sample was well balanced (51.1% percent and 48.9% percent) between males and females, • Of these, 102 (43%) obtained a positive COVID-19 diagnosis from RT-PCR exam. • with an average age of 49	**N/A**
[19]	• Blood sample database of 404 infected patients in the Wuhan Area • Of the remaining 404 patients, 213 recovered from the infection, while the other 191 died	**N/A**

	• Data originating from patients younger than the age 18, breast-feeding women and pregnant, and at least 80% of data recordings, were excluded • 58.7 precent of males,375 patients were 58.83 ± 16.46 years old • Test and Training datasets all together considered 375 patients, and 3 features.	
[20]	• Collected 279 patients with COVID-19 symptoms from Hospital (Milan, Italy) • the rRT-PCR test was screened with and 177 positives, while 102 received a negative response.	**N/A**

8. Recommendation

The aim of the present investigation was to address some of the problems and limitations that have been addressed in the academic literature with suggested and recommended solutions for future studies. All study indicated the efficiency of classifiers can be improved by dataset larger and more reliable [6] [14-20].[14] Recommended developed AI tools to be clinically applicable. They also indicated that increasing the dataset, from different the severity spectrums, would improve the model's predictive ability and make it a valuable tool to distinguish early from the many with COVID-19.[15] proposed the use of SVM classifier as hybridize method with other methods. Another study [16] recommended the instrument these models of classification in hospitals in with aim of identifying COVID-19 patients in a quick, safe method, in order to a reduction spread of rapid to the virus. Another study [6] indicated, the disease can be diagnosis on the basis of gender like a way that we can gain information about whether females is influenced more or male. They also suggested increase feature engineering is required for best findings and in future can be utilized deep learning [19]. Recommended the further studies are needed to consider more clinical confounding factors relevant COVID-19 disease. Another study [17] has shown that it is suspected that multiple forms of COVID-19 outbreaks will continue to occur in animals. Therefore, constant research is needed in the study of present and future coronaviruses. [20] recommended to the inclusion of more hematochemical parameters, and the cases whose probability to be COVID-positive is almost 100% by two or more positive swabs, to evaluate the ability of the classification model to predict the positivity of COVID-19, and improve the sensitivity further.

Table 4. Summary of the perspectives of works and case study.

Ref	Case study types		AI techniques		Type of datasets
	Primary data	Secondary data	ML	NLP	
[14]	√		√		Clinical data
[15]	×	√		√	Clinical data/textual
[16]	√	×	√	×	Questionnaire clinical data
[6]	×	√	√	√	Clinical data/textual
[17]	√	×	√	×	Clinical data
[18]	√	×	√	×	Clinical data
[19]	√	×	√	×	Electronic records
[20]	√	×	√	×	Clinical data

9. Case study

Previous academic literature has been discussed on the basis of different perspectives in terms of aspects related to the data set and the methods and mechanisms of artificial intelligence, in this section we will highlight the case study. Based on the details in Tables 3 and 4, the case study that was used to diagnose COVID-19 disease can be classified as primary and secondary. The primary data set is the real data set that was gathered through the search from the real cases of patients affected by COVID-19 disease in hospitals, on the contrary, the

secondary data set was obtained online, and the researchers published it to help researchers for apply their experiments. When examining the main characteristics of the data used, it was found that the training data set tends to be relatively small. In addition to its small size, the training data is usually obtained from a few and not various institutions (no multicenter). On the other hand, the vast majority of the literature has focused on specific features and categories in the detection and diagnosis of COVID-19 disease, as show Table 5. An unorganized collection of data in form textual is found in two studies for the COVID-19 disease [6][15]. In [16], the data were collected by online survey questionnaire during the outbreak of COVID-19 in Jordan. In [14] [17-20], COVID-19 cases were registered from many hospital analytical papers emphasised on the test, early symptoms and image reports of this virus. As offered in Table 5, only three investigations concentrated on three different features of each study, only one study considered blood tests as diagnostic attributes for COVID-19 disease, while the other studies focused on attributes ranging from 10 to 24 basic attributes in the diagnosis. Only in study [15] mostly included features and categories from 71 cases reported from the archive of the Italian Society of Medical and Interventional Radiology (SIRM) repository. Moreover, since COVID-19 can interfere with other diseases as we mentioned previously, comprehensive data and large reliable sample are of great interest, because extracting features related to COVID-19 disease has a great effect on ranking in terms of developing accuracy and reducing the error rate. Consequently, all the scenarios mentioned in this discussed literature will have a significant impact on the classification results of COVID-19 disease. Therefore, generating a large number of clinical data is required in order to provide comprehensive and accurate training.

Table 5. Summary Features and classes utilized in the SLR

Ref	Features and classes
[14]	Three features:(ALT) (a liver enzyme), the presence of aches body, and an elevated the red blood cells
[15]	Epidemiological+Symptom+Exposure/Travel History (Spatial/ Location) Comorbidity (diseases)+Laboratory Tests+TreatmentRadiological
[16]	signs and symptoms (13 influential features)
[6]	24 features
[17]	10 influential features
[18]	15 influential features
[19]	300 features, while only Three influential features namely: Lymphocyte and High-sensitivity C-reactive protein (hs-CRP) lactic dehydrogenase (LDH).
[20]	13 attributes, while the influential features namely: white blood cells count, and the platelets, CRP, AST, ALT, GGT, ALP, LDH plasma levels)

10. Conclusion

The current outbreak of COVID-19 disease has a great effect on the lives of population all over the world, and the number of those affected by this infection has increased to a very large number. Therefore, all countries of the world and scientists are trying to control this crisis, as many different medical tests and many studies in the science and technology sector have been used to help detect and identify COVID-19 cases. This systematic review focused on a comprehensive survey of the literature that is consistent with the research questions based on the function of artificial intelligence in extracting information from medical texts using NLP and machine learning methods in the detection and identification of COVID-19 disease. Distinct information was indicated, such as extracting information and gaining knowledge from clinical data, is a prerequisite for society to find effective and robust discoveries of COVID-19 disease cases in the fastest time so that patients receive appropriate care in the shortest time possible. Although the information extraction approach is necessary for clinical applications, it has not been used effectively in diagnosing COVID-19 infection, nor did NLP tasks have broad clinical applications in this review. On the other hand, this study referred to developments in machine learning with a comprehensive dataset and reliable. Moreover, intelligence swarms and deep learning algorithms were not used, and this is due to the limited availability of clinical data for researchers due to privacy law and institutional concerns to access electronic health records. Therefore, joint efforts are needed to issue clinical data and encourage researchers to contribute to automatic information extraction research and the use of clinical

NLP technology in order to assist health care in controlling this pandemic (COVID-19). There are some limitations in this review. First, the review is restricted to articles written in English, and likely articles written in other languages would also provide useful information. Second, the databases identified in this review may not be adequate and may introduce bias in this review. Third, articles that use medical images to diagnose COVID-19 disease were not considered in this review. Fourth, given the speed with which more studies appear in the short term because it is an emerging and new field, it is likely that there will be new applications and techniques on COVID-19 disease.

References

[1] N. Zhu et al., "A novel coronavirus from patients with pneumonia in China, 2019," N. Engl. J. Med., vol. 382, no. 8, pp. 727–733, 2020, doi: 10.1056/NEJMoa2001017.

[2] K. Liu et al., "Clinical characteristics of novel coronavirus cases in tertiary hospitals in Hubei Province," Chin. Med. J. (Engl)., vol. 133, no. 9, pp. 1025–1031, 2020, doi: 10.1097/CM9.0000000000000744.

[3] D. Wang et al., "Clinical Characteristics of 138 Hospitalized Patients with 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China," JAMA - J. Am. Med. Assoc., vol. 323, no. 11, pp. 1061–1069, 2020, doi: 10.1001/jama.2020.1585.

[4] Z. Hu et al., "Clinical characteristics of 24 asymptomatic infections with COVID-19 screened among close contacts in Nanjing, China," Sci. China Life Sci., vol. 63, no. 5, pp. 706–711, 2020, doi: 10.1007/s11427-020-1661-4.

[5] J. C. Sipior, "Considerations for Development and Use of AI in Response to COVID-19," Int. J. Inf. Manage., p. 102170, 2020, doi: 10.1016/j.ijinfomgt.2020.102170.

[6] A. M. U. D. Khanday, S. T. Rabani, Q. R. Khan, N. Rouf, and M. MohiUd Din, "Machine learning based approaches for detecting COVID-19 using clinical text data," Int. J. Inf. Technol., 2020, doi: 10.1007/s41870-020-00495-9.

[7] A. Alimadadi, S. Aryal, I. Manandhar, P. B. Munroe, B. Joe, and X. Cheng, "Artificial intelligence and machine learning to fight covid-19," Physiol. Genomics, vol. 52, no. 4, pp. 200–202, 2020, doi: 10.1152/physiolgenomics.00029.2020.

[8] S. Dahbour, R. Qutteneh, Y. Al-shafie, and I. Tumar, Selecting Accurate Classifier Models for a MERS-CoV Dataset, vol. 1. Springer International Publishing, 2019.

[9] J. Edinson, C. Saire, J. E. C. Saire, and J. Oblitas, "Covid19 Surveillance in Peru on April using Text Mining," Medrxiv, no. 31, pp. 4–7, 2020, doi: 10.1101/2020.05.24.20112193.

[10] S. Kocbek et al., "Text mining electronic hospital records to automatically classify admissions against disease: Measuring the impact of linking data sources," J. Biomed. Inform., vol. 64, no. 2016, pp. 158–167, 2016, doi: 10.1016/j.jbi.2016.10.008.

[11] B. X. Tran et al., "Studies of novel coronavirus disease 19 (Covid-19) pandemic: A global analysis of literature," Int. J. Environ. Res. Public Health, vol. 17, no. 11, pp. 1–20, 2020, doi: 10.3390/ijerph17114095.

[12] M. Dixon-woods, A. Booth, T. Miller, and A. J. Sutton, "How can systematic reviews incorporate qualitative research? A critical perspective," Qual. Res., vol. 6, no. 1, pp. 27–44, 2006, doi: 10.1177/1468794106058867.

[13] M. Modi, A. B. Ibrahim, and N. Abdul, "Systematic review of artificial intelligence techniques in the detection and classification of COVID-19 medical images in terms of evaluation and benchmarking: Taxonomy analysis, challenges, future solutions and methodological aspects," J. Infect. Public Health, 2020, doi: 10.1016/j.jiph.2020.06.028.

[14] X. Jiang, M. Coffee, A. Bari, J. Wang, and X. Jiang, "Towards an Artificial Intelligence Framework for Data-Driven Prediction of Coronavirus Clinical Severity," Comput. Mater. Contin., vol. 63, no. 1, pp. 537–551, 2020, doi: 10.32604/cmc.2020.010691.

[15]O. N. Oyelade and A. E. Ezugwu, "A case-based reasoning framework for early detection and diagnosis of novel coronavirus," Informatics Med. Unlocked, vol. 20, no. July, p. 100395, 2020, doi: 10.1016/j.imu.2020.100395.

[16]E. Fayyoumi, S. Idwan, and H. Aboshindi, "Machine learning and statistical modelling for prediction of Novel COVID-19 patients case study: Jordan," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 5, pp. 122–126, 2020, doi: 10.14569/IJACSA.2020.0110518.

[17]M. R. H. Mondal, S. Bharati, P. Podder, and P. Podder, "Data analytics for novel coronavirus disease," Informatics Med. Unlocked, vol. 20, p. 100374, 2020, doi: 10.1016/j.imu.2020.100374.

[18]B. Afm, M. Jl, D. Thr, and C. F. Adp, "COVID-19 diagnosis prediction in emergency care patients: a machine learning approach," medRxiv, 2020, doi: 10.1101/2020.04.04.20052092.

[19]L. Yan et al., "A machine learning-based model for survival prediction in patients with severe COVID-19 infection," medRxiv, 2020, doi: 10.1101/2020.02.27.20028027.

[20]D. Brinati, A. Campagner, D. Ferrari, M. Locatelli, G. Banfi, and F. Cabitza, "Detection of COVID-19 Infection from Routine Blood Exams with Machine Learning: A Feasibility Study," J. Med. Syst., pp. 1–12, 2020.

[21]W. C. Dai et al., "CT Imaging and Differential Diagnosis of COVID-19," Can. Assoc. Radiol. J., vol. 71, no. 2, pp. 195–200, 2020, doi: 10.1177/0846537120913033.

[22]W. Naudé, "Artificial intelligence vs COVID-19: limitations, constraints and pitfalls," AI Soc., no. 0123456789, 2020, doi: 10.1007/s00146-020-00978-0.

[23]Q.-V. Pham, D. C. Nguyen, T. Huynh-The, W.-J. Hwang, and P. N. Pathirana, "Artificial Intelligence (AI) and Big Data for Coronavirus (COVID-19) Pandemic: A Survey on the State-of-the-Arts," IEEE Access, vol. 8, no. Cdc, pp. 130820–130839, 2020, doi: 10.1109/access.2020.3009328.

[24]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, doi: 10.1016/j.jiph.2016.09.007.

[25]M. B. Jamshidi et al., "Artificial Intelligence and COVID-19: Deep Learning Approaches for Diagnosis and Treatment," IEEE Access, vol. 8, no. December 2019, pp. 109581–109595, 2020, doi: 10.1109/access.2020.3001973.

[26]A. Gupta, I. Banerjee, and D. L. Rubin, "Automatic information extraction from unstructured mammography reports using distributed semantics," J. Biomed. Inform., vol. 78, no. November 2017, pp. 78–86, 2018, doi: 10.1016/j.jbi.2017.12.016.

[27] F. Faisal, S. A. Bhuiyan, F. Bin Ashraf, and A. R. M. Kamal, "A framework for disease identification from unstructured data using text classification and disease knowledge base," 2019 5th Int. Conf. Adv. Electr. Eng. ICAEE 2019, pp. 547–554, 2019, doi: 10.1109/ICAEE48663.2019.8975447.

[28]O. Paper, G. Lippi, A. Simundic, and M. Plebani, "Potential preanalytical and analytical vulnerabilities in the laboratory diagnosis of coronavirus disease 2019 (COVID-19)," Clin Chem Lab Med, vol. 58, no. 7, pp. 1070–1076, 2020, doi: 10.1515/cclm-2020-0285