

# Machine learning in handling disease outbreaks: a comprehensive review

Dianadewi Riswantini, Ekasari Nugraheni

Research Center for Information and Data Sciences, National Research and Innovation Agency, Bandung, Indonesia

## Article Info

### Article history:

Received Jan 11, 2022

Revised Apr 22, 2022

Accepted Jun 6, 2022

### Keywords:

Disease outbreak

Healthcare

Infectious disease

Machine learning

Review

## ABSTRACT

The changes in the global environment have made impact on the evolution of infectious diseases, virus mutations, or new diseases which are challenging to be tackled with new technological advances. This work aims to identify and analyze previous studies on machine learning applications in handling disease outbreaks. Bibliometric analysis was conducted on 3,447 scientific articles selected from the Scopus database. Further, latent dirichlet analysis (LDA) method was applied to identify the topic hotspots in attempting to deepen the analysis. The LDA results identified twelve topic hotspots that can be classified into three themes: COVID-19 disease, miscellaneous diseases, and public opinion on disease outbreaks for discussion. The study reveals that the scientific structure of this domain is dominated by machine learning research on COVID-19 diseases and miscellaneous diseases caused by pathogens or some genetic factors. A huge amount of multimodal medical data was used by previous studies for prediction, forecasting, classification, or screening purposes to resolve many problems of diseases, including epidemiological surveillance, diagnosis, treatment, health monitoring, epidemic management, viral infection, and pathogenesis. Public opinions toward new diseases are also an interesting topic in addition to the public perceptions in response to the health protocol and policies.

*This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.*



## Corresponding Author:

Dianadewi Riswantini

Research Center for Information and Data Sciences, National Research and Innovation Agency

Bandung, Indonesia

Email: dianadewi.riswantini@brin.go.id

## 1. INTRODUCTION

Human behavior or negligence has impacted significant changes in the global ecosystem. The changes in the global environment have resulted in the emergence of pandemic threats caused by existing infectious diseases, virus mutations, or new diseases. The pathogen intrusion has evolved rapidly and threatened the human population through various infectious diseases. To tackle the threats, scientists and policymakers have conducted some action to prevent the spread of the disease by reducing the risks [1]. Moreover, unhealthy lifestyles have influenced human body conditions that may lead to internal and genetic diseases. World Health Organization (WHO) reported that heart disease, pulmonary and stroke are the top three leading causes of death globally [2]. Hence, the emergence of COVID-19 has worsened human health conditions and changed health data globally. The COVID-19 pandemic has affected various life aspects and still poses many challenges due to the virus mutation [3]. This phenomenon has accelerated research in epidemiology to study the pattern of disease spread or health-related occurrence and the factors that can influence the disease. In recent days of the pandemic, this knowledge has been beneficial in mapping the spread pattern of COVID-19.

Over the years, the study of epidemiology has evolved, and a new paradigm has developed both in public health services and scientific research. The data analysis methods are improved, requiring a multi-disciplinary approach to overcome various disease outbreaks. More and more advanced technology has been developed for epidemiology to any levels of analysis from the population, individuals, organs, cells to DNA, to elaborate on many concepts of health and diseases [4]. The development of information and communication technology (ICT) has speeded up the data analytic processes and increased the datasets from disparate data sources in the healthcare sector. Artificial intelligence has enabled the computer to imitate human intelligence and process huge amounts of data beyond human capability. The advances of artificial intelligence (AI) have resulted in tremendous machine learning approaches developed to achieve automated analysis in supporting the real-time decision-making process and discovering solutions to complex healthcare problems by learning the patterns through algorithms and statistical models. The artificial intelligence community developed machine learning by utilizing statistical methods to solve many research problems. AI and machine learning have revolutionized the transition from traditional to modern epidemiology by delivering solutions for the analysis of complex clinical data applicable in many applications, including disease diagnosis, drug repurposing, and discovery, personalized health treatment, health risk identification, outbreak prediction, and intelligent health system development [5], [6].

The relevance of this review study is intended to explore the evolution of machine learning application in supporting the scientific discipline of epidemiology. Bibliometric analysis was first conducted to get insight into the scientific mapping of the research domain. Further, systematic content analysis was performed to deepen the analysis of the contributions of machine learning in tackling the disease outbreaks. We address that this work has the following contributions: 1) present an overview of the role of machine learning in handling disease outbreaks; 2) describe the development and state-of-the-art of this domain research through the network and evolution keywords analysis; and 3) present the insight of infectious diseases that get thoughtful attention from researchers and the summarized view of the advanced techniques of machine learning used for solving diseases outbreak problems through content analysis.

This article's organizational structure starts with the description of the context and the research relevance mentioned in section 1. Previous works related to the topic of this article were presented in section 2, including the research gap that may be filled by the study. Section 3 describes the study workflow and materials used in this review. The results of the studies are discussed in section 4. At the beginning of section 4, some bibliometric measurements are explained in two subsections: scientific production and research interest and evolution. The last part of section 4 describes 12 topic hotspots selected through the topic modeling process. These topics are categorized into three groups, namely COVID-19 disease, miscellaneous diseases, and public opinion on disease outbreaks. Finally, the limitations and conclusion of this work will be drawn in section 5 and section 6.

## 2. RELATED WORKS

The growing use of machine learning in the health sector has encouraged the development of health science in dealing with disease outbreaks. Advances in AI and machine learning (ML) have helped society solve global health challenges and accelerated the achievement of sustainable healthcare. Research on this technology has been done for the health sector with various objectives, including disease diagnosis, health risk assessment, prediction and surveillance of disease outbreaks, and health management strategies [7]-[9]. The emergence of AI-based digital epidemiology supported by machine learning has dramatically contributed to the improvement of public health and disease outbreak handling [10], [11]. Many previous studies examining the application of machine learning in the health sector have been carried out. The emergence of a pandemic outbreak has accelerated the development of machine learning for epidemiology, the study dealing with disease transmission, and the factors that influence the disease. The following paragraphs will discuss the previous works related to this review study.

Research on the spread of infectious diseases has increased since the COVID-19 pandemic struck globally. Alfred and Obit [12] discussed the role of machine learning in handling the disease outbreak attempting to reduce its spreading. The study focused on detecting and predicting disease attacks by applying various classification and prediction models using structured and unstructured data. A systematic review of surveillance systems using social media data for similar purposes was conducted by [13]. Moreover, the development of text mining and the increasing use of social media deliver research opportunities to implement machine learning for health computing, covering the identification of risk factors and symptoms, health crisis phases, and public health responses to pandemics [14]-[16].

Some disease outbreaks were examined in epidemiological areas of research. Sak and Suchodolska [17] reviewed the potential application of machine learning for nutritional epidemiology to identify the influence of nutrients on the health and disease of the human body. Basu *et al.* [18] discussed the application

of machine learning for diabetes clinical epidemiology to improve risk stratification and prediction. Special issues of review studies have been conducted for particular diseases, such as cancer risk assessment [19], mosquito-borne disease transmission [20], COVID-19 diagnosis [21], and others.

From the analysis of previous studies, we capture that the existing research discusses the use of ML for specific disease problems or certain purposes such as disease detection, risk prediction, and spread estimation. There are still few studies that discuss the application of ML in dealing with disease outbreaks from a helicopter view. Our research examines a broader perspective than what has been done in previous studies to fill the gap. We explore the following issues: the evolution of research topics, research advances in infectious diseases, and the machine learning model that is currently being used.

### 3. MATERIALS AND METHOD

This review study used the Scopus citation database records as bibliographic data for analysis. Scopus was chosen because it covers a more expanded spectrum and a superior number of peer-reviewed publications and provides reliable bibliographic data [22]. The study was carried out in three stages: study design, data preparation, and data analysis. The study workflow is presented in Figure 1. At first, the study design stage determined the citation database and software tools objectives of the study. The tools were employed to devise the scientific mapping of the research domain. Then, we fixed searching keywords and inclusion/exclusion criteria to select the desired bibliometric data.

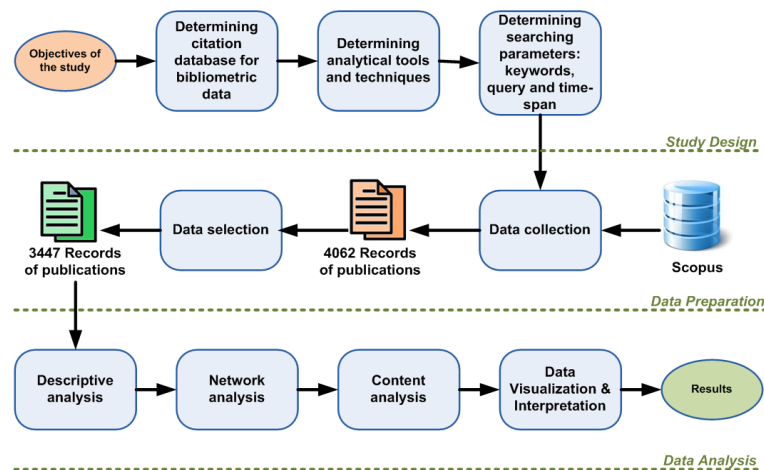


Figure 1. Study workflow

The bibliographic data were collected by employing two sets of terms related to the study domain. The first set contains 'machine learning' and 'deep learning' search keywords. The keyword 'deep learning' was included because this is an advanced form of machine learning that is currently growing due to the increasing data variety and volume. We used additional search keywords of 'pandemic', 'epidemic', 'endemic', and 'disease outbreak' to enrich the epidemiology term.

The bibliometric data were collected from the Scopus citation database using the determined keywords derived from the previous stage in the early data preparation stage. The data collection process resulted in 4,062 records of published articles. We applied some inclusion criteria for data selection, taking in articles published in 2000-2021 and written in English. Only publications in the form of journal papers, books, book chapters, conference papers, and reviews were included for analysis. Duplicated and incomplete records were removed from the dataset. Then we selected articles that were concerned with human diseases and excluded animal and plant diseases related articles. The details of the data selection process is presented in Figure 2. At the end of the data preparation process, 3,447 articles were selected for the next stage.

The bibliometric analysis utilized the Bibliometrix software package and Biblioshiny tools running under the R environment. The package provides tools for quantitative analysis for examining and visualizing bibliometric data. We conducted descriptive and network analysis in the last stage to develop knowledge maps and the research field's conceptual and intellectual structures [23]. Then, a machine learning approach called latent dirichlet allocation (LDA) was executed in this content analysis to explore the most prominent topics and clustered them into several issues. This approach was performed by extracting the abstracts of all articles that work based on the distributions of words. Using LDA, similar topics were examined across

articles and clustered by words that appeared salient in the dataset. We selected clusters that can shape issues related to the research domain. At the end of the stage, a complete reading of the articles was conducted to determine the proper topic hotspot for a particular cluster for data analysis.

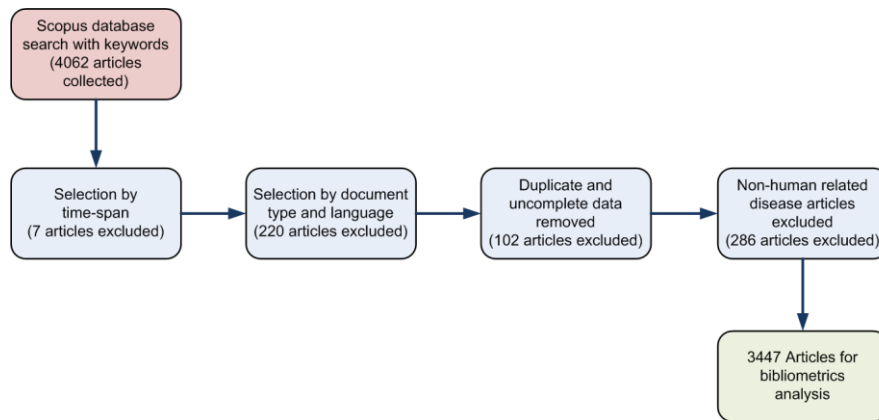


Figure 2. Selection process

## 4. RESULTS AND DISCUSSION

### 4.1. Scientific production

The dataset consists of 3,447 publications that are more than ninety per cent dominated by journal articles and proceeding papers, with few publications in books or book chapters. Figure 3 plots the evolution of the publications in this research area during the last two decades. In this period, it has been recorded in the history of epidemiology that there has been occurred several infectious disease outbreaks, including human immunodeficiency virus (HIV), hemagglutinin1 and neuraminidase1 (H1N1) Influenza, severe acute respiratory syndrome (SARS), middle east respiratory syndrome (MERS), Dengue, Chikungunya, and Zika [1] in addition to COVID-19. The figure shows that the volume of publications grew constantly and steadily between 2000 and 2012 and slightly increased starting from 2013. The increase was driven by the emergence of outbreaks of MERS, Ebola, and Zika. After Coronavirus was founded for the first time at the end of 2019, the trend had exponential growth. The COVID-19 pandemic has accelerated academic research in this field, yielding almost two-thirds of the articles in the dataset.

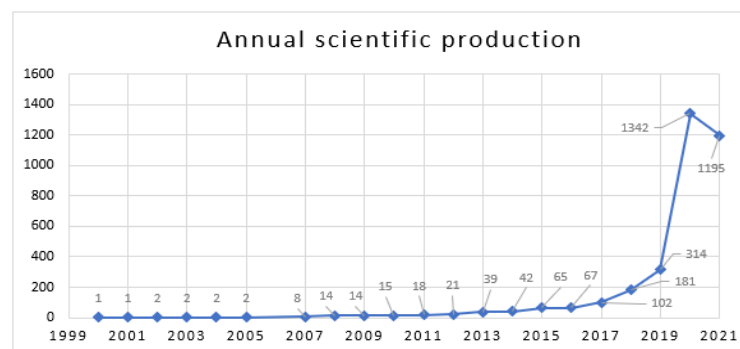


Figure 3. Annual production of publication

Scopus categorized articles in several subject areas. An article can be classified in more than one subject area, and it will be counted individually for each field. The top 15 subject areas with the most articles in the data set are shown a Table 1 arranged in descending order. The three major subject areas are computer science, medicine and engineering contributing more than 50% of the total articles. The rest of the articles are spread over various subject areas. Machine learning was developed by the community engaged in computer science and engineering. Hence as shown in the table, this subject area has contributed more than one-third to this study domain. It was noted that IEEE Access had published the most articles related to this subject area.

Table 1. Most research areas according to the number of articles

Subject Area	Number of articles	Percentage (%)
Computer Science	1,672	21.98
Medicine	1,380	18.14
Engineering	892	11.73
Biochemistry, Genetics & Molecular Biology	495	6.50
Mathematics	494	6.49
Decision Sciences	322	4.23
Environmental Science	248	3.26
Physics and Astronomy	246	3.23
Social Sciences	222	2.91
Agricultural and Biological Sciences	220	2.89
Material Science	195	2.56
Multidisciplinary	192	2.52
Immunology and Microbiology	134	1.76
Health Professions	133	1.74
Energy	107	1.40

4.2. Research interest and evolution

Research interest evolution begins by analyzing the occurrence of keywords, titles' words, and abstracts words that the authors most often used. The analysis results show that of the 6,855 author keywords in the dataset, two author keywords appear more than a thousand times, namely "machine learning" and "COVID-19". Regarding the topic of epidemiology and the world situation in the last two years, several authors use the terms "SARS-COV2" and "coronavirus" which refer to COVID-19 as frequently used keywords. Authors' keywords that occur frequently are also shown in Table 2. Figure 4 shows the word cloud that presents the most used author keywords indicated by text size. We can see that convolution neural network (CNN), long short term memory (LSTM), random forest, and support vector machine (SVM) are the authors' machine learning algorithms more widely used. Data mining techniques of prediction appeared prominent in the word cloud, including classification, sentiment analysis, and forecasting. Moreover, natural language processing (NLP), the branch of AI developed to process text or voice data, is also widely used in this study domain.

Table 2. Most frequently used words and keywords in the literature

Author keywords	n	Words in titles	n
machine learning	1,034	covid	1,401
COVID-19	1,028	learning	1,280
deep learning	571	machine	774
artificial intelligence	257	based	557
coronavirus	181	deep	505
epidemiology	156	data	410
sars-cov2	142	prediction	351
pandemic	121	detection	321
classification	98	analysis	314
prediction	96	pandemic	271



Figure 4. Word cloud

Research themes can be identified by analyzing a co-word network, a network analysis that aims to develop a conceptual structure of a particular scientific domain by mapping and clustering terms extracted from the collection of text data [23]. Figure 5 shows the co-word network using the author's keywords as the unit of analysis to build the conceptual structure of the dataset. There are 6,855 keywords identified that would be too many to fit on a chart. So, only 50 nodes with a three-occurrence threshold were set to get a

readable chart. The thickness of the connecting line indicates a proximity measure of the association of two keywords. The node size indicates the weight of the occurrence of the word.

The keywords are grouped into four interrelated clusters. The green "COVID-19" and the red "machine learning" have the largest node sizes, indicating that they are the most frequently occurring keywords. They also have the highest proximity, which indicates how often these keywords occur together. The red cluster consists of 15 words, the majority are machine learning techniques and methods, and the others are related to infectious diseases and outbreaks such as "dengue", "COVID-19 pandemic", and "epidemiology". The closest connections to the "machine-learning" node are the keywords "prediction", "classification" and "random forest".

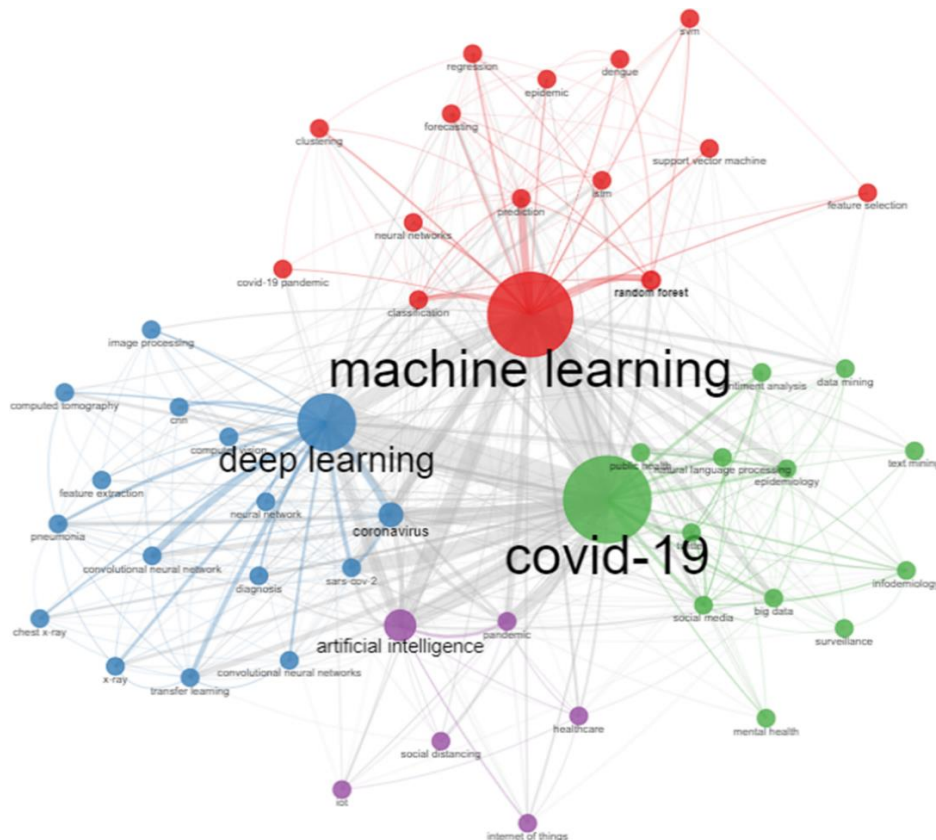


Figure 5. Co-word network

The green cluster corresponds to COVID-19 disease that can be divided into three sub-clusters. The first sub-cluster concerns data processing, including "data mining", "sentiment analysis", "natural language processing", "big data", and "text mining". The keywords "social media" and "Twitter" can be interpreted as the most widely used data sources. The last sub-group is related to pandemics such as "infodemiology", "surveillance", "epidemiology", "mental health", and "public health". Furthermore, we can see that "big data" has a strong relationship with "machine learning" and COVID-19 in this domain.

The blue cluster has the highest number of keywords consisting of 16 terms. The keywords were more related to the diagnosis of COVID-19 using deep learning interpreted by the words "coronavirus", "computer vision", "image processing", "computed tomography", "pneumonia", "x-ray", and others. It is identified that "convolutional neural networks" and "transfer learning" are associated more strongly with "deep learning" in addition to "coronavirus". Finally, the purple cluster is the smallest cluster on the co-word chart. In this cluster, "artificial intelligence" is the dominant keyword and associated closest with "pandemic". The keywords "health care", "social distancing", and "internet of things" were included in the cluster.

Co-word network analysis showed that machine learning has contributed to research in handling infectious diseases in the last two decades. Research topics of prediction, forecasting, classification, and diagnosis are the most frequently conducted in previous research. Classical algorithms such as the support vector machine and random forest were still interesting and used to solve many medical problems. The high

ability of deep learning to provide a transferable solution has driven the research in this domain to cover more complex problems. The convolution neural network algorithm and its advances have been widely used for analyzing medical images. In addition to medical records and image analysis, machine learning was also used for natural language processing and text mining for analysis of opinions posted on social media related to public health and mental health that were prominent when the world was facing the pandemic.

Figure 6 demonstrates the evolution of keywords and research direction over several periods. We divided the observed period into the time spans of 2000-2009, 2010-2019, 2020-2020, and 2021-2021. We considered the last two periods by allowing only one year because, during these two periods, the articles discussing machine learning for epidemiology were growing exponentially due to the COVID-19 pandemic. In the early period of 2000-2009, the Bayesian network and neural network were widely used in the research. Epistasis became a topic of interest in that period that elaborated on the interaction between different genes applied to molecular and quantitative genetics research for exploring complex diseases [24]. In this period, researchers were more concerned with epistasis and genetic problems.

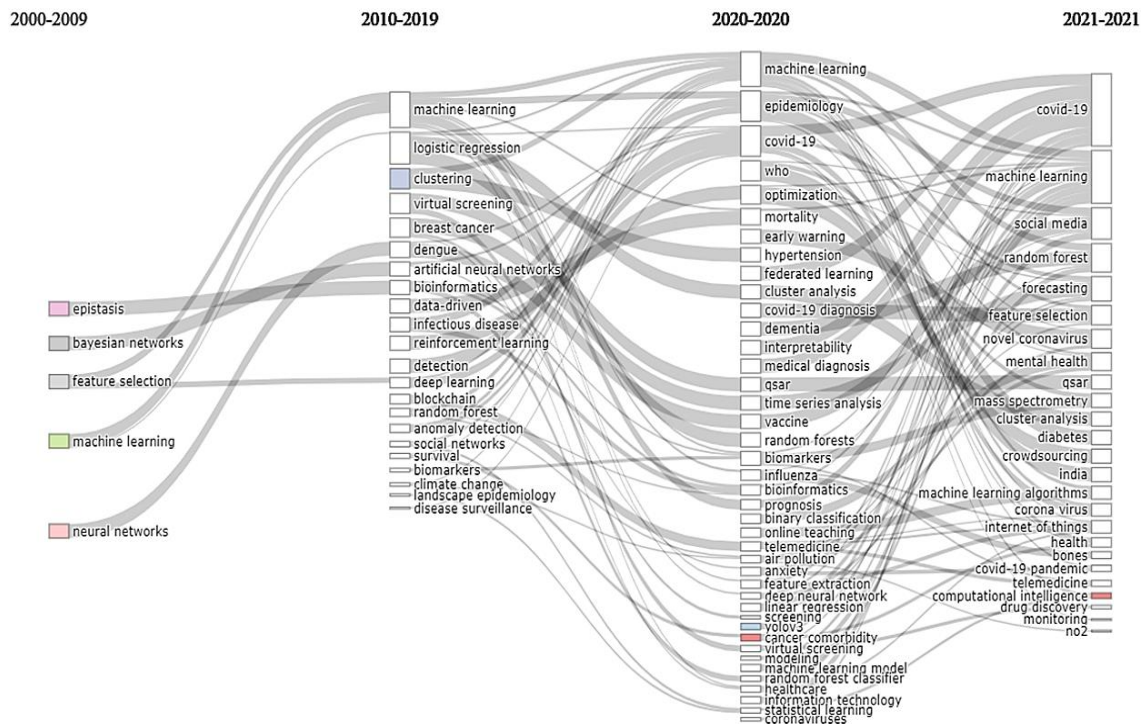


Figure 6. Evolution of keywords

During the period 2010-2019, logistic regression and random forest are widely applied in the research in addition to the artificial neural network. Analysis shows that many studies used data mining techniques of classification, clustering, and anomaly detection on health and bio-sciences. Deep learning and reinforcement learning have also facilitated the studies to process multimodal data from large and complex datasets. Breast cancer and dengue were two diseases that got more attention in machine learning research. The blockchain technology that was growing in the last of this period was applied to develop varied intelligent healthcare systems. Additionally, social media data was becoming a prominent source for disease surveillance.

When COVID-19 was identified as a global pandemic in 2020, the research on machine learning for this deadly virus was increased dramatically to date. Traditional machine learning and deep learning were still widely employed in COVID-19 research. The existence of immense data, decentralized in many distributed locations or remote devices, has encouraged researchers to find proper methods to cover many pandemic issues by harnessing the data. Federated learning is a machine learning method that involves statistical models trained over remote devices or soiled data centers while keeping data localized. This method is a distributed machine learning approach that enables models trained on a large corpus of decentralized data.

The analysis of the keyword evolution showed that the application of bioinformatics using machine learning became increasingly advanced after 2010. In these early years, research on cancer and dengue was prominent. The application of machine learning to treat cancer disease was concerned about the patient's survival. While dengue research focused on landscape epidemiology and the effects of climate change on the disease. Since the COVID-19 outbreak occurred, machine learning algorithms have been increasingly applied to deal with disease treatments and reduce the risk of comorbidities caused by hypertension, diabetes, cancer, and others. Because the symptoms of COVID-19 are almost similar to those of influenza, machine learning research on influenza has been continuously conducted to devise patient treatments and to discover drugs and vaccines for COVID-19. During this pandemic, research has also paid attention to the early warning of disease spread and health policy to prevent the wider spread of this outbreak.

### 4.3. Topic hotspots

The last stage of the study is to perform content analysis by applying LDA to the abstracts collected from the dataset. This machine learning-based text mining technique was developed to uncover hidden thematic structures and find the topic hotspots of the domain. This unsupervised probabilistic method employed a hierarchical Bayesian model providing a set of topics and the probabilities that represent the topics' strengths across the dataset [25]. This content analysis is intended to shed light on the main research areas that get more attention. Therefore, the research landscape can be exposed so that the future direction of the research can be drawn.

Initially, the analysis was designed to generate 20 topics from the abstracts of selected documents. After cleaning and stemming the dataset, the LDA model yielded topics that must be reshaped to get some prominent themes. We identified 12 topics regarding the domain of the study, as listed in Table 3. Some closest topics were merged due to their similar themes, and some others were removed because they were too weak to build a particular theme. The extracted topics were then distinguished into three clusters: 1) COVID-19 disease; 2) Miscellaneous diseases; and 3) public opinion on disease outbreaks. Content analysis was conducted to select the eligible documents that were strongly related to the identified topics. The following subsections discuss these clusters.

Table 3. Related topics covered by the selected articles

Topic	Top 20 prominent words	
Cluster 1: COVID-19 disease		
1	learning	
1	Case trend and spread prediction	covid, patient, prediction, model, mortality, risk, hospital, severe, spread, disease, infectious, confirm, clinic, death, identification, trend, forecast, outbreak, rate, and increase
2	Disease detection and health monitoring	covid, health, care, monitor, digital, service, device, IoT, solution, intelligence, sensor, smart, mobile, collect, medical, community, remote, real, time, and internet.
3	Medical imaging diagnosis	covid, image, model, detection, x-ray, cxr, accuration, disease, diagnosis, chest, patient, classification, pneumonia, case, infection, severe, medical, scan, screening, and spread.
4	Medicine, vaccine and immunity	covid, disease, drug, development, intelligence, prediction, medical health, new, model, spread, analysis, vaccine, effect, advance, world, virus, treatment, emergency, and clinic.
5	Viral infection	covid, antibody, patient, resistance, sequence, viral, response, severe, biomark, protein, pathogen, bacteria, immune, genome, antigen, detection, phenotype, strain, antibiotic, and infection.
6	Health protocol and policies	covid, image, mask, detection, classification, human, face, infection, risk, distance, social, control, model, identification, time, people, lockdown, policy, government, and spread.
Cluster 2: Miscellaneous disease		
7	Internal disease and risks	diabetic, risk, disease, liver, heart, cardiovascular, biomark, ecg, retina, age, organ, transplantation, factor, kidney, obesity, nutrition, eye, tissue, metabolism, and diet.
8	Cancer disease	cancer, disease, patient, clinic, prediction, survival, classification, risk, age, treatment, symptom, severe, observation, breast, death, care, brain, mortality, diagnosis, and epidemiology.
9	Influenza disease	factor, classification, prediction, influenza, health, flu, host, risk, human, anxiety, disease, country, pandemic, season, transmission, outbreak, area, virus, public, and environment.
10	Mosquito-borne disease	dengue, malaria, prediction, epidemic, blood, patient, analysis, climate, factor, spread, detection, fever, diagnostic, effect, classification, infection, incident, region, mosquito, and identification.
11	HIV and drug abuse	risk, HIV, association, estimation, identification, smoke, population, factor, effect, health, age, activity, birth, prediction, suicide, social, drug overdose, behavior, and opioid.
12	Cluster 3: Public opinion on disease outbreaks	health, social, tweet, media, public, covid, twitter, topic, identification, opioid, sentiment, pandemic, emotion, disease, analysis, online, surveillance, news, concern, and mental

#### 4.3.1. COVID-19 disease

The COVID-19 outbreak has opened up many research opportunities in many multi-discipline studies. Approximately 45% of the total articles in the dataset cover COVID-19 published in 2020-2021. Six



topics related to the COVID-19 disease cluster were obtained from the results of LDA topic modelling, namely: 1) case trend and spread prediction; 2) disease detection and health monitoring; 3) medical imaging diagnosis; 4) medicine, vaccine and immunity; 5) viral infection; and 6) health protocol and policies.

The studies on the spread of novel coronavirus in the early phase of the pandemic mainly focused on predicting new cases, recovery, and mortality, for global [26], [27] or country-specific scope [28]. Severity analysis of contaminated areas was also a concern for researchers [29]. Risk assessment and mapping for disaster management towards the pandemic were analyzed to prevent wider transmission and reduce the number of cases [30]. Regression and time series models were dominated in the previous studies complemented by statistical and susceptible, infected, and recovered (SIR) mathematical models [31] for dynamic or real-time prediction/forecasting purposes. Overall, the survey identified 344 articles in topic 1 discussing the coronavirus spread from the dataset. The journal of Chaos, Solitons, and Fractal were the most prolific journal on this topic.

Early detection of COVID-19 and health monitoring of patients are urgent issues to be concerned by health service authorities in order to isolate infected people, provide timely treatment [32], and also to prevent the spread of the disease [30]. There were 210 articles identified from the dataset discussing this topic. The detection of infected patients is very important for some prediction purposes, including patients with highly required immediate respiratory support [33], patients with a high risk of mortality [34], the mortality rate in comorbidities patients [35], patients' severity risk [36], and required intensive care unit (ICU) admission for high-risk patients [37]. A specimen test with real-time polymerase chain reaction (RT-PCR) is currently recognized as the standard test for confirmation of SARS-CoV-2 infection, delivering the highest accuracy for infectious detection compared to other testing methods [38]. However, this method still had some limitations because the test relies on a sample with a relatively high false-negative ratio and its expensive cost for countries with a lack of sufficient resources [39]. Previous studies on COVID-19 detected the virus infection based on data analysis of chest X-rays (CXR) and computerized tomography (CT) images [40], patients' voice, cough and breathing patterns [41], [42], electronic health records (EHR) [37], blood test results [43], and serum samples [38], [44].

AI, machine learning, and internet of thing (IoT) have been widely used for health monitoring during the SARS-CoV-2 outbreak. Telemedicine and telehealth could be adopted to monitor patients remotely, especially patients with medical emergencies. The human health condition can be monitored by tracking the health data such as body temperature, cough rate, respiratory rate, and blood oxygen saturation through an intelligent smartphone application that applied fog-based machine learning for diagnosis purposes [45]. Furthermore, the increasing number of COVID-19 patients has significantly impacted the needs of health facilities. Health care providers must manage overload conditions and allocate sufficient equipment during this emergency situation. Machine learning contributed to predicting the needs of hospital facilities such as ambulances availability [46], the required number of beds and mechanical ventilators [33], [47], hospital hygiene required [48], and optimization of critical medical supplies redistribution [49]. The most prolific journals for this topic are the Journal of Medical Internet Research, PLoS One, and the International Journal of Environmental Research and Public Health.

Whether someone is exposed to COVID-19 disease could be detected by analyzing his/her medical chest image [40], [50], [51]. This image-based analysis was used for screening the type of patients' respiratory disease and their severity. The results of this analysis were integrated with other patients' symptoms data to fit out the diagnosis. Multimodal diagnosis applications were developed by combining the medical image data with breathing sound and clinical data [52]. Moreover, the severity scoring of COVID-19 patients could be detected by analyzing their lung ultrasound videos [53]. Overall, there are more than 10% of articles in the dataset (391 articles) related to image-based COVID-19 diagnosis. The amount of research in this area increased significantly in the second year of the pandemic due to the soaring of cases. So, the need for rapid diagnosis with high accuracy provided by data processing techniques such as machine learning is necessary. The Journal of Scientific Reports, several journals under the IEEE publisher, and the Journal of Computers in Biology and Medicine were the top journals that published articles regarding this field of study. Deep learning techniques were the primary method that was dominantly applied in extracting complex information from chest medical images. Transfer learning was also applied in some past studies (29 articles) for the automatic detection of infected patients. Transfer learning is an alternative solution for the rapid development of smart applications, considering the rapid changes in virus mutation that can be more infectious and spread faster. The technique utilized a pre-trained model of a particular area, such as research on pneumonia, to be modified effectively for applications of a new disease, such as COVID-19 [54].

During the COVID-19 pandemic, machine learning has contributed to discovering medicines and vaccines to increase patients' survival rates. This review study identified 71 articles that covered the field. Bioinformatics applications have been developed for discovering, repositioning and repurposing drugs to find effective clinical treatments [55], [56]. The studies analyzed the interactions between compounds of drugs and proteins of SARS-CoV-2 and predicted the bio-activities that occurred. Analysis of the interaction

between the protein of the virus against drugs was widely done in past studies. The studies elaborated on the biochemical features and mutations of SARS-CoV-2 proteins regarding their impacts to develop drugs [57]. In attempting to suppress the severity, researchers also paid attention to identifying the potential inhibitors of SARS-CoV-2 main protease [58]. Some artificial intelligence applications were developed to repurpose available drugs for devising therapeutic strategies against COVID-19 [59]. In addition to drugs development, the role of vaccines is very important in reducing the number of cases and the severity caused by COVID-19. Computational intelligence using machine learning was applied in many past studies to design vaccines for COVID-19 [60]. Research on vaccine development aimed to predict and mitigate mutation threats of new variants of SARS-CoV-2 [61]. People's reaction toward the COVID-19 vaccine became the concern of researchers in observing to what extent the public responded to the announcement of the vaccine [62].

Virus characteristics and how the human body responds to this were the primary concern of viral infection topics. Forty-two articles discussed the topic, consisting of 19 published in 2020 and 23 in 2021 (until mid-year). Analytical Chemistry, the Indian Journal of Medical Research, Scientific Reports, and The Lancet Microbe were the most prolific source. SARS-CoV-2 has evolved to adapt the environmental changes through genetic mutations. Knowledge of virus evolution and transmission is essential during the pandemic to develop appropriate intervention strategies for the virus spreading control [63]. Several previous studies related to the SARS-CoV-2 genome signature evolution were: 1) identification of differences and similarities of viral variants based on genome sequence analysis [64]; 2) identification of protein interactions [65] and determination of the functions and pathways of proteins in biological processes [66]; 3) evaluation of viral mutation based on the protein sequence [67]; and 4) prediction of the SARS-CoV-2 mutation infectivity [68].

Furthermore, the severity of this SARS-CoV-2 infection is highly dependent on the host (human body) factors, such as age and immunity level [69]. Diagnostics of the host response to COVID-19 could be used for viral ribo nucleic acid (RNA) profiling [70] and identifying proteins as biomarkers, and the pathogenesis of COVID-19 [71]. People who have recovered from COVID-19 or received the COVID-19 vaccine were detected to have SARS-CoV-2 antibodies in their blood. Serological diagnosis based on antibody response shows that antibody and immunity of post-COVID-19 infected persons are increased [44]. However, the antibody response and the duration of immunity in post-COVID patients vary greatly depending on the individual health condition [72].

The main concerns discussed by previous research in topic 6 included the adherence to wearing masks, social. The main concerns discussed by previous research in topic 6 included the adherence to wearing masks, social distancing, and lockdown policy regarding the COVID-19 pandemic. The topic contained 63 articles published in 2020 (31 articles) and 2021 (32 articles until a mid-year). PLoS ONE, JMIR Public Health and Surveillance, and the International Journal of Environmental Research and Public Health were the most prolific sources on this topic. Twenty-one past studies have exposed the detection of face-mask-wearing conditions for human safety in coping with the pandemic. Most of them applied deep learning techniques for image analysis employing various architectures such as Yolov and MobileNet [73], [74]. To get automatic real-time data, the face-mask detection system was integrated with an IoT system [75], [76]. Finally, the study on topic mining to explore the crucial insights on public discourse against wearing a mask was performed by [77]. Based on the user-generated content on Twitter, a topic modelling technique of LDA was applied to address public concern about this health protocol. The results can support the decision-maker in reshaping the policy reducing public risk and improving public resilience.

#### 4.3.2. Miscellaneous diseases

The miscellaneous diseases cluster contains research on various types of infectious disease outbreaks for the last two decades. Approximately 619 of the total 3,447 articles cover topics in this cluster. This cluster's most widely discussed topic is an internal disease, with 198 articles. About 141 articles on cancer disease, 109 articles on the topic of influenza disease, 96 articles on mosquito-borne disease, and 75 articles on HIV and drug abuse.

Internal disease deals with various diseases and health problems affecting the human internal organs. Of the 198 articles, the two most discussed in internal disease are cardiovascular disease (heart and blood vessels) with 70 articles and endocrine disease (disorders of the endocrine system) with 51 articles. The rest discussed geriatrics (care of the elderly), gastroenterology (digestive system, liver, and gallbladder), nephrology (kidneys), and obesity which affected the health of internal organs. Cardiovascular diseases affect the heart and blood-vascular system. Heart attacks and strokes were recognized as the main cause (estimated 85%) of the 17.9 million deaths in 2019 [2]. Machine learning and deep learning have been widely used to identify, detect, predict, and forecast several types of cardiovascular disease. Some previous studies explored cardiology diseases, including heart disease detection [78], prediction of heart failure [79], diagnosis of atrial fibrillation [80], and cardiovascular mortality risk [81]. In addition to cardiovascular disease, endocrine system disorders, especially diabetes mellitus, were also widely discussed. According to WHO, diabetes is

one of the leading causes of death in the world, with 1.6 million deaths every year [2]. Most of the cases occurred in low-and middle-income countries [82]. Research on diabetes using machine learning and deep learning approaches has been carried out for the detection of diabetes types [83], investigation of the association between obesity and diabetes risk [84], algorithms for image classification for detection and screening of diabetic retinopathy [85]. PLoS ONE was the journal that published the most articles on internal disease, followed by the American Journal of Cardiology, Journal of Medical Internet Research, BMJ Open Diabetes Research and Care, Diabetologia, and Journal of the American College of Cardiology.

Cancer is the sixth leading cause of death globally, being the second most discussed topic of miscellaneous disease with 141 articles. This topic included 24 articles on carcinoma, 22 articles on breast cancer, 20 articles on lung cancer, 11 articles on prostate cancer, and the rest were brain cancer, leukemia, osteosarcoma, thyroid cancer, cervical cancer, and others. The top sources that published the topic are Cancers (MDPI) and the International Journal of Medical Informatics, Frontiers in Oncology, and Journal of the American Medical Informatics Association. Most of the articles on this topic use machine learning and deep learning approaches for prediction (81 articles), detection and diagnosis (16 articles). In particular, regarding the prediction approaches, some previous research elaborated the following predictions: patient survival likelihood [86], cancer prognosis [87], treatment plans [88], and metastasis (cancer spreads to a different body part) [89]. And the rest were the predictions of radiation risk in breast cancer [90], mortality risk [91], malignancy (local spreading) [92], and recurrences [93].

Historical data on global health shows that influenza pandemics have hit the world several times, infecting many people and causing many deaths. Approximately 75% of confirmed cases are caused by infection of the type A influenza viruses, the viruses that are capable of infecting animals [94]. Type A influenza viruses were highly contagious through poultry and human contact and often become epidemics in tropical countries. Over the past two decades, 2020 was the most prolific publication year on this topic, with 33 articles from a total of 96 articles. Most of the articles discussed machine learning and deep learning approach to predict and forecast virus spread [95], the likelihood of vaccinated patients [96], and the vaccine effectiveness [97]. The rest articles discussed the classification of tropism protein signature for influenza virus identification [98], classification of symptoms for diagnosis of suspected people [99], and early warning detection of infected patients [100].

Mosquito-borne diseases are transmitted through the bite of infected mosquitoes to human bodies that may cause epidemics, especially in countries with tropical and subtropical climates [101]. Diseases carried by mosquitoes are caused by parasites (malaria) or viruses (dengue, zika, yellow fever, West Nile, and others). The disease has caused many deaths in several countries over the world. In the last two decades, dengue cases have increased more than 8-fold, from 505,430 cases and 960 deaths in 2000 to 5,2 million cases and 4,032 deaths in 2019 [2]. In 2019 there were 229 million cases of new malaria infection worldwide, with 94% occurrences in the African region [2]. The number of articles on this topic began to increase in 2018 (25 articles) and reached 55 articles in 2020. PLoS Neglected Tropical Diseases published the most articles, followed by Malaria Journal and the Journal of Communications in Computer and Information Science. This review study identified that climate, meteorology, geography, and environment were important predictor factors that affected the distribution of mosquitoes as disease transmitters [102]. Thirty-nine articles in the dataset discussed these predictor factors. These factors were used for the prediction and forecasting of the number of cases, disease spread, and the impact of disease infection [103]. Further, the historical case data, population density, and human mobility were employed to predict the number of cases and areas at risk [104]. Machine learning and deep learning methods were also used for classification of infection severity [105], diagnosis of malaria parasites determined by the analysis of blood smears [106], detection of antibody responses to vaccines [107], and the role of mosquitoes as disease vectors [108].

HIV and drug abuse are the last topics in the miscellaneous disease cluster discussed in this subsection. HIV is a virus that damages the human body's immune system. The virus infections can lead to acquired immune deficiency syndrome (AIDS) if the infected people are not treated immediately. Around 36.3 million human lives have been lost due to HIV. In 2020, about 1.5 million people have been infected [2]. Research on this topic has been conducted since 2009 for the last two decades. AIDS (London, UK), PLoS ONE, and BMJ are the most sourced publishing on this topic. Currently, a cure for HIV disease has not yet been found, making it a serious global public health problem that could lead to epidemics. Injected drug addicts [109] and opioid abuse [110] mainly triggered the occurrence of the HIV epidemic. Another severe problem in dealing with the HIV epidemic is drug resistance. The SVM classification algorithm has been applied to predict drug resistance, and anti-HIV-1 [111]. Other learnings and deep learning algorithms have been widely used to: 1) predict HIV incidence and infection [112]; 2) predict drugs addiction and overdose tendencies [113]; 3) detect the likelihood of suicide for opioid users [114]; and 4) identify people or areas at high risk of HIV infection [115].

### 4.3.3. Public opinion on disease outbreaks

The last cluster consisted of 232 out of 3,447 total articles discussing public concerns related to disease outbreaks. Research on this topic increased starting in 2011. The topic became fascinating to researchers due to the various public responses to health protocols when the pandemic occurred. User-generated contents posted on social media were the prevalent dataset used by previous research because social media is recognized as the fastest, easiest, and most widely used platform to share news, opinions, and emotions. Big data technology and NLP enabled to process of massive amounts of social media data to identify some interesting public concerns. Sentiment analysis could reveal public presumptions and emotions of people related to the COVID-19 outbreak [116], [117]. Other public opinions were widely discussed in this cluster were public mental health [118], vaccines and anti-vaccines opinion [119], [120], HIV prevention [121], fake news and misinformation detection [122], and hate speech and racial bullying related to the outbreak [123]. Moreover, social media data was also widely used for disease spread predictions [124], new cases and event detection [125], and illicit opioid and drug abuse detection [126].

This survey study showed that machine learning had covered many research areas in tackling disease outbreaks using various datasets and algorithms. Private or public medical data were employed by the algorithms to solve medical problems. Table 4 presents the selected machine learning approaches learned from this study categorized into similar data types (see in Appendix).

## 5. LIMITATIONS

We acknowledge that this work has some limitations in presenting the knowledge. First, our survey on the contribution of machine learning to tackle disease outbreaks is restricted to publications in the English language. Publications in other languages certainly should be part of the domain research body of knowledge. Furthermore, the articles to be reviewed were selected by criteria of pre-determined keywords and a particular period of publications, which leads to the possibility that more relevant articles should be considered in building the knowledge. Even though the Scopus citation database covers more articles compared with other databases such as Web of Science or PubMed, the use of only the Scopus database becomes another limitation. Lastly, the discussions of the topic hotspots referred to prominent keywords derived from the LDA process only involve articles with the most relevant and significant contribution from the authors' point of view.

## 6. CONCLUSION

The study of epidemiology is progressing to protect public health and deliver the highest possible public health care services, enabling machine learning to contribute more to tackling many disease outbreaks. The emergence of new viruses and virus mutations has accelerated research in this domain since 2013. The trend was extremely increased when SARS-CoV-2 was detected. Traditional machine learning and deep learning have been widely used in previous research with various supervised, unsupervised, and reinforcement learning techniques. Additionally, federated learning, transfer learning, and ensemble learning were applied in many studies to reduce the complexity and get higher accuracy. Integrated with IoT technology, machine learning has encouraged the development of telehealth and telemedicine. The review study reveals that the scientific structure of this domain is dominated by machine learning research on COVID-19 diseases and miscellaneous diseases caused by pathogens or some genetic factors. A huge amount of multimodal medical data was used by previous studies to predict, forecast, classify, or screen resolving many problems of diseases, including epidemiological surveillance, diagnosis, treatment, health monitoring, epidemic management, viral infection, and pathogenesis. Public opinions towards new diseases are also an interesting topic for researchers in addition to the public perceptions in response to the health protocol and policies to prevent the spread of diseases. Research on epidemiology is still challenging, and bioinformatics applying machine learning still has many opportunities to provide solutions for health and medicine. Virus genomics and evolution open up to be studied. Pathogen and drug discoveries inquired to face new threats of diseases in the future. Nevertheless, patient management and public health require continuous improvement. Hence, machine learning is necessary to be harnessed in epidemiology for disease outbreak handling.

## ACKNOWLEDGEMENTS

The authors wish to thank the members of the Information Retrieval Research Group at the Research Centre for Data and Information Sciences, National Agency of Research and Innovation (Indonesia) for their support throughout this work.

## APPENDIX

Table 4. Machine learning approaches for disease outbreaks

No.	Study aims	Data	Model/Algorithm
<b>Electronic medical record (EMR) dataset</b>			
1.	To predict mechanical ventilation and mortality COVID-19 patients [33]	EMR	XGBoost and CatBoost
2.	To predict physiological damage and death up to the next 20 days [34]	EMR	Ensemble model (logistic regression (LR), SVM, gradient boosting, decision tree (DT), and neural network)
3.	To predict cardiovascular diseases [78]	Cardiac disease dataset	Recursion enhanced random forest (RF)-improved linear model (RFRF-ILM), and internet of medical things (IoMT)
4.	To predict the mortality risk and heart failure in hospitalized patients [80]	treatment of preserved cardiac function heart failure with an aldosterone antagonist (TOPCAT) data	LR, RF, and gradient descent boosting
5.	To predict the fatality of acute myocardial infarction (AMI) patients [81]	AMI data	Deep learning for AMI (DAMI), LR, and RF
6.	To classify diabetes disease with a new hybrid intelligent system [83]	Pima Indian Diabetes dataset	EM clustering, incremental PCA and incremental SVM
7.	To prognosis treatment decisions by analyzing pathological microscopic features [87]	Sun Yat-Sen University Cancer Center (data on patients with nasopharyngeal carcinoma (NPC))	Deep feed-forward neural network (DeepSurv)
8.	To predict Hepatocellular carcinoma (HCC) recurrences [93]	Clinical data of patients with HCC	Novel Bayesian network
9.	To detect malaria-infected red blood cells [105], [106]	National Institute of Health (NIH) malaria dataset, Broad Institute malaria dataset.	VGG16 and CNN
10.	To identify person at high risk of HIV [115]	HIV testing data from rural Kenya and Uganda	Super learner ensemble
<b>Surveillance dataset</b>			
11.	To predict global or country-based COVID-19 cases [26], [28], [29], [31]	Time series COVID-19 dataset	ANN-grey wolf optimizer (GWO), marine predators algorithm (MPA)-ANFIS, convolutional auto encoder (CAE) and AL (autoencoder LSTM)-CNN, SIR model MLP
12.	To forecast the number of beds required as Cov-19 cases [47]	Kaggle data	MLP
13.	To optimize the redistribution of critical medical supplies [49]	Statistics data from Centers for Disease Control and Prevention	LSTM
14.	To predict the survival rate of patients with oral and pharyngeal cancer (OPCs) [66]	the surveillance, epidemiology, and end results (SEER) database	Survival tree (ST), RF, conditional inference forest (CF), cox proportional, and hazard models
15.	To predict lymph node metastases for colorectal cancer patients [89]	SEER database	LR, XGBoost, k-NN, regression trees, SVM, ANN, and RF
16.	To predict whether a person has been vaccinated against H1N1 and seasonal flu [96]	The National H1N1 Flu survey (NHFS)	MIBox, TPOT, polynomial feature, RF, MLP, LR, DT, XGBoost and CatBoost
17.	To predict dengue outbreaks [102]	Dengue surveillance data	SVR, gradient boosting, GAM, negative binomial regression, LASSO, and LR
18.	To describe HIV trends and predict their occurrence [112]	Chinese Center for Disease Control and Prevention Database	LSTM, ARIMA, GRNN, and exponential smoothing
<b>Images Dataset</b>			
19.	To early detect COVID-19 patient [40]	Chest X-ray (CXR) images	ResNet and CNNet
20.	To analyze and categorize COVID-19 X-ray images [54]	X-Ray images dataset	EfficientNet, MobileNet, Xception, InceptionV3, and VGG19
21.	To extract CT masks to improve the diagnosis of COVID-19 [51]	Computed tomography (CT) images	Multi-agent deep reinforcement learning
22.	to diagnose COVID-19 using breath sounds, chest X-ray (CXR) [52]	CXR images data	InceptionV3-MLP
23.	To identify the species of the mosquitoes gender [108]	Data images of anesthetized/dead mosquito	YOLO
<b>Voice Dataset</b>			
24.	To detect Covid-19 through voice analysis of speech or coughed [41], [42]	Coswara and virufy database	Naïve Bayes (NB), BayesNet (BN), SVM, stochastic gradient descent (SGD), KNN, locally weighted learning (LWL), RNN, Adaboost, Bagging algorithms, One-R, decision table, DT, and REPTree,
<b>IoT generated/video dataset</b>			
25.	To apply IoT in healthcare and physical distance monitoring for pandemic situations [45]	Sensor data	ANFIS, DT, and SVM
26.	To detect face masks for infection spread control [74], [75]	Camera-generated data	VGG-16, MobileNetV2, inception V3, ResNet-50, and CNN-transfer learning

Table 5. Machine learning approaches for disease outbreaks (continue)

No.	Study aims	Data	Model/Algorithm
27.	To predict the dengue incidence based on human mobility [104]	Mobile network big data	ANN and XGBoost
<b>Protein/genome dataset</b>			
28.	To identify the protein-ligand interactions for a specific drug [57]	Genome sequence dataset	Feed-Forward DNN
29.	To predict the anti-viral activities of resultant compounds [58]	Protein data bank	AutoQSAR algorithm
30.	To predict COVID-19 vaccine candidates [60]	NCBI GenBank	Vaxign-ML
31.	To analyses genomic mutations of SARS-CoV-2 [63]	NCBI GenBank	RNN
32.	To predicts the protein interactions between SARS-COV-2 virus and human proteins [65]	Non-structural protein data	BiRNN
33.	To predict the effect of mutations on SARS-CoV-2 infectivity [68]	Protein and CFR (case fatality rate) data	DeepNEU: RNN, cognitive maps (CM), SVM and genetic algorithm (GA).
34.	To improve understanding and treatment of obesity and diabetes [84]	Qatar Biobank	PCA, RF, and gradient boosting
35.	To identify possible zoonotic influenza-A viruses [98]	Protein sequence-Influenza Research Database	RF
36.	To predict anti-HIV-1 peptides [111]	Amino acid sequences data	SVM
<b>user generated contents/news dataset</b>			
37.	To detect dengue and flu outbreaks [73]	Twitter data	RF, KNN, SVM, and DT
38.	To detect suicidality among opioid users [114]	Reddit data	RNN and CNN
39.	To identify HIV-associated patterns from large sets of social data [121]	Large collections of social media data	LR, RF, and ridge regression
40.	To detect misleading information related to COVID-19 [122]	WHO, UNICEF news/report	DT, KNN, LR, SVM, Multinomial NB, Bernoulli NB, Perceptron NN, Ensemble RF, XGBoost
41.	To detect cyberbullying regarding the pandemic [123]	Instagram, Facebook, Twitter, and Youtube	CNN, capsule network (CAPSNET), and deep NN
<b>Fusion dataset</b>			
42.	To measure the number of emergency ambulance required during an emergency situation [46]	Integrated data: environmental data, the localization data of mobile phone users, and the historical EAD data	LSTM

## REFERENCES

- [1] L. A. Reperant and A. D. M. E. Osterhaus, "AIDS, Avian flu, SARS, MERS, Ebola, Zika... what next?," *Vaccine*, vol. 35, no. 35, pp. 4470–4474, Aug. 2017, doi: 10.1016/j.vaccine.2017.04.082.
- [2] WHO, "World Health Assembly." [Online] Available: <https://www.who.int/>.
- [3] D. Riswantini, E. Nugraheni, A. Arisal, P. H. Khotimah, D. Munandar, and W. Suwarmingsih, "Big data research in fighting COVID-19: Contributions and techniques," *Big Data and Cognitive Computing*, vol. 5, no. 3, p. 30, 2021, doi: 10.3390/bdcc5030030.
- [4] N. Pearce, "Traditional epidemiology, modern epidemiology and public health," *Epidemiologia e Prevenzione*, vol. 21, no. 2, pp. 678–683, 1997, doi: 10.2105/AJPH.86.5.678.
- [5] M. Supriya and V. K. Chattu, "A review of artificial intelligence, big data, and blockchain technology applications in medicine and global health," *Big Data and Cognitive Computing*, vol. 5, no. 3, p. 41, 2021, doi: 10.3390/bdcc5030041.
- [6] Z. Ahmed, K. Mohamed, S. Zeeshan, and X. Q. Dong, "Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine," *Database*, vol. 2020, pp. 1–35, 2020, doi: 10.1093/database/baaa010.
- [7] N. Schwalbe and B. Wahl, "Artificial intelligence and the future of global health," *The Lancet*, vol. 395, no. 10236, pp. 1579–1586, May 2020, doi: 10.1016/S0140-6736(20)30226-9.
- [8] J. A. Roth, M. Battegay, F. Juchler, J. E. Vogt, and A. F. Widmer, "Introduction to Machine Learning in Digital Healthcare Epidemiology," *Infection Control & Hospital Epidemiology*, vol. 39, no. 12, pp. 1457–1462, 2018, doi: 10.1017/ice.2018.265.
- [9] Z. A. A. Alyasseri et al., "Review on COVID-19 diagnosis models based on machine learning and deep learning approaches," *Expert systems*, vol. 39, no. 3, pp. 1–32, 2022, doi: 10.1111/exsy.12759.
- [10] L. R. Jorm, "Commentary: Towards machine learning-enabled epidemiology," *International Journal of Epidemiology*, vol. 49, no. 6, pp. 1770–1773, 2020, doi: 10.1093/ije/dyaa242.
- [11] S. Rose, "Intersections of machine learning and epidemiological methods for health services research," *International Journal of Epidemiology*, vol. 49, no. 6, pp. 1763–1770, 2020, doi: 10.1093/ije/dyaa035.
- [12] R. Alfred and J. H. Obit, "The roles of machine learning methods in limiting the spread of deadly diseases: A systematic review," *Heliyon*, vol. 7, no. 6, p. e07371, Jun. 2021, doi: 10.1016/j.heliyon.2021.e07371.
- [13] A. Gupta and R. Katarya, "Social media based surveillance systems for healthcare using machine learning: A systematic review," *Journal of Biomedical Informatics*, vol. 108, p. 103500, Aug. 2020, doi: 10.1016/j.jbi.2020.103500.
- [14] J. Jeon, G. Baruah, S. Sarabadani, and A. Palanica, "Identification of risk factors and symptoms of COVID-19: Analysis of biomedical literature and social media data," *Journal of medical Internet research*, vol. 22, no. 10, p. e20509, 2020, doi: 10.2196/20509.
- [15] L. Fernandez-Luque and M. Imran, "Humanitarian health computing using artificial intelligence and social media: A narrative literature review," *International Journal of Medical Informatics*, vol. 114, pp. 136–142, Jun. 2018, doi: 10.1016/j.ijmedinf.2018.01.015.
- [16] S. F. Tsao, H. Chen, T. Tisseverasinghe, Y. Yang, L. Li, and Z. A. Butt, "What social media told us in the time of COVID-19: a

- scoping review," *The Lancet Digital Health.*, vol. 3, no. 3, pp. e175–e194, Mar. 2021, doi: 10.1016/S2589-7500(20)30315-0.
- [17] J. Sak and M. Suchodolska, "Artificial intelligence in nutrients science research: A review," *Nutrients*, vol. 13, no. 2, pp. 1–17, 2021, doi: 10.3390/nu13020322.
- [18] S. Basu, K. T. Johnson, and S. A. Berkowitz, "Use of Machine Learning Approaches in Clinical Epidemiological Research of Diabetes," *Current diabetes reports*, vol. 20, no. 12, pp. 1–19, 2020, doi: 10.1007/s11892-020-01353-5.
- [19] G. Kim and M. Bahl, "Assessing Risk of Breast Cancer: A Review of Risk Prediction Models," *Journal of breast imaging*, vol. 3, no. 2, pp. 144–155, 2021, doi: 10.1093/jbi/wbab001.
- [20] S. A. Lee, C. I. Jarvis, W. J. Edmunds, T. Economou, and R. Lowe, "Spatial connectivity in mosquito-borne disease models: A systematic review of methods and assumptions," *Journal of the Royal Society Interface*, vol. 18, no. 178, p. 20210096, 2021, doi: 10.1098/rsif.2021.0096.
- [21] M. H. Alsharif, Y. H. Alsharif, S. A. Chaudhry, M. A. Albream, A. Jahid, and E. Hwang, "Artificial intelligence technology for diagnosing COVID-19 cases: A review of substantial issues," *European Review for Medical and Pharmacological Sciences*, vol. 24, no. 17, pp. 9226–9233, 2020, doi: 10.26355/eurrev\_202009\_22875.
- [22] A. A. Chadegani *et al.*, "A comparison between two main academic literature collections: Web of science and scopus databases," *Asian social science*, vol. 9, no. 5, pp. 18–26, 2013, doi: 10.5539/ass.v9n5p18.
- [23] M. Aria and C. Cuccurullo, "bibliometrix: An R-tool for comprehensive science mapping analysis," *Journal of Informetrics*, vol. 11, no. 4, pp. 959–975, Nov. 2017, doi: 10.1016/j.joi.2017.08.007.
- [24] H. J. Cordell, "Epistasis: What it means, what it doesn't mean, and statistical methods to detect it in humans," *Human molecular genetics*, vol. 11, no. 20, pp. 2463–2468, 2002, doi: 10.1093/hmg/11.20.2463.
- [25] D. M. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *Advances in Neural Information Processing Systems*, no. 14, 2002.
- [26] S. Ardabili, A. Mosavi, S. S. Band, and A. R. Varkonyi-Koczy, "Coronavirus Disease (COVID-19) Global Prediction Using Hybrid Artificial Intelligence Method of ANN Trained with Grey Wolf Optimizer," *2020 IEEE 3rd International Conference and Workshop in Obuda on Electrical and Power Engineering (CANDO-EPE)*, 2020, pp. 000251–000254, doi: 10.1109/CANDO-EPE51100.2020.9337757.
- [27] O. I. Obaid, M. A. Mohammed, and S. A. Mostafa, "Long short-term memory approach for coronavirus disease prediction," *Journal of Information Technology Management*, vol. 12, pp. 11–21, 2021, doi: 10.22059/jitm.2020.79187.
- [28] M. A. A. Al-Qaness, A. A. Ewees, H. Fan, L. Abualigah, and M. A. Elaziz, "Marine predators algorithm for forecasting confirmed cases of COVID-19 in Italy, USA, Iran and Korea," *International journal of environmental research and public health*, vol. 17, no. 10, pp. 1–14, 2020, doi: 10.3390/ijerph17103520.
- [29] I. H. Kao and J. W. Perng, "Early prediction of coronavirus disease epidemic severity in the contiguous United States based on deep learning," *Results in Physics*, vol. 25, p. 104287, Jun. 2021, doi: 10.1016/j.rinp.2021.104287.
- [30] R. Elmoro *et al.*, "Risk and Protective Factors in the COVID-19 Pandemic: A Rapid Evidence Map," *Frontiers in public health*, vol. 8, no. 787, pp. 1–12, 2020, doi: 10.3389/fpubh.2020.582205.
- [31] S. A. Alanazi, M. M. Kamruzzaman, M. Alruwaili, N. Alshammari, S. A. Alqahtani, and A. Karime, "Measuring and Preventing COVID-19 Using the SIR Model and Machine Learning in Smart Health Care," *Journal of healthcare engineering*, vol. 2020, 2020, doi: 10.1155/2020/8857346.
- [32] N. Darapaneni *et al.*, "A novel machine learning based screening method for high-risk covid-19 patients based on simple blood exams," *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, 2021, pp. 1–6, doi: 10.1109/IEMTRONICS52119.2021.9422534.
- [33] L. Yu *et al.*, "Machine learning methods to predict mechanical ventilation and mortality in patients with COVID-19," *PLoS One*, vol. 16, no. 4, pp. 1–18, Apr. 2021, doi: 10.1371/journal.pone.0249285.
- [34] Y. Gao *et al.*, "Machine learning based early warning system enables accurate mortality risk prediction for COVID-19," *Nature communications*, vol. 11, no. 1, pp. 1–10, 2020, doi: 10.1038/s41467-020-18684-2.
- [35] M. Kivrak, E. Guldogan, and C. Colak, "Prediction of death status on the course of treatment in SARS-COV-2 patients with deep learning and machine learning methods," *Computer Methods and Programs in Biomedicine*, vol. 201, p. 105951, Apr. 2021, doi: 10.1016/j.cmpb.2021.105951.
- [36] S. Tabik *et al.*, "COVIDGR Dataset and COVID-SDNet Methodology for Predicting COVID-19 Based on Chest X-Ray Images," in *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 12, pp. 3595–3605, Dec. 2020, doi: 10.1109/JBHI.2020.3037127.
- [37] J. L. Izquierdo, J. Ancochea, and J. B. Soriano, "Clinical Characteristics and Prognostic Factors for Intensive Care Unit Admission of Patients with COVID-19: Retrospective Study Using Machine Learning and Natural Language Processing," *Journal of medical Internet research*, vol. 22, no. 10, p. e21801, 2020, doi: 10.2196/21801.
- [38] F. M. Nachtigall, A. Pereira, O. S. Trofymchuk, and L. S. Santos, "Detection of SARS-CoV-2 in nasal swabs using MALDI-MS," *Nature biotechnology*, vol. 38, no. 10, pp. 1168–1173, 2020, doi: 10.1038/s41587-020-0644-7.
- [39] G. Yin *et al.*, "An efficient primary screening of COVID-19 by serum Raman spectroscopy," *Journal of Raman Spectroscopy*, vol. 52, no. 5, pp. 949–958, 2021, doi: 10.1002/jrs.6080.
- [40] A. Keles, M. B. Keles, and A. Keles, "COVID-19-CNNNet and COVID-19-ResNet: Diagnostic Inference Engines for Early Detection of COVID-19," *Cognitive Computation*, pp. 1–11, Jan. 2021, doi: 10.1007/s12559-020-09795-5.
- [41] L. Verde, G. De Pietro, A. Ghoneim, M. Alrashoud, K. N. Al-Mutib, and G. Sannino, "Exploring the Use of Artificial Intelligence Techniques to Detect the Presence of Coronavirus Covid-19 through Speech and Voice Analysis," *IEEE Access*, vol. 9, pp. 65750–65757, 2021, doi: 10.1109/ACCESS.2021.3075571.
- [42] K. Feng, F. He, J. Steinmann and I. Demirkiran, "Deep-learning Based Approach to Identify Covid-19," *SoutheastCon 2021*, 2021, pp. 1–4, doi: 10.1109/SoutheastCon45413.2021.9401826.
- [43] N. C. Cady *et al.*, "Multiplexed detection and quantification of human antibody response to COVID-19 infection using a plasmon enhanced biosensor platform," *Biosensors and Bioelectronics*, vol. 171, p. 112679, Jan. 2021, doi: 10.1016/j.bios.2020.112679.
- [44] J. Rosado *et al.*, "Multiplex assays for the identification of serological signatures of SARS-CoV-2 infection: an antibody-based diagnostic and machine learning study," *The Lancet Microbe*, vol. 2, no. 2, pp. e60–e69, Feb. 2021, doi: 10.1016/S2666-5247(20)30197-X.
- [45] S. S. Vedaei *et al.*, "COVID-SAFE: An IoT-based system for automated health monitoring and surveillance in post-pandemic life," *IEEE Access*, vol. 8, pp. 188538–188551, Oct. 2020, doi: 10.1109/ACCESS.2020.3030194.
- [46] E. A. Rashed, S. Kodera, H. Shirakami, R. Kawaguchi, K. Watanabe, and A. Hirata, "Knowledge discovery from emergency ambulance dispatch during COVID-19: A case study of Nagoya City, Japan," *Journal of Biomedical Informatics*, vol. 117, no. February, p. 103743, May 2021, doi: 10.1016/j.jbi.2021.103743.
- [47] S. Nagpal, V. A. Athavale, A. K. Saini, and R. Sharma, "Indian Health Care System is Ready to Fight Against COVID-19 A Machine Learning Tool for Forecast the Number of Beds," *2020 Sixth International Conference on Parallel, Distributed and Grid*

- Computing (PDGC)*, 2020, pp. 61–65, doi: 10.1109/PDGC50313.2020.9315825.
- [48] A. M. Qahtani, B. M. Alouffi, H. Alhakami, S. Abuayeid, and A. Baz, "Predicting Hospitals Hygiene Rate during COVID-19 Pandemic," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 12, pp. 815–823, 2020, doi: 10.14569/IJACSA.2020.0111294.
- [49] B. P. Bednarski, A. D. Singh, and W. M. Jones, "On collaborative reinforcement learning to optimize the redistribution of critical medical supplies throughout the COVID-19 pandemic," *Journal of the American Medical Informatics Association*, vol. 28, no. 4, pp. 874–878, 2021, doi: 10.1093/jamia/ocaa324.
- [50] J. N. Hasoon *et al.*, "COVID-19 anomaly detection and classification method based on supervised machine learning of chest X-ray images," *Results in Physics*, vol. 31, p. 105045, 2021, doi: 10.1016/j.rinp.2021.105045.
- [51] H. Allioui *et al.*, "A Multi-Agent Deep Reinforcement Learning Approach for Enhancement of COVID-19 CT Image Segmentation," *Journal of personalized medicine*, vol. 12, no. 2, p. 309, 2022, doi: 10.3390/jpm12020309.
- [52] U. Sait *et al.*, "A deep-learning based multimodal system for Covid-19 diagnosis using breathing sounds and chest X-ray images," *Applied Soft Computing*, vol. 109, p. 107522, Sep. 2021, doi: 10.1016/j.asoc.2021.107522.
- [53] F. Mento *et al.*, "Deep learning applied to lung ultrasound videos for scoring COVID-19 patients: A multicenter study," *The Journal of the Acoustical Society of America*, vol. 149, no. 5, pp. 3626–3634, 2021, doi: 10.1121/10.0004855.
- [54] H. Lu, S. A. Hewakananage, and Y. Miao, "Transfer Learning from Pneumonia to COVID-19," *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, 2020, pp. 1–6, doi: 10.1109/CSDE50874.2020.9411550.
- [55] K. Cooper *et al.*, "Novel Development of Predictive Feature Fingerprints to Identify Chemistry-Based Features for the Effective Drug Design of SARS-CoV-2 Target Antagonists and Inhibitors Using Machine Learning," *ACS Omega*, vol. 6, no. 7, pp. 4857–4877, 2021, doi: 10.1021/acsomega.0c05303.
- [56] S. Mohanty, M. Harun AI Rashid, M. Mridul, C. Mohanty, and S. Swayamsiddha, "Application of Artificial Intelligence in COVID-19 drug repurposing," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 14, no. 5, pp. 1027–1031, Sep.-Oct. 2020, doi: 10.1016/j.dsx.2020.06.068.
- [57] N. Yuvaraj, K. Srihari, S. Chandragandhi, R. A. Raja, G. Dhiman, and A. Kaur, "Analysis of protein-ligand interactions of SARS-CoV-2 against selective drug using deep neural networks," *Big Data Mining and Analytics*, vol. 4, no. 2, pp. 76–83, 2021, doi: 10.26599/BDMA.2020.9020007.
- [58] T. Muthu Kumar, K. Rohini, N. James, V. Shanthi, and K. Ramanathan, "Discovery of potent Covid-19 main protease inhibitors using integrated drug-repurposing strategy," *Biotechnology and applied biochemistry*, vol. 68, no. 4, pp. 712–725, 2021, doi: 10.1002/bab.2159.
- [59] S. Mohapatra *et al.*, "Repurposing therapeutics for COVID-19: Rapid prediction of commercially available drugs through machine learning and docking," *PLoS One*, vol. 15, no. 11, p. e0241543, 2020, doi: 10.1371/journal.pone.0241543.
- [60] E. Ong, M. U. Wong, A. Huffman, and Y. He, "COVID-19 Coronavirus Vaccine Design Using Reverse Vaccinology and Machine Learning," *Frontiers in immunology*, vol. 11, p. 1581, 2020, doi: 10.3389/fimmu.2020.01581.
- [61] J. Chen, K. Gao, R. Wang, and G. W. Wei, "Prediction and mitigation of mutation threats to COVID-19 vaccines and antibody therapies," *Chemical science*, vol. 12, no. 20, pp. 6929–6948, 2021, doi: 10.1039/d1sc01203g.
- [62] L. A. Cofas, C. Delcea, I. Roxin, C. Ioanăș, D. S. Gherai, and F. Tajariol, "The Longest Month: Analyzing COVID-19 Vaccination Opinions Dynamics from Tweets in the Month following the First Vaccine Announcement," *IEEE Access*, vol. 9, pp. 33203–33223, 2021, doi: 10.1109/ACCESS.2021.3059821.
- [63] T. T. Nguyen *et al.*, "Genomic mutations and changes in protein secondary structure and solvent accessibility of SARS-CoV-2 (COVID-19 virus)," *Scientific Reports*, vol. 11, no. 1, p. 3487, 2021, doi: 10.1038/s41598-021-83105-3.
- [64] G. S. Randhawa, M. P. M. Soltysiak, H. El Roz, C. P. E. de Souza, K. A. Hill, and L. Kari, "Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study," *PLoS One*, vol. 15, no. 4, p. e0232391, 2020, doi: 10.1371/journal.pone.0232391.
- [65] T. B. Alakus and I. Turkoglu, "A Novel Protein Mapping Method for Predicting the Protein Interactions in COVID-19 Disease by Deep Learning," *Interdisciplinary Sciences: Computational Life Sciences*, vol. 13, no. 1, pp. 44–60, 2021, doi: 10.1007/s12539-020-00405-4.
- [66] H. Du, F. Chen, H. Liu, and P. Hong, "Network-based virus-host interaction prediction with application to SARS-CoV-2," *Patterns*, vol. 2, no. 5, p. 100242, May 2021, doi: 10.1016/j.patter.2021.100242.
- [67] J. Chen, R. Wang, M. Wang, and G. W. Wei, "Mutations Strengthened SARS-CoV-2 Infectivity," *Journal of molecular biology*, vol. 432, no. 19, pp. 5212–5226, Sep. 2020, doi: 10.1016/j.jmb.2020.07.009.
- [68] S. Esmail and W. R. Danter, "Lung organoid simulations for modelling and predicting the effect of mutations on SARS-CoV-2 infectivity," *Computational and Structural Biotechnology Journal*, vol. 19, pp. 1701–1712, 2021, doi: 10.1016/j.csbj.2021.03.020.
- [69] Z. Pang, G. Zhou, J. Chong, and J. Xia, "Comprehensive meta-analysis of covid-19 global metabolomics datasets," *Metabolites*, vol. 11, no. 1, p. 44, 2021, doi: 10.3390/metabo11010044.
- [70] D. L. Ng *et al.*, "A diagnostic host response biosignature for COVID-19 from RNA profiling of nasal swabs and blood," *Science advances*, vol. 7, no. 6, p. eabe5984, 2021, doi: 10.1126/sciadv.abe5984.
- [71] T. Shu *et al.*, "Plasma Proteomics Identify Biomarkers and Pathogenesis of COVID-19," *Immunity*, vol. 53, no. 5, pp. 1108–1122.e5, Nov. 2020, doi: 10.1016/j.immuni.2020.10.008.
- [72] W. N. Chia *et al.*, "Dynamics of SARS-CoV-2 neutralising antibody responses and duration of immunity: a longitudinal study," *The Lancet Microbe*, vol. 2, no. 6, pp. e240–e249, Jun. 2021, doi: 10.1016/S2666-5247(21)00025-2.
- [73] P. N. Amin, S. S. Moghe, S. N. Prabhakar and C. M. Nehete, "Deep Learning Based Face Mask Detection and Crowd Counting," *2021 6th International Conference for Convergence in Technology (I2CT)*, 2021, pp. 1–5, doi: 10.1109/I2CT51068.2021.9417826.
- [74] S. Hussain *et al.*, "IoT and deep learning based approach for rapid screening and face mask detection for infection spread control of COVID-19," *Applied Sciences*, vol. 11, no. 8, p. 3495, 2021, doi: 10.3390/app11083495.
- [75] X. Kong *et al.*, "Real-Time Mask Identification for COVID-19: An Edge-Computing-Based Deep Learning Framework," in *IEEE Internet of Things Journal*, vol. 8, no. 21, pp. 15929–15938, Nov. 2021, doi: 10.1109/IJOT.2021.3051844.
- [76] M. M. Rahman, M. M. H. Manik, M. M. Islam, S. Mahmud, and J. -H. Kim, "An Automated System to Limit COVID-19 Using Facial Mask Detection in Smart City Network," *2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, 2020, pp. 1–5, doi: 10.1109/IEMTRONICS51293.2020.9216386.
- [77] M. Al-Ramahi, A. Elnoshokaty, O. El-Gayar, T. Nasralah, and A. Wahbeh, "Public discourse against masks in the COVID-19 Era: Infodemiology study of twitter data," *JMIR Public Health and Surveillance*, vol. 7, no. 4, p. e26780, 2021, doi: 10.2196/26780.
- [78] C. Guo, J. Zhang, Y. Liu, Y. Xie, Z. Han, and J. Yu, "Recursion Enhanced Random Forest with an Improved Linear Model (RERF-ILM) for Heart Disease Detection on the Internet of Medical Things Platform," *IEEE Access*, vol. 8, pp. 59247–59256, 2020, doi: 10.1109/ACCESS.2020.2981159.
- [79] S. Angraal *et al.*, "Machine Learning Prediction of Mortality and Hospitalization in Heart Failure With Preserved Ejection






- Fraction," *JACC: Heart Failure*, vol. 8, no. 1, pp. 12–21, 2020, doi: 10.1016/j.jchf.2019.06.013.
- [80] P. M. Buscema, E. Grossi, G. Massini, M. Breda, and F. Della Torre, "Computer Aided Diagnosis for atrial fibrillation based on new artificial adaptive systems," *Computer Methods and Programs in Biomedicine*, vol. 191, p. 105401, Jul. 2020, doi: 10.1016/j.cmpb.2020.105401.
- [81] J. myoung Kwon *et al.*, "Deep-learning-based risk stratification for mortality of patients with acute myocardial infarction," *PLoS One*, vol. 14, no. 10, p. e0224502, 2019, doi: 10.1371/journal.pone.0224502.
- [82] J. Ramsingh and V. Bhuvanawari, "A Big Data Framework to Analyze Risk Factors of Diabetes Outbreak in Indian Population Using a Map Reduce Algorithm," *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2018, pp. 1755–1760, doi: 10.1109/ICCONS.2018.8663143.
- [83] M. Nilashi, O. B. Ibrahim, A. Mardani, A. Ahani, and A. Jusoh, "A soft computing approach for diabetes disease classification," *Health Informatics Journal*, vol. 24, no. 4, pp. 379–393, 2018, doi: 10.1177/1460458216675500.
- [84] E. Ullah, R. Mall, R. Rawi, N. M. Moustaid, A. A. Butt, and H. Bensmail, "Harnessing Qatar Biobank to understand type 2 diabetes and obesity in adult Qataris from the First Qatar Biobank Project," *Journal of translational medicine*, vol. 16, no. 1, pp. 1–10, 2018, doi: 10.1186/s12967-018-1472-0.
- [85] Y. Zhang *et al.*, "Artificial intelligence-enabled screening for diabetic retinopathy: A real-world, multicenter and prospective study," *BMJ Open Diabetes Research and Care*, vol. 8, no. 1, pp. 1–11, 2020, doi: 10.1136/bmjdr-2020-001596.
- [86] M. Du, D. G. Haag, J. W. Lynch, and M. N. Mittinty, "Comparison of the tree-based machine learning algorithms to cox regression in predicting the survival of oral and pharyngeal cancers: Analyses based on seer database," *Cancers*, vol. 12, no. 10, p. 2802, 2020, doi: 10.3390/cancers12102802.
- [87] K. Liu *et al.*, "Deep learning pathological microscopic features in endemic nasopharyngeal cancer: Prognostic value and potential role for individual induction chemotherapy," *Cancer medicine*, vol. 9, no. 4, pp. 1298–1306, 2020, doi: 10.1002/cam4.2802.
- [88] K. Syed *et al.*, "Machine-Learning Models for Multicenter Prostate Cancer Treatment Plans," *Journal of Computational Biology*, vol. 28, no. 2, pp. 166–184, 2021, doi: 10.1089/cmb.2020.0188.
- [89] J. H. Ahn *et al.*, "Development of a Novel Prognostic Model for Predicting Lymph Node Metastasis in Early Colorectal Cancer: Analysis Based on the Surveillance, Epidemiology, and End Results Database," *Frontiers in oncology*, vol. 11, p. 999, 2021, doi: 10.3389/fonc.2021.614398.
- [90] S. Lee *et al.*, "Machine learning on genome-wide association studies to predict the risk of radiation-associated contralateral breast cancer in the WECARE Study," *PLoS One*, vol. 15, no. 2, p. e0226157, 2020, doi: 10.1371/journal.pone.0226157.
- [91] A. A. Elfiky, M. J. Pany, R. B. Parikh, and Z. Obermeyer, "Development and Application of a Machine Learning Approach to Assess Short-term Mortality Risk Among Patients With Cancer Starting Chemotherapy," *JAMA network open*, vol. 1, no. 3, p. e180926, 2018, doi: 10.1001/jamanetworkopen.2018.0926.
- [92] N. Nazeer, B. Wajid, I. Nazir and F. Gohar, "Prediction of Malignancy of Brain Cancer on SEER Dataset using Random Forest, SVM, and Naive Bayes Classifiers," *2020 IEEE 23rd International Multitopic Conference (INMIC)*, 2020, pp. 1–5, doi: 10.1109/INMIC50486.2020.9318156.
- [93] D. Xu, J. Q. Sheng, P. J. H. Hu, T. S. Huang, and W. C. Lee, "Predicting hepatocellular carcinoma recurrences: A data-driven multiclass classification method incorporating latent variables," *Journal of Biomedical Informatics*, vol. 96, p. 103237, Aug. 2019, doi: 10.1016/j.jbi.2019.103237.
- [94] M. Nyirenda, R. Omori, H. L. Tessmer, H. Arimura, and K. Ito, "Estimating the lineage dynamics of human influenza B viruses," *PLoS One*, vol. 11, no. 11, p. e0166107, 2016, doi: 10.1371/journal.pone.0166107.
- [95] S. Venkatraman *et al.*, "Forecasting influenza activity using machine-learned mobility map," *Nature communications*, vol. 12, no. 1, pp. 1–12, 2021, doi: 10.1038/s41467-021-21018-5.
- [96] S. S. Ayachit, T. Kumar, S. Deshpande, N. Sharma, K. Chaurasia, and M. Dixit, "Predicting H1N1 and Seasonal Flu: Vaccine Cases using Ensemble Learning approach," *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, 2020, pp. 172–176, doi: 10.1109/ICACCCN51052.2020.9362909.
- [97] M. A. Zeller, P. C. Gauger, Z. W. Arendsee, C. K. Souza, A. L. Vincent, and T. K. Anderson, "Machine Learning Prediction and Experimental Validation of Antigenic Drift in H3 Influenza A Viruses in Swine," *mSphere*, vol. 6, no. 2, p. e00920-20, 2021, doi: 10.1128/msphere.00920-20.
- [98] C. L. P. Eng, J. C. Tong, and T. W. Tan, "Distinct host tropism protein signatures to identify possible zoonotic influenza A viruses," *PLoS One*, vol. 11, no. 2, p. e0150173, 2016, doi: 10.1371/journal.pone.0150173.
- [99] V. K. Jain and S. Kumar, "Rough set based intelligent approach for identification of H1N1 suspect using social media," *Kuwait Journal of Science*, vol. 45, no. 2, pp. 8–14, 2018.
- [100] P. Chen, E. Chen, L. Chen, X. J. Zhou, and R. Liu, "Detecting early-warning signals of influenza outbreak based on dynamic network marker," *Journal of cellular and molecular medicine*, vol. 23, no. 1, pp. 395–404, 2019, doi: 10.1111/jcmm.13943.
- [101] L. S. Jayashree, R. L. Devi, N. Papandrianos, and E. I. Papageorgiou, "Application of Fuzzy Cognitive Map for geospatial dengue outbreak risk prediction of tropical regions of Southern India," *Intelligent Decision Technologies*, vol. 12, no. 2, pp. 231–250, 2018, doi: 10.3233/IDT-180330.
- [102] P. Guo *et al.*, "Developing a dengue forecast model using machine learning: A case study in China," *PLoS neglected tropical diseases*, vol. 11, no. 10, p. e0005973, 2017, doi: 10.1371/journal.pntd.0005973.
- [103] F. Y. Nejad and K. D. Varathan, "Identification of significant climatic risk factors and machine learning models in dengue outbreak prediction," *BMC Medical Informatics and Decision Making*, vol. 21, no. 1, pp. 1–12, 2021, doi: 10.1186/s12911-021-01493-y.
- [104] K. G. S. Dharmawardana *et al.*, "Predictive model for the dengue incidences in Sri Lanka using mobile network big data," *2017 IEEE International Conference on Industrial and Information Systems (ICIIS)*, 2017, pp. 1–6, doi: 10.1109/ICIINFS.2017.8300381.
- [105] O. S. Zhao *et al.*, "Convolutional neural networks to automate the screening of malaria in low-resource countries," *PeerJ*, vol. 8, p. e9674, 2020, doi: 10.7717/peerj.9674.
- [106] A. Maqsood, M. S. Farid, M. H. Khan, and M. Grzegorzec, "Deep malaria parasite detection in thin blood smear microscopic images," *Applied Sciences*, vol. 11, no. 5, pp. 1–19, 2021, doi: 10.3390/app11052284.
- [107] C. Dobanõ *et al.*, "RTS,S/AS01E immunization increases antibody responses to vaccine-unrelated Plasmodium falciparum antigens associated with protection against clinical malaria in African children: A case-control study," *BMC medicine*, vol. 17, no. 1, pp. 1–19, 2019, doi: 10.1186/s12916-019-1378-6.
- [108] V. Kittichai *et al.*, "Deep learning approaches for challenging species and gender identification of mosquito vectors," *Scientific reports*, vol. 11, no. 1, pp. 1–14, 2021, doi: 10.1038/s41598-021-84219-4.
- [109] C. Mwangi, S. Karanja, J. Gachohi, V. Wanjihia, and Z. Ngang'A, "Depression, injecting drug use, and risky sexual behavior syndemic among women who inject drugs in Kenya: A cross-sectional survey," *Harm reduction journal*, vol. 16, no. 1, pp. 1–11, 2019, doi: 10.1186/s12954-019-0307-5.




- [110] R. Mojtabai, M. Amin-Esmacili, E. Nejat, and M. Olfson, "Misuse of prescribed opioids in the United States," *Pharmacoepidemiology and drug safety*, vol. 28, no. 3, pp. 345–353, 2019, doi: 10.1002/pds.4743.
- [111] N. Poorinmohammad, H. Mohabatkar, M. Behbahani, and D. Biria, "Computational prediction of anti HIV-1 peptides and in vitro evaluation of anti HIV-1 activity of HIV-1 P24-derived peptides," *Journal of peptide science*, vol. 21, no. 1, pp. 10–16, 2015, doi: 10.1002/psc.2712.
- [112] G. Wang *et al.*, "Application of a long short-term memory neural network: A burgeoning method of deep learning in forecasting HIV incidence in Guangxi, China," *Epidemiology & Infection*, vol. 147, 2019, doi: 10.1017/S095026881900075X.
- [113] X. Dong *et al.*, "Predicting opioid overdose risk of patients with opioid prescriptions using electronic health records based on temporal deep learning," *Journal of Biomedical Informatics*, vol. 116, p. 103725, Apr. 2021, doi: 10.1016/j.jbi.2021.103725.
- [114] H. Yao, S. Rashidian, X. Dong, H. Duanmu, R. N. Rosenthal, and F. Wang, "Detection of suicidality among opioid users on reddit: Machine learning-based approach," *Journal of medical internet research*, vol. 22, no. 11, p. e15293, 2020, doi: 10.2196/15293.
- [115] L. B. Balzer *et al.*, "Machine Learning to Identify Persons at High-Risk of Human Immunodeficiency Virus Acquisition in Rural Kenya and Uganda," *Clinical Infectious Diseases*, vol. 71, no. 9, pp. 2326–2333, 2020, doi: 10.1093/cid/ciz1096.
- [116] N. Chintalapudi, G. Battineni, and F. Amenta, "Sentimental analysis of COVID-19 tweets using deep learning models," *Infectious Disease Reports*, vol. 13, no. 2, pp. 329–339, 2021, doi: 10.3390/IDR13020032.
- [117] S. Albahli *et al.*, "COVID-19 Public Sentiment Insights: A Text Mining Approach to the Gulf Countries," *Cmc-Computers Materials & Continua*, vol. 67, no. 2, pp. 1613–1627, 2021, doi: 10.32604/cmc.2021.014265.
- [118] D. A. Bowen, J. Wang, K. Holland, B. Bartholow, and S. A. Sumner, "Conversational topics of social media messages associated with state-level mental distress rates," *Journal of mental health*, vol. 29, no. 2, pp. 234–241, 2020, doi: 10.1080/09638237.2020.1739251.
- [119] X. Zhou, E. Coiera, G. Tsafnat, D. Arachi, M. S. Ong, and A. G. Dunn, "Using social connection information to improve opinion mining: Identifying negative sentiment about HPV vaccines on Twitter," *Stud. Health Technol. Inform.*, vol. 216, pp. 761–765, 2015, doi: 10.3233/978-1-61499-564-7-761.
- [120] X. Huang *et al.*, "Can online self-reports assist in real-time identification of influenza vaccination uptake? A cross-sectional study of influenza vaccine-related tweets in the USA, 2013–2017," *BMJ Open*, vol. 9, no. 1, p. e024018, 2019, doi: 10.1136/bmjopen-2018-024018.
- [121] S. D. Young, W. Yu, and W. Wang, "Toward automating HIV identification: Machine learning for rapid identification of HIV-related social media data," *Journal of acquired immune deficiency syndromes*, vol. 74, no. 3, pp. S128–S131, 2017, doi: 10.1097/QAI.0000000000001240.
- [122] M. K. Elhadad, K. F. Li, and F. Gebali, "Detecting misleading information on COVID-19," *IEEE Access*, vol. 8, pp. 165201–165215, 2020, doi: 10.1109/ACCESS.2020.3022867.
- [123] A. Kumar and N. Sachdeva, "Multimodal cyberbullying detection using capsule network with dynamic routing and deep convolutional neural network," *Multimedia Systems*, pp. 1–10, Feb. 2021, doi: 10.1007/s00530-020-00747-5.
- [124] S. Amin, M. I. Uddin, D. H. Alsaeed, A. Khan, and M. Adnan, "Early Detection of Seasonal Outbreaks from Twitter Data Using Machine Learning Approaches," *Complexity*, vol. 2021, 2021, doi: 10.1155/2021/5520366.
- [125] E. Nugraheni, P. H. Khotimah, A. Arisal, A. F. Rozie, D. Riswantini, and A. Purwarianti, "Classifying aggravation status of COVID-19 event from short-text using CNN," *2020 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET)*, 2020, pp. 240–245, doi: 10.1109/ICRAMET51080.2020.9298674.
- [126] T. Katsuki, T. K. Mackey, and R. Cuomo, "Establishing a link between prescription drug abuse and illicit online pharmacies: Analysis of twitter data," *Journal of medical Internet research*, vol. 17, no. 12, pp. 1–12, 2015, doi: 10.2196/jmir.5144.

## BIOGRAPHIES OF AUTHORS



**Dianadewi Riswantini**    received a scholarship from the Overseas Fellowship Program, a collaboration between the Indonesian Government and World Bank, for a Bachelor's and Master's degree in computer science from the Delft University of Technology, the Netherlands, completed in 1994. She is currently a Ph.D. candidate in the School of Business and Management, Bandung Institute Technology, Indonesia, engaged in Big Data analytics for business and management. She is joining the Information Retrieval Research Group at the Research Center for Data and Information Sciences, National Agency of Research and Innovation (Indonesia). Her research interests include data analytics, text mining, natural language processing, and machine learning in the fields of social and medical informatics. She can be contacted at the email dianadewi.riswantini@brin.go.id.



**Ekasari Nugraheni**    obtained a Bachelor's degree in information management from the AKRIND Institute of Technology Yogyakarta in 1996. She received a scholarship from the Indonesian Ministry of Research and Technology for her Master's degree in informatics from the Bandung Institute of Technology, Indonesia, completed in 2016. Currently, she is joining the Information Retrieval Research Group at the Research Center for Data and Information Sciences, National Research and Innovation Agency (Indonesia). Her research interests include data analysis, data mining, deep learning, and natural language processing. She can be contacted at email: ekasari.nugraheni@brin.go.id.