IJAMEC
e-ISSN: 2147-8228

INTERNATIONAL JOURNAL OF APPLIED MATHEMATICS ELECTRONICS AND COMPUTERS

www.dergipark.org.tr/ijamec

International Open Access

Volume 10
Issue 01

March 2022

Research Article

Scientometric Analysis of COVID-19 Scholars Publication using Machine Learning

David Opeoluwa Oyewola [a], ***Emmanuel Gbenga Dada*** [b,*]

[a]*Department of Mathematics & Computer Science, Federal University Kashere, Gombe, Nigeria*
[b]*Department of Mathematical Sciences, Faculty of Science, University of Maiduguri, Maiduguri, Nigeria*

ARTICLE INFO

Article history:
Received 18 August 2021
Accepted 1 March 2022

Keywords:
Bayesian Generalized Linear Model,
Heteroscedastic Discriminant Analysis,
Multivariate Adaptive Regression Spline,
Naïve Bayes,
Scientometric.

ABSTRACT

The global health crisis that started in December 2019 resulted in an outbreak of coronavirus named COVID-19. Scientists worldwide are working to demystify the transmission and pathogenic mechanisms of the deadly coronavirus. The World Health Organization has declared COVID-19 a pandemic in March 2020, which makes it essential to track and analyse the research state of COVID-19 for guidance on further research. This research was conducted using scientometric analysis, knowledge-mapping analysis, COVID-19 studies and journal classifications. The publications used in this study include over 3000 COVID-19 papers made available to the public from 1 January 2018 to 15 April 2021 in the PubMed databases. In this study, it was discovered that the rapid reaction of researchers worldwide resulted in a fast growth trend between 2019 and 2021 in the number of publications related to COVID-19. It was discovered that the largest number of studies is in the United States of America, which is one of the countries most affected by a pandemic. The method adopted for this study involved the use of documents such as Case Reports (CAT), Journal Article (JAT), letter (LTR), EAT, and Editorial (EDT). This is followed by the classification of COVID-19 related publications that were retrieved from PubMed between 2019 and 2021 using machine learning (ML) models such as Naïve Bayes (NB), Bayesian Generalized Linear Model (BGL), Heteroscedastic Discriminant Analysis (HDA) and Multivariate Adaptive Regression Spline (MAR). Simulation results show that the classification accuracy of MAR is better than that of other ML models used in this study. The sensitivity of the MAR is within the range of 100%. This shows that MAR performs better than NB, BGL and HDA. MAR performs better with an overall accuracy of 89.62%. Our results show a high degree of strong collaboration in coronavirus research and the exchange of knowledge in the global scientific community.

1. Introduction

The emergence of COVID-19 virus on 30th December 2019 in a hospital in Wuhan, China marked a paradigm shift in medical research worldwide. Scientists have swiftly risen to the challenge to address and prevent the further spread of the pandemic. Clinical investigations and laboratory experiments identified a novel coronavirus (2019-nCoV) as the pathogenic agent responsible for the disease, which was consequently called COVID-19 [1]. Many areas of human lives have been affected by this extremely infectious disease-causing virus [2]. Predominantly, the COVID-19 pandemic has caused the death of about 3,046,229 people as of April 20, 2021 [3]. To control the negative fruits of COVID-19, it is required that governments and institutions put into operation effective public healthcare procedures inextricably linked to COVID-19 monitoring, detection, cures, injections, and research [4].

As a result of the newness and, subsequently, the unfamiliarity with the symptoms of the illness, scientific research have a very important role to play in combating the COVID-19 pandemic. There has been an astronomical increase in scientific publications in this research area [4],

* Corresponding author. E-mail address: *gbengadada@unimaid.edu.ng*
DOI: 10.18100/ijamec.984201

[5]. Many measures have already been adopted by different organisations to make research related to COVID-19 publicly available as much as possible. It is now clearly acknowledged that medical emergencies are also information emergencies [6], [7], [8], [9]. For instance, WHO keeps a record of important research which is updated every day [10]. This is in addition to a website that gathers information from different sources to make them available to the public [11], resembling that of the European Commission [9]. Journal and book publishers are granting free access to appropriate journals and books, while some publishers that require payments of article processing charges (APC) before publishing papers online totally waive publication fees for a similar purpose. An additional inventiveness is the making public of COVID-19 Open Research Dataset (CORD-19) [12]. CORD-19 is a fast expanding corpus that contains both old and new research publications on COVID-19. It updates its record of COVID-19 publications on daily basis. The dataset provided by CORD-19 has motivated researchers to use recent progress in natural language processing to produce fresh understandings in combating this highly contagious disease.

Doing a wide-ranging review of several COVID-19 related literature manually is an intimidating and laborious job. Machine learning (ML) techniques which is a sub-discipline of artificial intelligence (AI) can perform a crucial role in quickly analysing the huge number of literature and extracting important understandings from them. Hence, in March 2020, the US government highly advised researchers to take advantage of AI techniques in research to solve the COVID-19 pandemic [13]. Since ML has emerged as a beneficial tool in the fight against the pandemic, it is essential to evaluate the scientific advancement in this field by appraising the advances of research publications that enlighten on how ML is being applied to solving COVID-19 problems.

In this paper, ML algorithms were applied to perform the scientometric analysis of COVID-19 literature. ML algorithms have proven their efficacy in handling large data. It can therefore be used for the scientometric of the COVID-19 epidemic [23]. The major contributions of this work include:

i. A survey of different ML models and their application to systematizing, harmonizing, and classifying COVID-19 related publications was presented.
ii. Application of ML models to explore COVID-19 research topics that have received attention from the research community to give precedence to research gaps.
iii. Exploit the power of ML algorithms to comprehend how COVID-19 literature have evolved over time
iv. Apply ML techniques to identify the foremost authors, authors collaboration, countries, average number of publications per year, the average number of citations per year, most productive authors, most relevant sources; and
v. Use ML algorithms to investigate the associations between COVID-19 research topics and other related areas.

The rest of this paper is organized as follows: related works is done in Section 2. The methodology employed for this work as well as performance measurements is discussed in Section 3. The results and the discussion are presented in section 4, and we conclude in section 5.

2. Related Works

Many research works have used ML techniques to carry out bibliometric analysis of related COVID-19 publications. Aristovnik, Raveslj, Umek [14] did a bibliometric analysis of COVID-19 literature in the area of science and non-science research terrain. The study used the Scopus database together with every applicable and up-to-date information on COVID-19 literature that totalled 16,866 in the first six months of 2020. The downside of their work is that many publications on the subject of COVID-19 after June 2020 was not analysed. Haghani et al. [15] performed scientometric analysis and explorative study of COVID-19 related publications. The weakness of their work is that several recently published COVID-19 literature was not included in the analysis. Doanvo et al. [16] applied ML techniques to analyse the real content of coronavirus publication summaries to detect research intersection between COVID-19 and other coronavirus infections, research areas that have drawn much interest, and the ones that are demanding an investigation. The downside of their work is that several pieces of literature on the subject of COVID-19 after June 2020 was not analysed.

Dong et al. [17] used topic modelling to comprehend research flashpoints around COVID-19 and ailments caused by coronavirus variants. The shortcoming of this work is that it did not analyse many publications on COVID-19 after April 2020. Le Bras et al. [18] applied the COVID-19 and CORD-19 publication datasets to envisage COVID-19 research work done since the time the deadly virus was declared a pandemic till May 2020. The implication of this is that research on COVID-19 that are done after May 2020 was not analysed in the work. Mao et al. [19] performed a worldwide bibliometric and envisioned investigation on the significance and progress of coronavirus research. The authors' analysed coronavirus related literature from 2003 to the second month of 2020. Abd-Alrazaq et al. [20] did a machine learning-based bibliometric analysis of COVID-19 publications. The authors retrieved 196,630 literature from the CORD-19 database but only used 28,904 for their study. However, the authors only used ML to group subjects using the topical cluster. The limitation of the work is that it only covers COVID-19 publications within a period of 7 months (January to July 2020). Several valuable literature after this

period was not analysed. Moreover, the study only produced the percentage of topic and cluster dominance as output. There is no metric to measure the accuracy of ML models used in the proposed system. Colavizza et al. [21] provided a scientometric summary of the CORD-19 database. The authors investigated the description of publications incorporated into the CORD-19 database from a scientometric viewpoint. The limitation of this work is that publications analysed are the ones limited to May 2020. Thus, much COVID-19 research published subsequently were not analyzed.

In summary, the research gaps identified from this literature show that COVID-19 related publications investigated in the majority of the studies have limited dates which are roughly three months after the start of the COVID-19 pandemic. Therefore, several publications that came after that were not analyzed. Furthermore, many studies explored literature that are associated with all kinds of coronaviruses rather than concentrating on COVID-19. Consequently, the outcomes of such COVID-19 related studies were combined with the ones associated with other variants of coronavirus. Also, some studies included very meagre and insufficient number COVID-19 related publications. In addition, many studies did not investigate the subject that earlier studies had tackled, rather they just evaluated the metadata of those studies (such as countries, author name, author affiliation, total citations, bibliometric items, source journals, and others). Finally, classification of topics among different studies was done using manually rather than using machine learning techniques. To properly address the aforementioned research gaps, this paper intends to carry out a broad scientometric analysis of existing COVID-19 publications.

The COVID-19 pandemic, according to [50], has wreaked havoc in every country on the planet, causing major health, economic, and societal implications. The authors used VOSviewer software to conduct a scientometric examination of COVID-19. Text mining was applied, and the research included a review of 18,955 papers relevant to AI and COVID-19 from the Scopus database from March 2020 to June 2021. They employed an automated technique to extract research subjects of interest, and identified that the most essential current research lines are centred on patient-based treatments. The work also identified the most relevant publications in relation to the COVID-19 pandemic.

Malik et al. [51] conducted a scientometric analysis of coronavirus-related literature. Data were obtained from the Web of Science (WoS) and 28,846 publications was retrieved. According to the report, an increasing tendency of COVID-19 has been noticed for a long time, led by the United States and China, followed by the United Kingdom, Europe, and a few other countries. The last two decades provided approximately 39.5 per cent of total documents, while only the first six months of 2020 contributed approximately 46.5 per cent of total records.

3. Methodology

Extensive scientific data was obtained from PubMed on COVID-19-related research [22]. PubMed contains over 32 million quotations from MEDLINE, Life Science Journals, and online books for biomedical literature. PubMed is updated regularly to include the most recent COVID-19 published articles. We used the update of the timestamp of 1 January 2019 to 15 April 2021, which included over 86,000 articles on COVID-19 and CORONAVIRUS. In this study, we used the following search terms to retrieve studies on all coronaviruses: "COVID-19" and "CORONAVIRUS". The search was conducted on PubMed which was widely recognized in previous researches [23], [24] [25], [26]. The dataset downloaded from PubMed consists of information such as authors, abstract, document types, publication name, pagination, ISO source abbreviation, authors keywords, keywords associated with Scopus or ISI database, author address, cited references, time cited, year. In addition, five document types were considered for the machine learning process such as Journal Article (JAT), Case Reports (CAT), Letter (LTR), English Abstract (EAT) and Editorial (EDT) which scaled the number of documents to 3,000 records. In this analysis, quantitative tools such as PubMed, RISmed and bibliometrix were considered [27]. It offers several routines for importing COVID-19 PubMed data. The objective of PubMed is to collect metadata from a PubMed database based on NCBI REST APIs for publications, awards and clinical trials. First of all, we identify a question consisting of a COVID-19 compilation and coronavirus of articles published between 2019 and 2021 and submitted to the NCBI PubMed framework. In the interest of time and limits on the PubMed API, we limited the download to 3,000 entries. Finally, we convert the XML-structured object into a data frame, with cases that match Field Tags documents and variables as used in the R package bibliometrix [28].

3.1. Machine Learning

Machine Learning is a scientific discipline that focuses on programming systems to learn and improve experiences automatically [29]. In this sense, the learning of complex patterns is correlated with recognition and smart decision making. This study intends to classify the document types, such as Journal Article (JAT), Case Reports (CAT), Letter (LTR), English Abstract (EAT) and Editorial (EDT) of the COVID-19 data obtained from PubMed between 2019 and 2021 using machine learning. Table 1 shows the publication types, acronyms and scope notes of the COVID-19 scholars. The authors' dominance ranking proposed by [30] was used as input values while the Scholars publication types was used as the target values. Dominance ranking consists of Dominance Factor, Total Articles, Single Authored, Multi Authored, First Authored, Rank by Articles and Rank by Dominance Factor column [31]. Different classifiers such as Naïve Bayes, Bayesian Generalized Linear Model, Heteroscedastic Discriminant Analysis and Multivariate

Adaptive Regression Spline algorithm have been used to learn the proposed document types based on their performance in different literature. Before running the above methods, feature selection is performed. To determine whether all features are equally important or necessary to discriminate between the classes such as JAT, CAT, LTR, EAT and EDT.

Table 1. Publication Types of COVID-19 Scholars

Type of Document	Acronyms	Scope Note
Case Reports	CAT	Clinical lectures, accompanied by assessment trials leading to a diagnosis.
English Abstracts	EAT	Non-English language works Identifier for English abstracts.
Editorial	EDL	A report consisting of the views, values and policies of a journal's author or editor, normally on current medical or science topics of relevance for the health community or society as a whole. Published editorials by newspaper editors representing a company's or organizations official organ are usually important.
Journal Article	JAT	The most prevalent category for NLM databases for articles and other objects.
Letter	LTR	Work consists of written or printed correspondence between individuals or between individuals and company representatives. It may be personal or technical correspondence. The letter is typically from one or more writers to the editor of a journal or book that publishes the item commented or addressed in medical and other science journalists. COMMENT is sometimes followed by LETTER.

3.2 Naïve Bayes (NB)

The Naive Bayesian classifier is based on the theorem of Bayes with predictors of independence. A Bayesian Naïve model is simple to create, without complicated iterative estimation of parameters, which makes it useful especially for large datasets [32]. The Naive Bayesian classifier, despite its simplicity, is also surprisingly well used because of its often-advanced methods of classification. Naive Bayes presume that the effect on a given class (c) of the value of a predictor (d) is independent of that of other predictors. This presumption is referred to as class independence. The mathematical calculation is given as:

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)} \quad (1)$$

Where $P(c|d)$ is the posterior probability of class (target) given predictor (attribute), $P(c)$ is the prior probability of a class, $P(d|c)$ is the likelihood which is the probability of predictor given class and $P(d)$ is the prior probability of predictor.

3.3 Bayesian Generalized Linear Model (BGL)

The Bayes Generalized Linear Model (BGL) assumes preliminary or prior distribution based on preliminary or prior information and subsequent distribution is achieved through the integration of sample information with such prior information. In general, information collected from the post or posterior distribution is closer to true information, since it brings together sample data and expert views [33]. Assume that β, θ are independent, the prior distribution of the model is given as [34]:

$$\pi(\beta,\theta)\alpha\, exp(-\sum_{i=1}^{n} n_j\theta_i^{-1} - \sum_{j=1}^{p}\frac{\left(\beta_j - u_{\beta_j}\right)^2}{2\sigma_{\beta_j}^2}\prod^{n}\frac{\theta^{-1-m_i}}{\sqrt{\sigma_{\beta_j}^2}} \quad (2)$$

According to the Bayesian formula, it can be shown that the posterior distribution of parameters is directly proportional to the product of the prior distribution of parameters and the likelihood function of the model. Thus, when θ is fixed we have:

$$\pi(\beta|\theta,Y)\alpha L(Y|\beta,\theta)\pi(\beta,\theta)\alpha$$

$$\prod_{\tau(\theta_i\exp(\sum_{j=1}^{p}x_{ij}\beta_j))}^{n}\frac{\theta_i^{\theta_i\exp(\sum_{j=1}^{p}x_{ij}\theta_j)}}{}^{y_i}\quad \theta_i\left(\exp\left(\sum_{j=1}^{p}x_{ij}\beta_j\right)\right)^{-1}\exp(-\theta_i y_i)\exp\left(-\sum_{i=1}^{n}n_i\theta_i^{-1}-\sum_{j=1}^{p}\frac{(\beta_j-u_{\beta_j})}{2\sigma^2{}_{\beta_j}}\right)\prod_{i=1}^{n}\frac{\theta_i^{-1-m_i}}{\sqrt{\sigma_{\beta_j}^2}}$$

(3)

3.4 Heteroscedastic Discriminant Analysis (HDA)

The Heteroscedastic Discriminant Analysis (HDA) is a generalization of Linear Discriminant Analysis (LDA) which handles unequal samples of the class of covariance. We can regard HDA as the limiting maximum probability projection where the log probability of samples is maximized in the predicted space. The restriction is determined by maximizing the planned interclass dispersal volume by adding an updated HDA analysis objective function [35]. The updated objective function takes weighted contributions of each class into account which is given as [36]:

$$\prod^{c}\left(\left|\frac{W^t s_b W}{W^t \sum_i W}\right|\right)^{n_i} = \frac{|W^t s_b W|}{\prod_{i=1}^{c}|W^t \sum_i W|^{n_i}} \quad (4)$$

By taking log of (4) we get a discriminant function:

$$H(W) = \sum_{i=1}^{c} -n_i \log|W^t \sum_i W| + nlog|W^t s_b W| \quad (5)$$

$$S_i = \sum_{x\in C_i}(x - m_i)(x - m_i)^t \quad (6)$$

$$\sum_i = \frac{1}{n_i} S_i \quad (7)$$

$$m = \frac{1}{n}\sum_x x \quad (8)$$

$$m_i = \frac{1}{n_i}\sum_{x\in C_i} x \quad (9)$$

$$s_b = \sum_{i=1}^{c} n_i (m_i - m)(m_i - m)^t \quad (10)$$

Where W is the projection matrix, $\sum_i$ is the covariance matrix of class C_i, m_i is the mean vector between the class.

3.5 Multivariate Adaptive Regression Spline (MAR)

Multivariate Adaptive Regression Spline (MAR) is a versatile regression method and nonparametric approach that incorporates piecewise linear regression function referred to as basic function (*bf*). To estimate the performance of MAR, it uses basic functions (*bf)* for capturing the hidden nonlinear relations between independent input variables [37]. *Bf* is therefore the main component in the generation of a MAR model. The MAR can be defined as:

$$o_p = \varphi + \sum_{m=1}^{M} c_m bf_m(x) \quad (11)$$

$$bf_m(x) = \max(0, x - t) \quad (12)$$

or

$$bf_m(x) = \max(0, t - x) \quad (13)$$

Where o_p is the output of the MAR model, φ is a constant value, M is the number of basic functions contributing to the MAR model, $bf_m(x)$ is the basic functions, and c_m is the coefficient related to each $bf_m(x)$.

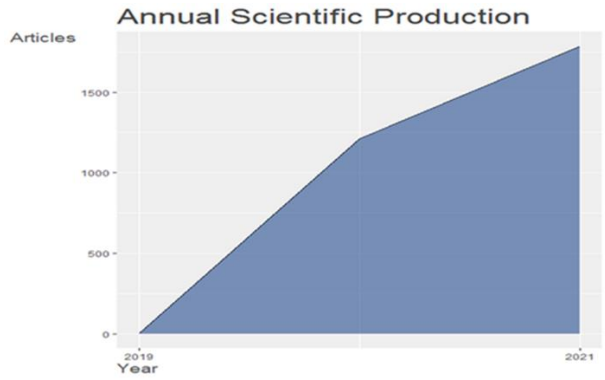

Figure 1. Annual Scientific Production from 2019-2021

4. Results and Discussion

A total of 3,000 documents were retrieved, covering a period of 2 years (2019–2021), since the COVID-19 virus was discovered on 30th December 2019 in a hospital in Wuhan, China which later become a global pandemic [38]. Interestingly, the number of documents obtained has grown exponentially. At the beginning of 2019, only two articles were published while in 2020 the COVID-19 publications have increased by 40.43% and in 2021 the articles increased by 59.51%. The highest productivity was observed in 2021 with a total of 1784 publications. This implies that interest in COVID-19 research is growing exponentially as shown in Table 2 and Figure 2.

Table 2. Growth Rate of COVID-19 Articles

Year	Published Articles	Percentage (%)
2019	2	0.0667
2020	1212	40.43
2021	1784	59.51

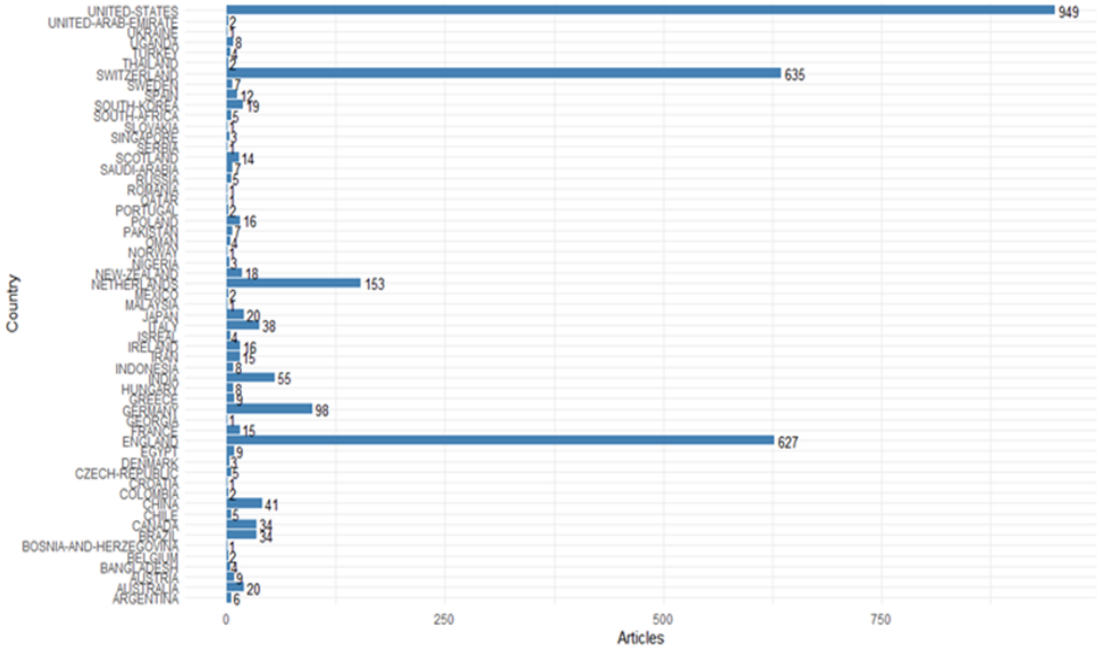

Figure 2. COVID-19 Scholars Countries by number of articles (2019-2021)

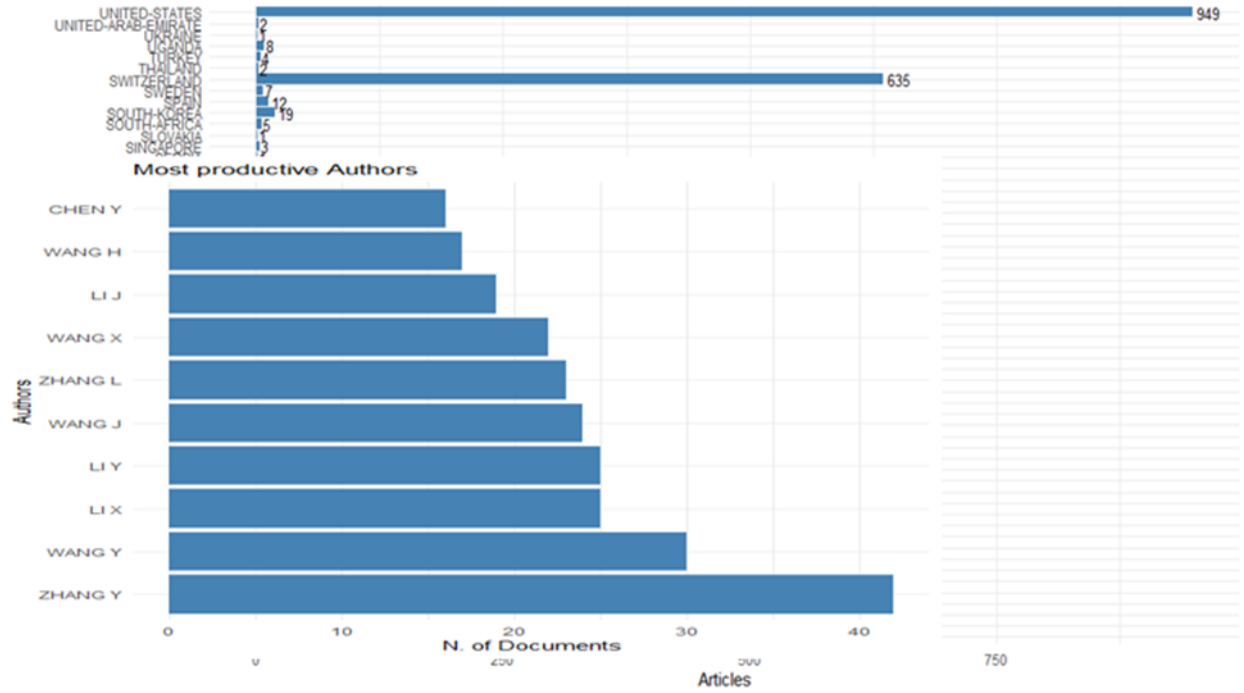

Figure 3. Most Productive Authors

Table 3. Top 10 Manuscripts per citation on COVID-19

Author	Year	Journal	DOI
CHEN B [39]	2020	Environmental Pollution	10.1016/j.envpol.2021.117074
ZAKI N [40]	2020	Journal Infection Public Health	10.1016/j.jiph.2021.01.019
CHEN Z [41]	2020	Journal Infection Public Health	10.1016/j.jiph.2021.02.002
PANIKAR S [42]	2020	Journal Infection Public Health	10.1016/j.jiph.2020.12.037
CHEDID M [43]	2020	Journal Infection Public Health	10.1016/j.jiph.2021.02.001
KHODEIR MM [44]	2020	Journal Infection Public Health	10.1016/j.jiph.2021.03.001
KITAJIM A H [45]	2020	Journal of Clinical Virology	10.1016/j.jcv.2021.104813
DOURAD O D [46]	2021	Biomedicine & Pharmacotherapy	10.1016/j.biopha.2021.111578
BARIK S [47]	2020	Infection, Genetics and Evolution	10.1016/j.meegid.2021.104858
CHEW NW [48]	2021	The Canadian Journal of Cardiology	10.1016/j.cjca.2021.04.003

Figure 2 presents COVID-19 Scholars countries based on articles published from 2019 to 2021. The most relevant country is the United States with 949 articles, Switzerland ranked second with 635 articles, England ranked third with 627 articles, the Netherlands fourth with 153 articles and Germany fifth with 98 articles.

A total of 20,071 authors have been identified in the PubMed database out of which 161 are single-authored documents while 19910 are multi-authored documents. The top 10 most productive authors are shown in Figure 3. Author Zhang Y has the highest number of published papers (42), followed

by Wang Y (30), Li X and Li Y have the same number of published papers (25) and Wang J (24) papers as shown in Fig. 3. Also, Table 3 shows the top ten most cited manuscripts based on COVID-19 research. According to the number of total citations, it is evident that Chen B in the Journal of Environmental Pollution and Zaki N in Journal of Infection Public Health are the most important authors involved in COVID-19 research in the year 2020.

Table 4. Most Relevant Keywords

Author Keywords	Articles (%)	Author Keywords-Plus	Articles-Plus (%)
COVID-19	1362 (49.19)	HUMANS	857 (22.05)
SARS-COV-2	566 (20.44)	COVID-19	849 (21.84)
CORONAVIRUS	340 (12.27)	SARS-COV-2	771 (19.84)
PANDEMIC	151 (5.45)	PANDEMICS	332 (8.54)
CORONAVIRUS DISEASE 2019	85 (3.07)	FEMALE	272 (6.99)
MENTAL HEALTH	58 (2.09)	MALE	238 (6.12)
PUBLIC HEALTH	58 (2.09)	ADULT	179 (4.61)
MORTALITY	54 (1.95)	MIDDLE-AGED	165 (4.24)
EPIDEMIOLOGY	52 (1.88)	AGED	130 (3.34)
SEVERE ACUTE RESPIRATORY SYNDROME CORONAVIRUS 2	43 (1.55)	CROSS-SECTIONAL STUDIES	94 (2.42)

Table 4 shows the most relevant keywords such as author keywords, articles, author keywords-plus, articles-plus. COVID-19 was the most frequently used keyword with 1362 (49.19%) occurrences while HUMANS was the most frequently keywords-plus with 857 (22.05%). SARS-COV-2 is the second most relevant keyword with 566 (20.44%) occurrences while COVID-19 is the second most frequent keyword-plus with 849 (21.84%). A density map is also generated for keywords with co-occurrence keywords as shown in Figure 4. COVID-19, HUMANS, SARS-COV-2 and PANDEMIC was the most frequently used keywords.

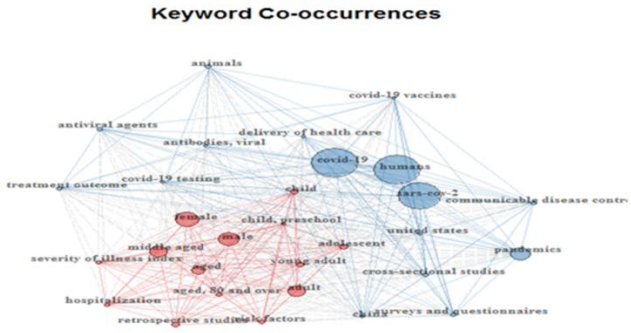

Figure 4. Co-occurrence of Keywords

A word clouds or tag clouds are graphical representations of word frequency that give greater prominence to words that appear more frequently in a source text. Figure 5 shows the word clouds of all the extracted COVID-19 databases in PubMed. Patients, respiratory, COV, SARS, infection, health, severe, acute, care, clinical, 2020, results, methods are frequently used words in the abstracts. Figure 6 is the word clouds of all the journal of COVID-19 documents. PloSOne has the highest number of COVID-19 publications followed by the Journal of Clinical Medicine, Frontiers in Immunology, Cureus, Frontiers in Psychology, Cardiothoracic imaging and Scientific reports.

Figure 5. Word Cloud of Author Abstract

Figure 6. Word Cloud of Journal Publication of COVID-19

The machine learning (ML) field is continuously evolving. It has been used in a different fields such as psychology, computer vision, computational complexity, control theory, information theory, neurobiology and so on [49]. In this study, the researcher intended to classify the document types, such as Journal Article (JAT), Case Reports (CAT), Letter (LTR), English Abstract (EAT) and Editorial (EDT) of the COVID-19 data obtained from PubMed between 2019 and 2021 using machine learning. Fig 7 shows the distribution of journal articles considered. Journal Article (JAT) has more articles than the remaining four articles. This implies that the author published more Articles than the rest such as CAT, LTR, EAT and EDT.

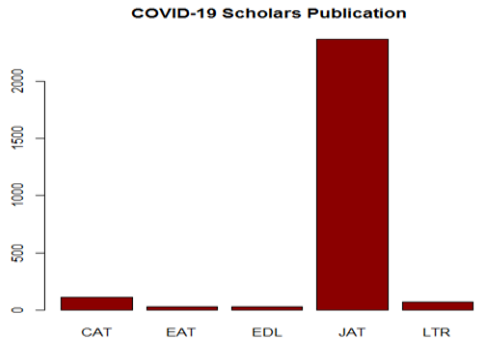

Figure 7. COVID-19 Scholars Publication

In this research, we considered four machine learning algorithms namely Naïve Bayes (NB), Bayesian Generalized Linear Model (BGL), Heteroscedastic Discriminant Analysis (HDA) and Multivariate Adaptive Regression Spline (MAR). Table 5 shows the performance of Machine Learning algorithms on the COVID-19 Journal Publication dataset. Different performance metrics were estimated including Specificity, Prevalence, Detection Rate and Balanced Accuracy. NB, BGL and HDA failed to classify journal publication datasets as shown in the performance metrics. The specificity of the NB, BGL and HDA is within the range of 33-100% as shown in Table 5. Simulation results show that the classification accuracy of MAR is better than that of other ML models used in this study. The sensitivity of the MAR is within the range of 100%. This shows that MAR performs better than NB, BGL and HDA. Comparing the overall statistics of the proposed ML models on the journal publication dataset as shown in Table 6, MAR performs better with an overall accuracy of 89.62%.

Table 5. Performance of Machine Learning

MODEL	PUBLICATION	Specificity	Prevalence	Detection Rate	Balanced Accuracy
NB	CAT	100.00	6.15	0.00	50.00
	EAT	75.67	0.38	0.00	37.84
	EDL	100.00	0.38	0.00	50.00
	JAT	33.33	89.62	68.85	55.08
	LTR	100.00	3.46	0.00	50.00
BGL	CAT	100.00	6.15	0.00	50.00
	EAT	46.72	0.38	0.00	23.35
	EDL	100.00	0.38	0.00	50.00
	JAT	62.96	89.62	43.08	55.52
	LTR	100.00	3.46	0.00	50.00
MAR	CAT	100.00	6.15	0.00	50.00
	EAT	100.00	0.38	0.00	50.00
	EDL	100.00	0.38	0.00	50.00
	JAT	00.00	89.62	89.62	50.00
	LTR	100.00	3.42	0.00	50.00
HAD	CAT	100.00	6.15	0.00	50.00
	EAT	35.13	0.038	0.38	67.57
	EDL	93.05	0.038	0.00	46.52
	JAT	66.67	89.62	19.23	44.06
	LTR	95.22	3.46	0.77	58.72

Table 6. Overall Statistics

MODEL	Accuracy	Kappa	95% CI
NB	68.85	0.0266	(0.6283, 0.7442)
MAR	43.08	0.0142	(0.3698, 0.4934)
MAR	89.62	0.0000	(0.8525, 0.9304)
HAD	20.38	-0.0052	(0.1566, 0.2580)

5. Conclusion

The COVID-19 pandemic is a typical emergency for public health where a high incidence of infection presents a major risk not only for global public health but also for economic and social growth. To solve these emergencies, the crisis, the impact on its various fields and the solutions to deal with possibly devastating consequences must be well understood, analysed and dealt with. Scientific awareness of COVID-19 is therefore vital as it contributes to responses to real-life issues. This review gives a complete overview of the literature of COVID-19. The main significance of this research is to investigate institutional productivity, institutional research rankings, and assessing the impact of top scholarly articles in COVID-19, as well as analysing the profiles of top authors in COVID-19 research, and applying machine learning to classify academic papers in COVID-19. This research analysis is carried out using the PubMed database. Specifically, we identified the main COVID-19–related topics such as the growing rate of COVID-19 production, COVID-19 Scholars countries, most productive authors, relevant keywords, word clouds of abstract and classification of document types of journal publication. Future research work is to develop a review study by using Google Scholar, Scopus, Web of Science and so on.

Acknowledgment

The authors want to thank PubMed for providing access to the COVID-19 related publications dataset which was used for the experiments conducted in this study.

References

[1] WHO, “Novel Coronavirus (2019-nCoV) Situation Report-1”, World Health Organization. Geneva, Switzerland; 2020. Available at: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200121-sitrep-1-2019-ncov.pdf?sfvrsn=20a99c10_4 [Accessed June 2021].

[2] ILO, FAO, IFAD and WHO, "Impact of COVID-19 on people's livelihoods, their health and our food systems"; 2020. Available at: https://www.who.int/news/item/13-10-2020-impact-of-covid-19-on-people%27s-livelihoods-their-health-and-our-food-systems [Accessed May 2021].

[3] Worldometer, “COVID-19 Coronavirus Pandemic”; 2021. Available at: https://COVID Live Update: 142,813,353 Cases and 3,046,229 Deaths from the Coronavirus - Worldometer (worldometers.info) [Accessed May 2021].

[4] P. Yang, X. Wang, "COVID-19: a new challenge for human beings", Cellular & molecular immunology, vol. 17, no. 5, pp. 555-557, 2020.

[5] A. Aristovnik, D. Ravšelj, L. Umek, “A bibliometric analysis of COVID-19 across science and social science research landscape”, Sustainability, vol. 12, no. 21, pp. 9132, 2020.

[6] B. Xie, D. He, T. Mercer, Y. Wang, D. Wu, K. R. Fleischmann, Y. Zhang, L. H. Yoder, K. K. Stephens, M. Mackert, M. K. Lee, “Global health crises are also information crises: A call to action”, Journal of the Association for Information Science and Technology, vol. 71, no. 12, pp. 1419-23, 2020.

[7] M. Cinelli, W. Quattrociocchi, A. Galeazzi, C. M. Valensise, E. Brugnoli, A. L. Schmidt, P. Zola, F. Zollo, A. Scala, “The COVID-19 social media infodemic”, Scientific Reports, vol. 10, no. 1, pp. 1-10, 2020.

[8] B. Swire-Thompson, D. Lazer, “Public health and online misinformation: Challenges and Recommendations”, Annual Review of Public Health, vol. 41, no. 1, pp. 433–451, 2020. https://doi.org/10.1146/annurevpublhealth-040119-094127 PMID: 31874069

[9] J. P. Ioannidis, “Coronavirus disease 2019: the harms of exaggerated information and non-evidence-based measures”, European Journal of clinical investigation, vol. 50, no. 4, 2020.

[10] J. Zarocostas, “How to fight an infodemic”, Lancet. 395(10225), 2020. https://doi.org/10.1016/S0140-6736(20)30461-X.

[11] EPI-WIN “WHO Information Network for Epidemics”, 2020; Available at: https://www.who.int/teams/risk-communication.

[12] L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Eide, K. Funk, R. Kinney, Z. Liu, W. Merrill, P. Mooney, “Cord-19: The covid-19 open research dataset”, ArXiv. Jul 9, 2020.

[13] A. Doanvo, X. Qian, D. Ramjee, H. Piontkivska, A. Desai, M. Majumder, "Machine learning maps research needs in covid-19 literature", Patterns, 1(9):100123, Dec 11, 2020.

[14] A. Aristovnik, D. Ravšelj, L. Umek, "A bibliometric analysis of COVID-19 across science and social science research landscape", Sustainability, 12(21):9132, Jan. 2020 [doi: 10.20944/preprints202006.0299.v1]

[15] M. Haghani, M. C. Bliemer, F. Goerlandt, J. Li, "The scientific literature on Coronaviruses, COVID-19 and its associated safety-related research dimensions: A scientometric analysis and scoping review", Safety Science, 1;129:104806, 2020 [doi: 10.1016/j.ssci.2020.104806]

[16] A. Doanvo, X. Qian, D. Ramjee, H. Piontkivska, A. Desai, M. Majumder, "Machine learning maps research needs in covid-19 literature", Patterns, 1(9):100123, 2020. [doi: 10.1101/2020.06.11.145425]

[17] M. Dong, X. Cao, M. Liang, L. Li, H. Liang, G. Liu, "Understand research hotspots surrounding COVID-19 and other coronavirus infections using topic modelling", medRxiv. Jan 2020. [doi: 10.1101/2020.03.26.20044164]

[18] P. Le Bras, A. Gharavi, D. A. Robb, A. F. Vidal, S. Padilla, M. J. Chantler, "Visualising COVID-19 research", arXiv preprint, 2020.

[19] X. Mao, L. Guo, P. Fu, C. Xiang, "The status and trends of coronavirus research: A global bibliometric and visualized analysis", Medicine, 29;99(22):e20137, 2020.

[20] A. Abd-Alrazaq, J. Schneider, B. Mifsud, T. Alam, M. Househ, M. Hamdi, Z. Shah, "A comprehensive overview of the COVID-19 literature: Machine learning-based bibliometric analysis", Journal of medical Internet research, 8;23(3):e23703, 2021.

[21] G. Colavizza, R. Costas, V. A. Traag, N. J. Van Eck, T. Van Leeuwen, L. Waltman, "A scientometric overview of CORD-19", PloS one, 7;16(1):e0244839, 2021. https://doi.org/10.1371/journal.pone.0244839

[22] NIH, "National Library of Medicine", National Centre for Biotechnology Information, 2021; Available from: https://pubmed.ncbi.nih.gov.

[23] Y. Gong, T. C. Ma, Y. Y. Xu, R. Yang, L. J. Gao, S. H. Wu, J. Li, M. L. Yue, H. G. Liang, X. He, T. Yun, "Early research on COVID-19: a bibliometric analysis", The Innovation, 1(2):100027, Aug 28, 2020. https://doi.org/10.1016/j.xinn.2020.100027.

[24] F. De Felice, A. Polimeni, "Coronavirus disease (COVID-19): a machine learning bibliometric analysis", in vivo, 34(3 suppl), 1613-1617, 2020.

[25] F. R. Nasab, "Bibliometric analysis of global scientific research on SARS-CoV-2 (Covid-19)", MedRxiv. Jan 1, 2020.

[26] H. Dehghanbanadaki, F. Seif, Y. Vahidi, F. Razi, E. Hashemi, M. Khoshmirsafa, H. Aazami, "Bibliometric analysis of global scientific research on Coronavirus (COVID-19)", Medical Journal of the Islamic Republic of Iran, 34:51, 2020.

[27] M. Aria, C. Cuccurullo, "Bibliometrix: An R-tool for comprehensive science mapping analysis", Journal of informetrics, 1;11(4):959-75, Nov 2017. https://doi.org/10.1016/j.joi.2017.08.007

[28] R Package, 2021. Available at: www.bibliometrix.org [Accessed January 20, 2020].

[29] E. Alpaydin, Introduction to machine learning, MIT press; 2020 Mar 17.

[30] S. Kumar, S. Kumar, "Collaboration in research productivity in oilseed research institutes of India", InProceedings of Fourth International Conference on Webometrics, Informetrics and Scientometrics, Vol. 28, Jul 28, 2008.

[31] K. Yuan, L. Gao, Z. Jiang, Z. Tang, "Formula Ranking within an Article", InProceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, pp. 123-126, May 23, 2018. doi:10.1145/3197026.3197061.

[32] M. J. Sánchez-Franco, A. Navarro-García, F. J. Rondán-Cataluña, "A naive Bayes strategy for classifying customer satisfaction: A study based on online reviews of hospitality services", Journal of Business Research, 101:499-506, Aug 1, 2019. doi:10.1016/j.jbusres.2018.12.051.

[33] G. Shi, C. Y. Lim, T. Maiti, "Bayesian model selection for generalized linear models using non-local priors", Computational Statistics & Data Analysis, 133:285-96, May 1, 2019. doi:10.1016/j.csda.2018.10.007.

[34] C. Gao, Q. Li, Z. Guo, "Automobile Insurance Pricing with Bayesian General Linear Model", In International Conference on Information and Management Engineering, pp. 359-365, Sep 17, 2011, Springer, Berlin, Heidelberg.

[35] K. S. Gyamfi, J. Brusey, A. Hunt, E. Gaura, "Linear classifier design under heteroscedasticity in Linear Discriminant Analysis", Expert Systems with Applications, 79:44-52, 2017 Aug 15. doi:10.1016/j.eswa.2017.02.039.

[36] K. Stąpor, T. Smolarczyk, P. Fabian, "Heteroscedastic discriminant analysis combined with feature selection for credit scoring", Statistics in Transition new series, 17(2):265-80, 2016.

[37] M. Samadi, M. H. Afshar, E. Jabbari, H. Sarkardeh, "Application of multivariate adaptive regression splines and classification and regression trees to estimate wave-induced scour depth around pile groups", Iranian Journal of Science and Technology, Transactions of Civil Engineering, 44(1):447-59, Oct. 2020. https://doi.org/10.1007/s40996-020-00364-2.

[38] D. O. Oyewola, A. F. Augustine, E. G. Dada, A. Ibrahim, "Predicting Impact of COVID-19 on Crude Oil Price Image with Directed Acyclic Graph Deep Convolution Neural Network", Journal of Robotics and Control (JRC), 2(2):103-109, Mar 19, 2021.

[39] B. Chen, J. Han, H. Dai, P. Jia, "Biocide-tolerance and antibiotic-resistance in community environments and risk of direct transfers to humans: Unintended consequences of community-wide surface disinfecting during COVID-19", Environmental Pollution, 117074, 2021 Apr 3. DOI: 10.1016/j.envpol.2021.117074.

[40] N. Zaki, E. A. Mohamed, "The estimations of the COVID-19 incubation period: A scoping reviews of the literature", Journal of infection and public health, 14(5):638-46, May 2021, DOI: 10.1016/j.jiph.2021.01.019

[41] Z. Chen, W. Xie, Z. Ge, Y. Wang, H. Zhao, J. Wang, Y. Xu, W. Zhang, M. Song, S. Cui, X. Wang, "Reactivation of SARS-CoV-2 infection following recovery from COVID-19", Journal of infection and public health, 14(5):620-627, May 2021.

[42] S. Panikar, G. Shoba, M. Arun, J. J. Sahayarayan, A. U. Nanthini, A. Chinnathambi, S. A. Alharbi, O. Nasif, H. J. Kim, "Essential oils as an effective alternative for the treatment of COVID-19: Molecular interaction analysis of protease (Mpro) with pharmacokinetics and toxicological properties", Journal of Infection and Public Health, 14(5):601-10, May 2021. DOI: 10.1016/j.jiph.2020.12.037.

[43] M. Chedid, R. Waked, E. Haddad, N. Chetata, G. Saliba, J. Choucair, "Antibiotics in treatment of COVID-19 complications: a review of frequency, indications, and efficacy", Journal of infection and public health, 14(5):570, May 2021, DOI: 10.1016/j.jiph.2021.02.001.

[44] M. M. Khodeir, H. A. Shabana, A. S. Alkhamiss, Z. Rasheed, M. Alsoghair, S. A. Alsagaby SA, Khan MI, Fernández N, Al Abdulmonem W. Early prediction keys for COVID-19 cases progression: A meta-analysis. Journal of infection and public health. 2021 Mar 5, DOI: 10.1016/j.jiph.2021.03.001.

[45] H. Kitajima, Y. Tamura, H. Yoshida, H. Kinoshita, H. Katsuta, C. Matsui, A. Matsushita, T. Arai, S. Hashimoto, A. Iuchi, T. Hirashima, "Clinical COVID-19 diagnostic methods: Comparison of reverse transcription loop-mediated isothermal amplification (RT-LAMP) and quantitative RT-PCR (qRT-PCR)", Journal of Clinical Virology, 1;139:104813, Jun 2021. DOI: 10.1016/j.jcv.2021.104813.

[46] D. Dourado, D. T. Freire, D. T. Pereira, L. Amaral-Machado, E. N. Alencar, A. L. de Barros, E. S. Egito, "Will curcumin nanosystems be the next promising antiviral alternatives in COVID-19 treatment trials?" Biomedicine & Pharmacotherapy, 6:111578, Apr. 2021. DOI: 10.1016/j.biopha.2021.111578.

[47] S. Barik, "Systematizing the genomic order and relatedness in the open reading frames (ORFs) of the coronaviruses", Infection, Genetics and Evolution, 92:104858, Aug 2021. DOI: 10.1016/j.meegid.2021.104858.

[48] N. W. Chew, Z. G. Ow, V. X. Teo, R. R. Heng, C. H. Ng, C. H. Lee, A. F. Low, M. Y. Chan, T. C. Yeo, H. C. Tan, P. H. Loh, "The Global Impact of the COVID-19 Pandemic on STEMI care: A Systematic Review and Meta-Analysis", Canadian Journal of Cardiology, 2021 Apr 20. DOI: 10.1016/j.cjca.2021.04.003.

[49] T. M. Mitchell, Machine learning, Burr Ridge, IL: McGraw Hill, 45(37):870-7, 1997.