COVID-19 DATASETS: A SURVEY AND FUTURE CHALLENGES

Junaid Shuja*
Department of Computer Science
COMSATS University Islamabad
Abbottabad Campus, Pakistan
And
Umm Al-Qura University
Makkah, Saudi Arabia

Eisa Alanazi
Computer Science Department
Umm Al-Qura University
Makkah, Saudi Arabia

Waleed Alasmary
Computer Engineering Department
Umm Al-Qura University
Makkah, Saudi Arabia

Abdulaziz Alashaikh
Computer Engineering & Networks Department
University of Jeddah
Jeddah, Saudi Arabia

May 20, 2020

ABSTRACT

In December 2019, a novel virus named as COVID-19 emerged in the city of Wuhan, China. In early 2020, the COVID-19 virus spread in all continents of the world except Antarctica causing widespread infections and deaths due to its contagious characteristics and no medically proven treatment. The COVID-19 pandemic has been termed as most consequential global crisis after the World Wars. The first line of defense against the COVID-19 spread are the non-pharmaceutical measures like social distancing and personal hygiene. On the other hand, the medical service providers are the first responders for infected persons with severe symptoms of COVID-19. The great pandemic affecting billions of lives economically and socially has motivated the scientific community to come up with solutions based on computer-aided digital technologies for diagnosis, prevention, and estimation of COVID-19. Some of these efforts focus on statistical and Artificial Intelligence-based analysis of the available data concerning COVID-19. All of these scientific efforts necessitate that the data brought to service for the analysis should be open-source to promote the extension, validation, and collaboration of the work in the fight against the global pandemic. Our survey is motivated by the open-source efforts that can be mainly categorized as: **(a)** COVID-19 diagnosis from CT scans and X-ray images, **(b)** COVID-19 case reporting, transmission estimation, and prognosis from epidemiological, demographic, and mobility data, **(c)** COVID-19 emotional and sentiment analysis from social media, and **(d)** knowledge-based discovery and semantic analysis from the collection of scholarly articles covering COVID-19. We review and critically analyze works in these directions that are accompanied by open-source data and code. We hope that the article will provide the scientific community with an initiative to start open-source extensible and transparent research in the collective fight against COVID-19.

Keywords COVID-19, coronavirus, pandemic, machine learning, artificial intelligence, open-source, data-sets.

1 Introduction

The COVID-19 virus has been declared a pandemic by the World Health Organization (WHO) with more than three million cases and 224172 deaths across the world as per WHO statistics of 1 May 2020 [1]. The cure to COVID-19 can take twelve months due to its clinical trials on humans of varying ages and ethnicity before approval. The cure to

*Corresponding author: junaidshuja@cuiatd.edu.pk

COVID-19 can be further delayed due to possible genetic mutations shown by the virus [2]. The pandemic situation is affecting billions of people socially, economically, and medically with drastic changes in social relationships, health policies, trade, work and educational environments. The global pandemic is a threat to human society and calls for immediate actions. The COVID-19 pandemic has motivated the research community to aid front-line medical service staff with cutting edge research for mitigation, detection, and prevention of the virus [3].

Scientific community has brainstormed to come up with ideas that can limit the crisis and help prevent future such pandemics. Other than medical science researchers and virology specialists, scientists supported with digital technologies have tackled the pandemic with novel methods. Two significant scientific communities aided with digital technologies can be identified in the fight against COVID-19. The main digital effort in this regard comes from the Artificial Intelligence (AI) community in the form of automated COVID-19 detection from Computed Tomography (CT) scans and X-ray images. The second such community aided by digital technologies is of mathematicians and epidemiologists who are developing complex virus diffusion and transmission models to estimate virus spread under various mobility and social distancing scenarios [4]. Besides these two major scientific communities, efforts are being made for analyzing social and emotional behavior from social media [5], collecting scholarly articles for semantic analysis [6], detection COVID-19 from cough samples [7], and automated contact tracing [2].

Data is an essential element for the efficient implementation of scientific methods. Two approaches are followed by the community while performing scientific research. The research methods and data are either closed-source to protect proprietary scientific contributions or open-source for higher usability, verifiability, transparency, and quality [8, 9]. The latter approach is also known as public or open-access approach. In the existing COVID-19 pandemic, the open-source approach is deemed more effective for mitigation and detection of the COVID-19 virus. We emphasize that the COVID-19 pandemic demands a unified approach with open-source data and methodology so that the scientific community across the globe can join hands with verifiable and transparent research [10].

The combination of AI and open-source data sets produces a practical solution for COVID-19 diagnosis. Automated CT scan based COVID-19 detection techniques work with training the learning model on existing CT scan data-sets that contain labeled images of COVID-19 positive and normal cases. Similarly, the detection of COVID-19 from cough samples requires a data-set of both normal and infected persons to learn and distinguish features of the infected person from a healthy person. Therefore, it is necessary to provide open-source data-sets and methods so that **(a)** researchers across globe can enhance and modify existing work to limit the global pandemic, **(b)** existing techniques are verified for correctness by researchers across the board before implementation in real-world scenarios, and **(c)** researchers collaborate and join hands in the global fight against the pandemic with community-oriented research and development [11, 12].

Artificial Intelligence (AI) and Machine Learning (ML) techniques have been prominently used to efficiently solve various computer science problems ranging from bio-informatics to image processing. ML is based on the premise that an intelligent machine should be able to learn and adapt from its environment based on its experiences without explicit programming [13, 14]. ML models and algorithms have been standardized across multiple programming languages such as, Python and R. The main challenge to the application of ML models is the availability of the open-source data [15, 16]. Given publicly available data sets, ML techniques can aid the fight against COVID-19 on multiple fronts. The principal such application is ML based COVID-19 diagnosis from CT scans and X-rays that can lower the burden on short supplies of reverse transcriptase polymerase chain reaction (RT-PCR) test kits [17, 18]. Similarly, statistical and epidemiological analysis of COVID-19 case reports can help find a relation between human mobility and virus transmission. Moreover, social media data mining can provide sentiment and socio-economic analysis in current pandemic for policy makers. Therefore, the COVID-19 pandemic has necessitated collection of new data sets regarding human mobility, epidemiology, psychology, and radiology to aid scientific efforts [19]. It must be noted that while digital technologies are aiding in the combat against COVID-19, they are also being utilized for spread of misinformation [20], hatred [21], propaganda [22], and online financial scams [23].

We make the first effort in surveying research works based on open-source data sets concerning COVID-19 pandemic. Researchers [18] surveyed AI-based techniques for data acquisition, segmentation, and diagnosis for COVID-19. The article was not focused on works that are accompanied by publicly available data sets. Moreover, the above-mentioned article focused only on the applications of ML based medical diagnosis. Authors [19] listed publicly available medical data sets for COVID-19. The work did not detail the AI applications of the data-set and data-sets on epidemiology and psychology of COVID-19. In contrast, we provide a comprehensive survey of the open-source COVID-19 data-sets (medical and textual) with details of the applied AI and statistical techniques. We are motivated by the fact that this survey will help researchers in the identification of appropriate open-source data-sets for their research. We provide a comparison of data-sets in terms of application, type, and size with similar items will provide valuable insights for data-set selection. Moreover, we highlight the future research directions in terms of missed research opportunities and missing data-sets so that the research community can work towards the public availability of the data.

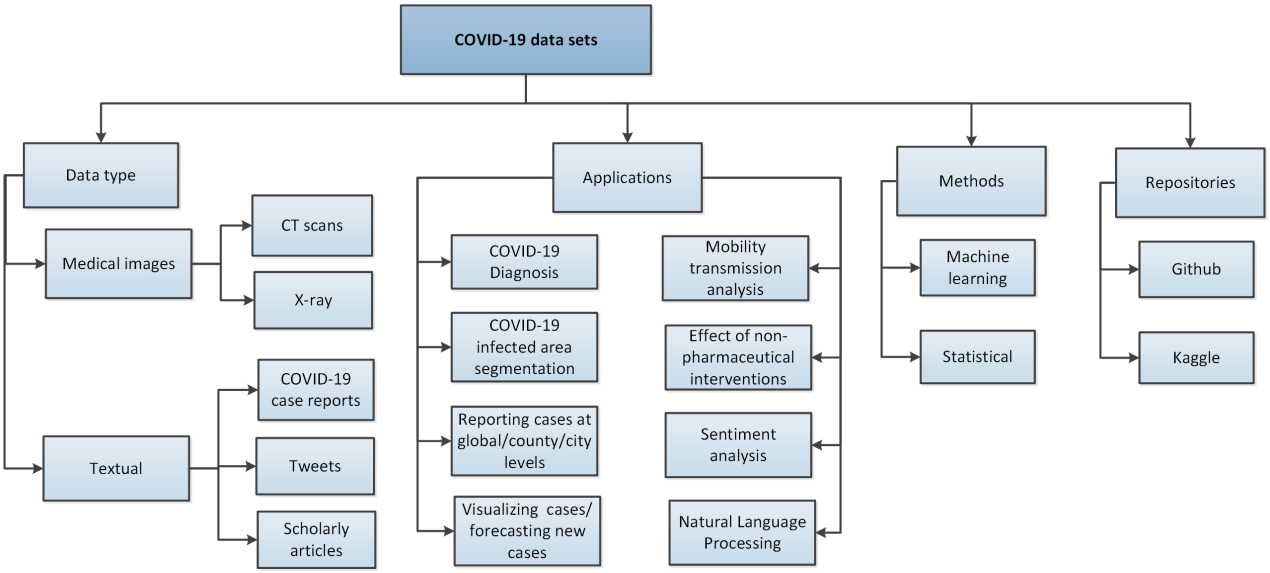

Figure 1: Taxonomy of COVID-19 open-source data sets

Most of the articles included in this survey have not been rigorously peer-reviewed and published as pre-prints. However, their inclusion is necessary as the current pandemic situation requires rapid publishing process. Moreover, the inclusion of non-peer-reviewed studies in this article is supported by their transparent open-source methods which can be independently verified. For the search of the relevant literature review, we searched the online databases of Google scholar, BioRxiv, and medRxiv. The keywords employed were "COVID-19" and "data-set". We separately searched two online open-source communities, i.e., Kaggle and Github for data-sets that are not yet part of any publication. We focused on articles with applications of computer science and mathematics in general. We hope that our efforts will be fruitful in limiting the spread of COVID-19 through elaboration of scientific fact-finding efforts.

We divide the data sets into two main categories, i.e., **(a)** medical images and **(b)** textual data. Medical images based data sets are mostly brought into service for screening and diagnosis of COVID-19. Medical images are either Chest CT scan or X-rays. Medical image data sets should consider patients consent and preserve patient privacy. Medical image based diagnosis lowers burden on conventional PCR based screening. Textual data sets serve three main purposes, i.e., **(a)** forecasting the transmission and spread of COVID-19 based on reported cases, **(b)** analyzing public sentiment/opinion by tracking/mining COVID-19 related keywords on popular social media platforms, and **(c)** collecting scholarly articles on COVID-19 for a centralized view on related research and application of information extraction/text mining. A consolidated view of the taxonomy of COVID-19 Open Source data sets is illustrated in Figure 1.

The rest of the article is organized as follows. Section 2 presents the comprehensive list of medical COVID-19 data sets divided into categories of CT scans and X-rays. Section 3 details a list of textual data sets classified into COVID-19 case report, social media, and scholarly article collections. In section 4, a comparison of listed data-sets is provided in terms of openness, application, and data-type. Section 5 discusses the dimensions that need attention from scholars and future perspectives on COVID-19 research. Section 6 provides the concluding remarks for the article.

2 COVID-19 Medical image data-sets

Medical images in the form of Chest CT scans and X-rays are essential for automated COVID-19 diagnosis. Some of the leading hospitals across the world are utilizing AI/ML algorithms to diagnose COVID-19 cases from CT scans/X-ray images after preliminary trails of the technology [2]. A generic work-flow of ML based COVID-19 diagnosis is illustrated in Figure 2. We discuss CT scan and X-ray image data-sets separately in the following subsections.

2.1 CT scans

Cohen et al. [11] describe the public COVID-19 image collection consisting of X-ray and CT-scans with ongoing updates. The data set consists of more than 125 images extracted from various online publications and websites. The

[2]https://www.bbc.com/news/business-52483082

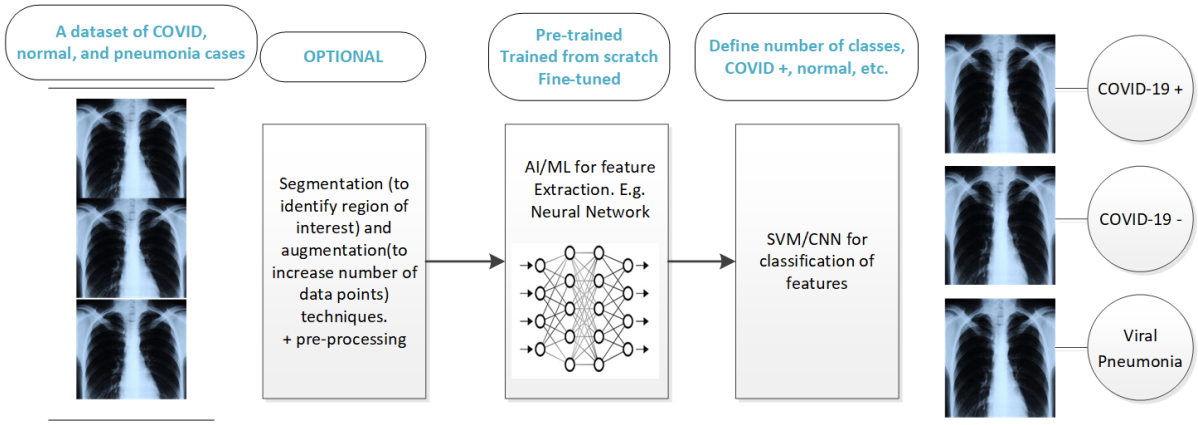

Figure 2: A generic work-flow of AI/ML based COVID-19 diagnosis

data set specifically includes images of COVID-19 cases along with MERS, SARS, and ARDS based images. The authors enlist the application of deep and transfer learning on their extracted data set for identification of COVID-19 while utilizing motivation from earlier studies that learned the type of pneumonia from similar images [24]. Each image is accompanied by a set of 16 attributes such as patient ID, age, date, and location. The extraction of CT scan images from published articles rather than actual sources may lessen the image quality and affect the performance of the machine learning model. Some of the data sets available and listed below are obtained from secondary sources. The public dataset published with this study is one of the pioneer efforts in COVID-19 detection and most of the listed studies utilize this data set.

Researcher [25] published a data set consisting of 275 CT scans of COVID-19 positive patients. The data set is extracted from 760 medRxiv and bioRxiv preprints about COVID-19. The authors also employed a deep convolutional network for training on the data set to learn COVID-19 cases for new data with an accuracy of around 85%. The model is trained on 183 COVID-19 positive CT scans and 146 negative cases. The model is tested on 35 COVID-19 positive CTs and 34 non-COVID CTs and achieves an F1 score of 0.85. Due to the small data set size, deep learning models tend to overfit. Therefore, the authors utilized transfer learning on the Chest X-ray data set released by NIH to fine-tune their deep learning model. The online repository is being regularly updated and currently consists of 349 CT images containing clinical findings of 216 patients.

Wang et al. [26] investigated a deep learning strategy for COVID-19 screening from CT scans. A total of 453 COVID-19 pathogen-confirmed CT scans were utilized along with typical viral pneumonia cases. The COVID-19 CT scans were obtained from various Chinese hospitals with an online repository maintained at [3]. Transfer learning in the form of pre-trained CNN model (M-inception) was utilized for feature extraction. A combination of Decision tree and Adaboost were employed for classification with 83.9% accuracy.

Segmentation helps health service providers to quickly and objectively evaluate the radiological images. Segmentation is a pre-processing step that outlines the region of interest, e.g., infected regions or lesions for further evaluation. Shan et al. [27] obtained CT scan images from COVID-19 cases based mostly in Shanghai for deep learning-based lung infection quantification and segmentation. However, their data set is not public. The deep learning-based segmentation utilizes VB-Net, a modified 3-D convolutional neural network, to segment COVID-19 infection regions in CT scans. The proposed system performs auto-contouring of infection regions, accurately estimates their shapes, volumes, and percentage of infection (POI). The system is trained using 249 COVID-19 patients and validated using 300 new COVID-19 patients. Radiologists contributed as a human in the loop to iteratively add segmented images to the training data set. Two radiologists contoured the infectious regions to quantitatively evaluate the accuracy of the segmentation technique. The proposed system and manual segmentation resulted in 90% dice similarity coefficients.

Other than the published articles, few online efforts have been made for image segmentation of COVID-19 cases. A COVID-19 CT Lung and Infection Segmentation Dataset is listed as open source [28]. The data set consists of 20 COVID-19 CT scans labeled into left, right, and infectious regions by two experienced radiologists and verified by another radiologist [4]. Three segmentation benchmark tasks have also been created based on the data set [5].

[3] `https://ai.nscc-tj.cn/thai/deploy/public/pneumonia_ct`
[4] https://zenodo.org/record/3757476
[5] https://gitee.com/junma11/COVID-19-CT-Seg-Benchmark

Another such online initiative is the COVID-19 CT segmentation dataset [6]. The segmented data set is hosted by two radiologists based in Oslo. They obtained images from a repository hosted by the Italian society of medical and interventional radiology (SIRM) [7]. The obtained images were segmented by the radiologist using 3 labels, i.e., ground-glass, consolidation, and pleural effusion. As a result, a data set that contains 100 axial CT slices from 60 patients with manual segmentations in the form of JPG images is formed. Moreover, the radiologists also trained a 2D multilabel U-Net model for automated semantic segmentation of images.

In the following paragraph, we list COVID-19 data set initiatives that are public but are not associated with any publication.

The Coronacases Initiative shares 3D CT scans of confirmed cases of COVID-19 [8]. Currently, the web repository contains 3D CT images of 10 confirmed COVID-19 cases shared for scientific purposes. The British Society of Thoracic Imaging (BSTI) in collaboration with Cimar UK's Imaging Cloud Technology deployed a free to use encrypted and anonymized online portal to upload and download medical images related to COVID-19 positive and suspected patients [9]. The uploaded images are sent to a group of BSTI experts for diagnosis. Each reported case includes data regarding the age, sex, PCR status, and indications of the patient. The aim of the online repository is to provide COVID-19 medical images for reference and teaching. The SIRM is hosting radiographical images of COVID-19 cases [10]. Their data set has been utilized by some of the cited works in this article. Another open-source repository for COVID-19 radiographic images is Radiopaedia [11]. Multiple studies [29, 30] employed this dataset for their research.

2.2 X-ray

Researchers [31] present COVID-Net, a deep convolutional network for COVID-19 diagnosis based on Chest X-ray images. Motivated by earlier efforts on radiography based diagnosis of COVID-19, the authors make their data set and code accessible for further extension. The data set consists of 13,800 chest radiography images from 13,725 patient cases from three open access data repositories. The COVID-Net architecture consists of two stages. In the first stage, residual architecture design principles are employed to construct a prototype neural network architecture to predict either of **(a)** normal, **(b)** non-COVID infection, and **(c)** COVID-19 infections. In the second stage, the initial network design prototype, data, and human-specific design requirements act as a guide to a design exploration strategy to learn the parameters of deep neural network architecture. The authors also audit the COVID-Net with the aim of transparency via examination of critical factors leveraged by COVID-Net in making detection decisions. The audit is executed with GSInquire which is a commonly used AI/ML explainability method [32].

Author in [33] utilized the data set of Cohen et al. [11] and proposed COVIDX-Net, a deep learning framework for automatic COVID-19 detection from X-ray images. Severn different deep convolutional neural network architecture, namely, VGG19, DenseNet201, InceptionV3, ResNetV2, InceptionResNetV2, Xception, and MobileNetV2, were utilized for performance evaluation. The VGG19 and DenseNet201 model outperform other deep neural classifiers in terms of accuracy. However, these classifiers also demonstrate higher training times.

Apostolopoulos et al [34] merged the data set of Cohen et al. [11], a data set from Kaggle [12], and a data set of common bacterial-pneumonia X-ray scans [35] to train convolutional neural networks (CNN) to distinguish COVID-19 from common pneumonia. Five CNNs namely VGG19, MobileNet v2, Inception, Xception, and Inception ResNet v2 with common hyper-parameters. Results demonstrate that VGG19 and MobileNet v2 perform better than other CNNs in terms of accuracy, sensitivity, and specificity.

The researchers extended their work in [30] to extract biomarkers from X-ray images using a deep learning approach. The authors employ MobileNetv2, a CNN is trained for the classification task for six most common pulmonary diseases. MobileNetv2 extracts features from X-ray images in three different settings, i.e., from scratch, with the help of transfer learning (pre-trained), and hybrid feature extraction via fine-tuning. A layer of Global Average Pooling was added over MobileNetv2 to reduce overfitting. The extracted features are input to a 2500 node neural network for classification. The data set include recent COVID-19 cases and X-rays corresponding to common pulmonary diseases. The COVID-19 images (455) are obtained from Cohen et al. [11], SIRM, RSNA, and Radiopaedia. The data set of common pulmonary diseases is extracted from a recent study [35] among other sources. The training from scratch strategy outperforms

[6] http://medicalsegmentation.com/covid19/

[7] https://www.sirm.org/en/category/articles/covid-19-database/

[8] https://coronacases.org

[9] https://www.bsti.org.uk/training-and-education/covid-19-bsti-imaging-database/

[10] https://www.sirm.org/en/category/articles/covid-19-database/

[11] https://radiopaedia.org/articles/covid-19-3

[12] https://www.kaggle.com/andrewmvd/convid19-X-rays

transfer learning with higher accuracy and sensitivity. The aim of the research is to limit exposure of medical experts with infected patients with automated COVID-19 diagnosis.

Researchers [36] merged the data set of Cohen et al. [11] (50 images)and a data set from Kaggle [13] (50 images) for application of three pre-trained CNNs, namely, ResNet50, InceptionV3 and InceptionResNetV2 to detect COVID-19 cases from X-ray radiographs. The dataset was equally divided into 50 normal and 50 COVID-19 positive cases. Due to the limited data set, deep transfer learning is applied that requires smaller data set to learn and classify features. The ResNet50 provided the highest accuracy for classifying COVID-19 cases among the evaluated models.

Authors in [37] propose Support Vector Machine based classification of X-ray images instead of predominately employed deep learning models. The authors argue that deep learning models require large data sets for training that are not available currently for COVID-19 cases. The data set brought to service in this article is an amalgam of Cohen et al. [24], a data set of Kaggle [14], and data set of Kermany et. [35]. Author of [34] also utilized data from same sources. The dataset consists of 127 COVID-19 cases, 127 pneumonia cases, and 127 healthy cases. The methodology classifies the X-ray images into COVID-19, pneumonia, and normal cases. Pre-trained networks such as AlexNet, VGG16, VGG19, GoogleNet, ResNet18, ResNet50, ResNet101, InceptionV3, InceptionResNetV2, DenseNet201, XceptionNet, MobileNetV2 and ShuffleNet are employed on this dataset for deep feature extraction. The deep features obtained from these networks are fed to the SVM classifier. The accuracy and sensitivity of ResNet50 plus SVM is found to be highest among CNN models.

Similar to Sethy et al. [37], Afshar et al. [38] also negated the applicability of DNNs on small COVID-19 data sets. The authors proposed a capsule network model (COVID-CAPS) for the identification of COVID-19 based on X-ray images. Each layer of a Capsule Network consists of several Capsules, each of which represents a specific image instance at a specific location with the help of several neurons. The length of a Capsule determines the existence probability of the associated instance. COVID-CAPS uses four convolutional and three capsule layers and was pre-trained with transfer learning on the public NIH dataset of X-rays images for common thorax diseases. COVID-CAPS provides a binary output of either positive or negative COVID-19 case. The COVID-CAPS achieved an accuracy of 95.7%, a sensitivity of 90%, and specificity of 95.8%.

Authors [39] contributed towards a single COVID-19 X-ray image database for AI applications based on four sources. The aim of the research was to explore the possibility of AI application for COVID-19 diagnosis. The source databases were Cohen et al. [11], Italian Society of Medical and Interventional Radiology dataset, images from recently published articles, and a data set hosted at Kaggle [15]. The cumulative data set contains 190 COVID-19 images, 1345 viral pneumonia images, and 1341 normal chest x-ray images. The authors further created 2500 augmented images from each category for the training and validation of four CNNs. The four tested CNNs are AlexNet, ResNet18, DenseNet201, and SqueezeNet for classification of Xray images into normal, COVID-19, and viral pneumonia cases. The SqueezeNet outperformed other CNNs with 98.3% accuracy and 96.7% sensitivity. The collective database can be found at [16].

Authors [40] utilized data augmentation techniques to increase the number of data points for CNN based classification of COVID-19 X-ray images. The proposed methodology adds data augmentation to basic steps of feature extraction and classification. The authors utilize the data set of Cohen et al. [11]. The authors design five deep learning model for feature extraction and classification, namely, custom-made CNNs trained from scratch, transfer learning-based fine-tuned CNNs, proposed novel COVID-RENet, dynamic feature extraction through CNN and classification using SVM, and concatenation of dynamic feature spaces (COVID-RENet and VGG-16 features) and classification using SVM. SVM classification is brought to serve to further increase the accuracy of the task. The results showed that the proposed COVID-RENet and Custom VGG-16 models accompanied by the SVM classifier show better performance with approximately 98.3% accuracy in identifying COVID-19 cases.

A comprehensive list of AI-based COVID-19 research can be found at [41]. A list open-source data sets on the Kaggle can be found at [17]. A simple search list of COVID-19 based works on Github can be found at [18].

3 COVID-19 Textual data-sets

COVID-19 case reports, global and county-level dashboards, epidemiological data, demographic data, mobility data, social media posts, and scholarly article collections are detailed in the following subsections.

[13] https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia
[14] Kaggle. https ://www.Kaggle e.com/andre wmvd/convid19-X-rays
[15] https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia
[16] https://www.kaggle.com/tawsifurrahman/covid19-radiography-database
[17] https://www.kaggle.com/covid-19-contributions
[18] https://github.com/search?q=covid-19

3.1 COVID-19 case reports

The earliest and most noteworthy data set depicting the COVID-19 pandemic at a global scale was contributed by John Hopkins University [42]. The authors developed an online real-time interactive dashboard first made public in January 2020 [19]. The dashboard lists the number of cases, deaths, and recoveries divided into country/provincial regions. A data is more detailed to the city level for the USA, Canada, and Australia. A corresponding Github repository of the data is also available [20]. The data collection is semi-automated with main sources are DXY (a medical community [21]) and WHO. The DXY community collects data from multiple sources and updated every 15 minutes. The data is regularly validated from multiple online sources and health departments. The aim of the dashboard was to provide the public, health authorities, and researchers with a user-friendly tool to track, analyze, and model the spread of COVID-19.

Kucharski et al. [4] modeled COVID-19 cases based on datasets from and outside Wuhan. The purpose of the study was to estimate human-to-human transmissions and virus outbreaks if the virus was introduced in a new region. The four time-series datasets used were: the daily number of new internationally exported cases, the daily number of new cases in Wuhan with no market exposure, the daily number of new cases in China, and the proportion of infected passengers on evacuation flights between December 2019 and February 2020. The study while employing stochastic modeling found that the R*0* declined from 2.35 to 1.05 after travel restrictions were imposed in Wuhan. The study also found that if four cases are reported in a new area, there is a 50% chance that the virus will establish within the community.

Author in [43] presented a framework for serial interval estimation of COVID-19. As the virus is easily transmitted in a community from an infected person, it is important to know the onset of illness in primary to secondary transmissions. The date of illness onset is defined as the date on which a symptom relevant to COVID-19 infection appears. The serial interval refers to the time between successive cases in a chain of disease transmission. The authors obtain 28 cases of pairs of infector-infectee cases published in research articles and investigation reports and rank them for credibility. A subset of 18 high credible cases are selected to analyze that the estimated median serial interval lies at 4.0 days. The median serial interval of COVID-19 is found to be smaller than SARS. Moreover, it is implied that contact tracing methods may not be effective due to the rapid serial interval of infector-infectee transmissions.

Tindale et al. [44] study the COVID-19 outbreak to estimate the incubation period and serial interval distribution based on data obtained in Singapore (93 cases) and Tianjin (135). The incubation period is the period between exposure to an infection and the appearance of the first symptoms. The data was made available to the respective health departments. The serial interval can be used to estimate the reproduction number (R*0*) of the virus. Moreover, both serial interval and incubation period can help identify the extent of pre-symptomatic transmissions. With more than a months data of COVID-19 cases from both cities, The mean serial interval was found to be 4.56 days for Singapore and 4.22 days for Tianjin. The mean incubation period was found to be 7.1 days for Singapore and 9 days for Tianjin.

Researchers [45] described an econometric model to forecast the spread and prevalence of COVID-19. The analysis is aimed to aid public health authorities to make provisions ahead of time-based on the forecast. A time-series database was built based on statistics from Johns Hopkins University dashborad [22] and made public. Auto-Regressive Integrated Moving Average (ARIMA) model prediction on the data to predict the epidemiological trend of the prevalence and incidence of COVID-2019. The ARIMA model consists of an autoregressive model, moving average model, and seasonal autoregressive integrated moving average model. The ARIMA model parameters were by autocorrelation function. ARIMA (1,0,4) model was selected for the prevalence of COVID-2019 while ARIMA (1,0,3) was selected as the best ARIMA model for determining the incidence of COVID-19. The research predicted that if the virus does not develop any new mutations, the curve will flatten in the near future.

Researchers [46] investigated the serial interval of COVID-19 based on publicly reported cases. A total of 468 COVID-19 transmission events reported in China outside of Hubei Province between January 21, 2020, and February 8, 2020 formulated the data set. The data is compiled from reports of provincial disease control centers. The data indicated that in 59 of the cases, the infectee developed symptoms earlier than the infector indicated pre-symptomatic transmission. The mean serial interval is estimated to be 3.96 with a standard deviation of 4.75. The mean serial interval of COVID-19 is found to be lower than similar viruses of MERS and SARS. The production rate (Ro) of the data set is found to be 1.32

Dey et al. [47] analyzed the epidemiological outbreak of COVID-19 using a visual exploratory data analysis approach. The authors utilized publicly available data sets from WHO, the Chinese Center for Disease Control and Prevention, and Johns Hopkins University for cases between 22 January 2020 to 16 February 2020 all around the globe. The data set consisted of time series information regarding the number of cases, origin country, recovered cases, etc. The main

[19] https://www.arcgis.com/apps/opsdashboard/index.html

[20] https://github.com/CSSEGISandData/COVID-19

[21] https://ncov.dxy.cn/ncovh5/view/pneumonia

[22] https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html

objective of the study is to provide time-series visual data analysis for the understandable outcome of the COVID-19 outbreak.

Researcher [48] investigated the transmission control measures of COVID-19 in China. The authors compiled and analyzed a unique data set consisting of case reports, mobility patterns, and public health intervention. The COVID-19 case data were collected from official reports of the health commission. The mobility data were collected from location-based services employed by Social media applications such as WeChat. The travel pattern from Wuhan during the spring festival was constructed from Baidu migration index [23]. The study found that the number of cases in other provinces after the shutdown of Wuhan can be strongly related to travelers from Wuhan. In cities with a lesser population, the Wuhan travel ban resulted in a delayed arrival (+2.91 days) of the virus. Cities that implemented the highest level emergency before the arrival of any case reported 33.3% lesser number of cases. The low level of peak incidences per capita in provinces other than Wuhan also indicates the effectiveness of early travel bans and other emergency measures. The study also estimated that without the Wuhan travel band and emergency measures, the number of COVID-19 cases outside Wuhan would have been around 740000 on the 50th day of the pandemic. In summary, the study found a strong association between the emergency measures introduced during spring holidays and the delay in epidemic growth of the virus. A global mobility data collected from Google location services can be found at Google [24] and Kaggle [25].

Liu et al. [49] formulated a spatio-temporal data set of COVID-19 cases in China on the daily and city levels. As the published health reports are in the Chinese language, the authors aim to facilitate researchers around the globe with data set translated to English. The data set also divides the cases to city/county level for analysis of city-wide pandemic spread contrary to other countries/province categorizations [26]. The data set consists of essential stats for academic research, such as daily new infections, accumulated infections, daily new recoveries, accumulated recoveries, daily new deaths, etc. Each of these statistics is compiled into a separate CSV file and made available on Github. The first two authors did cross-validation of their data extraction tasks to reduce the error rate.

Researcher [50] utilize reported death rates in South Korea in combination with population demographics for correction of under-reported COVID-19 cases in other countries. South Korea is selected as a benchmark due to its high testing capacity and well-documented cases. The author correlates the under-reported cases with limited sampling capabilities and bias in systematic death rate estimation. The author brings to service two datasets. One of the datasets is WHO statistics of daily country-wise COVID-19 reports. The second dataset is demographic database maintained by the UN. This dataset is limited from 2007 onwards and hosted on Kaggle for country wise analysis [27]. The adjustment in number of COVID-19 cases is achieved while comparing two countries and computing their Vulnerability Factor which is based on population ages and corresponding death rates. As a result, the Vulnerability Factor of countries with higher age population is greater than one leading to higher death rate estimations. A complete workflow of the analysis is also hosted on Kaggle [28].

Kraemer et al. [51] analyzed the effect of human mobility and travel restrictions on spread on COVID-19 in China. Real-time and historical mobility data from Wuhan and epidemiological data from each province were employed for the study (source: Baidu Inc.). The authors also maintain a list of cases in Hubei and a list of cases outside Hubei. The data and code can be found at [29]. The study found that before the implementation of travel restrictions, the spatial distribution of COVID-19 can be highly correlated to mobility. However, the correlation is lower after the imposition of travel restrictions. Moreover, the study also estimated that the late imposition of travel restrictions after the onset of the virus in most of the provinces would have lead to higher local transmissions. The study also estimated the mean incubation period to identify a time frame for evaluating early shifts in COVID-19 transmissions. The incubation period was estimated to be 5.1 days.

The aforementioned authors also list and maintain the epidemiological data of COVID-19 cases in China as a part of a separate study [8, 52]. The data set contains individual-level information of laboratory-confirmed cases obtained from city and provincial disease control centers. The information includes **(a)** key dates including the date of onset of disease, date of hospital admission, date of confirmation of infection, and dates of travel, **(b)** demographic information about the age and sex of cases, **(c)** geographic information, at the highest resolution available down to the district level, **(d)** symptoms, and **(e)** any additional information such as exposure to the Huanan seafood market. The data set is updated regularly. The aim of the open access line list data is to guide the public health decision-making process in the context of the COVID-19 pandemic.

[23] http://qianxi.baidu.com/

[24] https://www.google.com/covid19/mobility/

[25] https://www.kaggle.com/chaibapat/google-mobility

[26] https://coronavirus.jhu.edu/map.html

[27] https://www.kaggle.com/lachmann12/world-population-demographics-by-age-2019

[28] https://www.kaggle.com/lachmann12/correcting-under-reported-covid-19-case-numbers

[29] `https://github.com/Emergent-Epidemics/covid19_npi_china`

Killeen et al. [53] accounted for the county-level dataset of COVID-19 in the US. The machine-readable dataset contains more than 300 socioeconomic parameters that summarize population estimates, demographics, ethnicity, education, employment, and income among other healthcare system-related metrics. The data is obtained from the government, news, and academic sources. The authors obtain time-series data from [42] and augment it with activity data obtained from SafeGraph. SafeGraph is a digital footprint platform that aggregates location-based data from multiple applications. The journalistic data is used to infer the implementation of lock-down measures at the county level. The dataset is envisioned to serve the scientific community in general and ML applications specifically for epidemiological modeling.

Researchers [54] provide another study for evaluating the effects of travel restrictions on COVID-19 transmissions. The authors quantify the impact of travel restrictions in early 2020 with respect to COVID-19 cases reported outside China using statistical analysis. The authors obtained an epidemiological dataset of confirmed COVID-19 cases from government sources and websites. All confirmed cases were screened using RT-PCR. The quantification of COVID-19 transmission with respect to travel restrictions was carried out for the number of exported cases, the probability of a major epidemic, and the time delay to a major epidemic.

Lai et al. [55] quantitatively studied the effect of non-pharmaceutical interventions, i.e., travel bans, contact reductions, and social distancing on the COVID-19 outbreak in China. The authors modeled the travel network as Susceptible-exposed-infectious-removed (SEIR) model to simulate the outbreak across cities in China in a proposed model named Basic Epidemic, Activity, and Response COVID-19 model. The authors used epidemiological data in the early stage of the epidemic before the implementation of travel restrictions. This data was used to determine the effect of non-pharmaceutical interventions on onset delay in other regions with first case reports as an indication. The authors also obtained large scale mobility data from Baidu location-based services which report 7 billion positioning requests per day. Another historical dataset from Baidu was obtained for daily travel patterns during the Chinese new year celebrations which coincided with the COVID-19 outbreak. The study estimated that there were approximately 0.1 Million COVID-19 cases in China as of 29 February 2020. Without the implementation of non-pharmaceutical interventions, the cases were estimated to increase 67 fold. The impact of various restrictions was varied with early detection and isolation preventing more cases than the travel restrictions. In the case of a three-week early implementation of non-pharmaceutical interventions, the cases would have been 95% less. On the contrary, if the non-pharmaceutical interventions were implemented after a further delay of 3 weeks, the COVID-19 cases would have increased 18 times.

A study on a similar objective of investigating the impact of non-pharmaceutical interventions in European countries was carried out in [56]. At the start of pandemic spread in European countries, non-pharmaceutical interventions were implemented in the form of social distancing, banning mass gathering, and closure of educational institutes. The authors utilized a semi-mechanistic Bayesian hierarchical model to evaluate the impact of these measures in 11 European countries. The model assumes that any change in the reproductive number is the effect of non-pharmaceutical interventions. The model also assumed that the reproduction number behaved similarly across all countries to leverage more data across the continent. The study estimates that the non-pharmaceutical interventions have averted 59000 deaths up till 31 March 2020 in the 11 countries. The proportion of the population infected by COVID-19 is found to be highest in Spain followed by Italy. The study also estimated that due to mild and asymptomatic infections many fold low cases have been reported and around 15% of Spain population was infected in actual with a mean infection rate of 4.9%. The mean reproduction number was estimated to be 3.87. Real-time data was collected from ECDC (European Centre of Disease Control) for the study.

3.2 Social media data

Researchers [57] contributed towards a publicly available ground truth textual data set to analyze human emotions and worries regarding COVID-19. The initial findings were termed as Real World Worry Dataset (RWWD). In the current global crisis and lock-downs, it is very essential to understand emotional responses on a large scale. The authors requested multiple participants from UK on 6th and 7th April (Lock-down, PM in ICU) to report their emotions and formed a dataset of 5000 texts (2500 short and 2500 long texts). The number of participants was 2500. Each participant was required to provide a short tweet-sized text (max 240 characters) and a long open-ended text (min 500 characters). The participants were also asked to report their feelings about COVID-19 situations using 9-point scales (1 = not at all, 5 = moderately, 9 = very much). Each participant rated how worried they were about the COVID-19 situation and how much anger, anxiety, desire, disgust, fear, happiness, relaxation, and sadness they felt. One of the emotions that best represented their emotions was also selected. The study found that anxiety and worry were the dominant emotions. STM package from R was reported for topic modeling. The most prevalent topic in long texts related to the rules of lock-down and the second most prevalent topic related to employment and economy. In short texts, the most prominent topic was government slogans for lock-down.

Chen et al. [58] describe a multilingual coronavirus data set with the aim of studying online conversation dynamics. The social distancing measure has resulted in abrupt changes in the society with the public accessing social platforms

for information and updates. Such data sets can help identify rumors, misinformation, and panic among the public along with other sentiments from social media platforms. Using Twitter's streaming API and Tweepy, the authors began collecting tweets from January 28, 2020 while adding keywords and trending accounts incrementally in the search process. At the time of publishing, the data set consisted of over 50 million tweets and 450GB of raw data.

Authors in [59] collected a Twitter dataset of Arabic language tweets on COVID-19. The aim of the data set collection is to study the pandemic from a social perspective, analyze human behavior, and information spread with special consideration to Arabic speaking countries. The data set collection was started in March 2020 using twitter API and consists of more than 2,433,660 Arabic language tweets with regular additions. Arabic keywords were used to search for relevant tweets. Hydrator and TWARC tools are employed for retrieving the full object of the tweet. The data set stores multiple attributes of a tweet object including the ID of the tweet, username, hashtags, and geolocation of the tweet.

Researcher [5] analyzes a data set of tweets about COVID-19 to explore the policies and perceptions about the pandemic. The main objective of the study is to identify public response to the pandemic and how the response varies time, countries, and policies. The secondary objective is to analyze the information and misinformation about the pandemic is presented and transmitted. The dataset is collected using Twitter API and covers 22 January to 13 March 2020. The corpus contains 6,468,526 tweets based on different keywords related to the virus in multiple languages. The data set is being continuously updated. The authors propose the application of Natural Language Processing, Text Mining, and Network Analysis on the data set as their future work. Similar data sets of Twitter posts regarding COVID-19 can be found at Github [30] and Kaggle [31].

Zarei et al. [60] gather social media content from Instagram using hashtags related to COVID-19 (coronavirus, covid19, and corona etc.). The authors found that 58% of the social media posts concerning COVID-19 were in English Language. The authors proposed the application of fake new identification and social behavior analysis on their data set.

Sarker et al. [61] mined Twitter to analyze symptoms of COVID-19 from self-reported users. The authors identified 203 COVID-19 patients while searching Twitter streaming API with expressions related to self-report of COVID-19. The patients reported 932 different symptoms with 598 unique lexicons. The most frequently reported COVID-19 symptoms were fever (65%) and cough (56%). The reported symptoms were compared with clinical findings on COVID-19. It was found that anosmia (26%) and ageusia (24%) reported on Twitter were not found in the clinical studies.

3.3 Scholarly articles

The Allen Institute for AI with other collaborators started an initiative for collecting articles on COVID-19 research named CORD-19 [6]. The data set was initiated with 28K articles now contains more than 52k articles and 41k full texts. Multiple repositories such as PMC, BioRxiv, MedRxiv, and WHO were searched with queries related to COVID-19 ("COVID-19", "Coronavirus", "Corona virus", "2019-nCoV", etc.). Along with the article's data set, a metadata file is also provided which includes each article DOI and publisher among other information. The dataset is also divided into commercial and non-commercial subsets. The duplicate articles were clustered based on publication ID/DOI and filtered to remove duplication. Design challenges such as machine-readable text, copyright restrictions, and clean canonical metadata were considered while collecting data. The aim of the data set collection is to facilitate information retrieval, extraction, knowledge-based discovery, and text mining efforts focused on COVID-19 management policies and effective treatment. The dataset has been popular among the research community with more than 1.5 million views and more than 75k downloads. A competition at Kaggle based on information retrieval from the proposed data set is also active. On the other hand, several publishers have created separate sections for COVID-19 research and listed on their website.

Researchers [62] provided a scoping review of 65 research articles published before 31 January 2020 indicating early studies on COVID-19. The review followed a five-step methodological framework for the scoping review as proposed in [63]. The authors searched multiple online databases including bioRxiv, medRxiv, Google scholar, PubMed, CNKI, and WanFang Data. The searched terms included "nCoV", "2019 novel coronavirus", and "2019-nCoV" among others. The study found that approximately 90% of the published articles were in the English language. The largest proportion (38.5%) of articles studied the causes of COVID-19. Chinese authors contributed to most of the work (67.7%). The study also found evidence of virus origin from the Wuhan seafood market. However, specific animal association of the COVID-19 was not found. The most commonly reported symptoms were fever, cough, fatigue, pneumonia, and headache form the studies conduction clinical trails of COVID-19. The surveyed studies have reported masks, hygiene practices, social distancing, contact tracing, and quarantines as possible methods to reduce virus transmission. The article sources are available as supplementary resources with the article.

[30] https://github.com/BayesForDays/coronada

[31] https://www.kaggle.com/smid80/coronavirus-covid19-tweets

Researchers [12] detailed a systematic review and critical appraisal of prediction models for COVID-19 diagnosis and prognosis. Multiple publication repositories such as PubMed were searcher for articles that developed and validated COVID-19 prediction models. Out of the 2696 available titles and 27 studies describing 31 prediction models were included for the review. Three of these studies predicted hospital admission from pneumonia cases. Eighteen studies listed COVID-19 diagnostic models out of which 13 were ML-based. Ten studies detailed prognostic models that predicted mortality rates among other parameters. The critical analysis utilized PROBAST, a tool for risk and bias assessment in prediction models [64]. The analysis found that the studies were at high risk of bias due to poorly reported models. The study recommended that COVID-19 diagnosis and prognosis models should adhere to transparent and open-source reporting methods to reduce bias and encourage realtime application.

Researcher from Berkeley lab have developed a web search portal for dataset of scholarly articles on COVID-19 [32]. The data set is composed of several scholarly data sets including Wang et al. [6], LitCovid, and Elsevier Novel Coronavirus Information Center. The continuously expanding dataset contains approximately 60K articles with 16K specifically related to COVID-19. The search portal employs NLP to look for related articles on COVID-19 and also provides valuable insights regarding the semantic of the articles.

4 Comparison

In this section we provide a tabular and descriptive comparison of the surveyed open-source data sets. Table 1 presents the comparison of medical image data sets in terms of application, data type, and ML method in tabular form.

We can categorize and compare all of the listed works on their openness. Some of the works do not have data and code publicly available and it is difficult to validate their work [65]. Others have code or data publicly available [66]. Such studies are more relevant in the current pandemic for global actions concerning scientific research against COVID-19. On the other hand, some studies merge multiple data sets and mention the source of data but do not host it as a separate repository [67]. The highly relevant studies have made public both data and code [25, 53].

An equal number of reported works have utilized CT scans and X-ray images. However, segmentation techniques to identify infected areas have been only applied to CT scans [27]. Similarly, augmentation techniques to increase the size of the data set have been applied in one of the listed studies [40]. All of the works provided 2D CT scans except for one resource from the Coronacases Initiative [33]. Most of the COVID-19 diagnosis works employed CNNs for classification. Some of the works utilized transfer learning to further increase the accuracy of classification [26, 30]. Moreover, few works augmented CNNs with SVM for feature extraction and classification tasks [40, 37]. Higher accuracies were reported from works augmenting transfer learning and SVM with CNNs. CNNs and deep learning models are reported to overfit models due to the limited size of the dataset [25]. Therefore, authors also researched alternative approaches in the form of Capsule network [38] and SVM [37] for better classification on limited data sets of COVID-19 cases. Most of the COVID-19 diagnosis works distinguished between two outcomes of COVID-19 positive or negative cases [36]. However, some of the works utilized three outcomes, i.e., COVID-19 positive, viral pneumonia, and normal cases for applicability in real-world scenarios. Researchers [30] expanded the classification to six common types of pneumonia. Such methodologies require the extraction and compilation of data sets with other categories of pneumonia radiographs. The ML-based COVID-19 diagnosis is difficult to fully automate as a human in the loop (HITL) is required to label radiographic data [27].

ResNet, MobileNet, and VGG have been commonly employed as pre-trained CNNs for classification [33, 34]. AI/ML explainability methods have been seldom used to delineate the performance of CNNs [32]. Most of the works report accuracies greater than 90% for COVID-19 diagnosis [40, 37]. The data sets of Cohen et al. [11] is considered pioneering effort and is mostly utilized for the COVID-19 cases and Kermany et al. [35] is employed for common pneumonia cases.

Table 2 presents the comparison of textual data-sets in terms of application, data type, and statistical method in tabular form.

The textual data sets are applied for multiple purposes, such as, **(a)** reporting and estimated new COVID-19 cases [42, 4], **(b)** estimating community transmission [43], **(c)** correlating the effect of mobility on virus transmissions [51], **(d)** estimating effect of non-pharmaceutical interventions on COVID-19 cases [55, 56], **(e)** learning emotional and socio-economic issue from social media [57, 58], and **(f)** analyzing scholarly publications for semantics [62]. Most of the articles apply statistical techniques (stochastic, Bayesian, and regression) for estimation and correlation [56, 45]. There is scope for the application of AI/ML technique as proposed in some studies [5, 53]. However, only statistical techniques have been applied to textual data sets in most of the listed works. Most of the studies that estimate COVID-19

[32] https://covidscholar.org
[33] https://coronacases.org/

Table 1: Comparison of COVID-19 medical image data sets

Study	Application	Data Type	Machine Learning	Link
[11]	COVID-19 diagnosis	X-ray or CT Scan	Proposed Deep and transfer learning	https://github.com/ieee8023/covid-chestxray-dataset
[25]	COVID-19 diagnosis	CT scans	Deep Convolutional network	https://github.com/UCSD-AI4H/COVID-CT
[26]	COVID-19 diagnosis	CT scans	Deep Convolutional network, Transfer learning	https://ai.nscc-tj.cn/thai/deploy/public/pneumonia_ct
[27]	COVID-19 infected area segmentation	Segmented CT scans	Deep Convolutional Network	NA
[28]	COVID-19 infected area segmentation	Segmented CT scans	NA	https://zenodo.org/record/3757476
Medical segmentation	COVID-19 infected area segmentation	Segmented CT scans	U-Net model	http://medicalsegmentation.com/covid19/
Coronacases Initiative	COVID-19 diagnosis	3D CT scans	NA	https://coronacases.org/
BSTI	COVID-19 diagnosis and reference	Miscellaneous	NA	https://www.bsti.org.uk/training-and-education/covid-19-bsti-imaging-database/
SIRM	COVID-19 diagnosis and reference	Miscellaneous	NA	https://www.sirm.org/en/category/articles/covid-19-database/
Radiopaedia	COVID-19 diagnosis and reference	Miscellaneous	NA	https://radiopaedia.org/articles/covid-19-3
[31]	COVID-19 diagnosis	X-ray images	Deep Convolutional network, transfer learning	https://github.com/lindawangg/COVID-Net
[33]	COVID-19 diagnosis	X-ray	Deep learning	https://github.com/ieee8023/covid-chestxray-dataset
[34]	COVID-19 diagnosis	X-ray	CNN and transfer learning	https://github.com/ieee8023/covid-chestxray-dataset + [35] + Kaggle convid19-X-rays
[30]	COVID-19 diagnosis, extract biomarkers	X-ray	CNN and transfer learning	[11] + SIRM + RSNA + Radiopaedia + [35]
[36]	COVID-19 diagnosis	X-ray	CNN	[11] + https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia
[37]	COVID-19 diagnosis	X-ray	CNN + SVM	https://github.com/ieee8023/covid-chestxray-dataset + Kaggle + [35]
[38]	COVID-19 diagnosis	X-ray	Capsule network + Transfer learning	https://github.com/ShahinSHH/COVID-CAPS
[39]	COVID-19 diagnosis	X-ray	CNN	https://www.kaggle.com/tawsifurrahman/covid19-radiography-database
[40]	COVID-19 diagnosis	X-ray	CNN + SVM	Cohen et al. [11]

transmissions utilize COVID-19 case data collected from various governmental, journalistic, and academic sources [53]. The studies that have human emotions have utilized Twitter API to collect data [59, 5]. Studies estimating effect of non-pharmaceutical interventions on COVID-19 bring to service location/mobility, epidemiological, and demographic data [56]. The collection of scholarly articles have proposed potential of NLP techniques [6] while a demonstration of the same is available at COVIDScholar. However, the details of the semantic analysis algorithms applied by the COVIDScholar are not available. Github is the first choice of researchers to share open access data while Kaggle is seldom put to use [57, 36].

Most of the research on COVID-19 is currently not peer-reviewed and in the form of pre-prints. The COVID-19 pandemic is a matter of global concern and necessitates that any scientific work published should go through a rigorous review process. At the same time, the efficient diffusion of scientific research is also demanded. Therefore, this survey had to include pre-prints that are currently bypassing the review process to compile a comprehensive list of articles. The pre-prints contribute approximately 50% of the cited research in this survey. The credibility of surveyed work is supported by the open-source data sets and code accompanying the pre-prints.

5 Discussions and Future Challenges

The are multiple challenges to the open-source data sets and research in fight against COVID-19. The main challenges related to the theme of our article are, but not limited to, **(a)** forecasting COVID-19 cases and fatalities on city and

Table 2: Comparison of Textual COVID-19 data sets

Study	Application	Data Type	Statistical method	Link
[42]	Reporting global cases	COVID-19 cases	NA	https://github.com/CSSEGISandData/COVID-19
[4]	Estimating new cases	COVID-19 cases	stochastic transmission dynamic	https://github.com/adamkucharski/2020-ncov/
[43]	Community transmission	COVID-19 cases (dates)	Bayesian approach	https://github.com/aakhmetz/nCoVSerialInterval2020
[44]	Community transmission	COVID-19 cases	Expectation-maximization	https://github.com/carolinecolijn/ClustersCOVID19
[45]	COVID-19 spread	COVID-19 statistics	ARIMA	https://github.com/CSSEGISandData/COVID-19
[46]	Community transmission	COVID-19 cases (dates)	maximum likelihood fitting and the Akaike information criterion	https://github.com/MeyersLabUTexas/COVID-19
[47]	COVID-19 visual analysis	COVID-19 statistics	Exploratory data analysis	WHO + John Hopkins + Chinese Center for Disease Control and Prevention
[48]	COVID-19 transmission control analysis	COVID-19 statistics	regression analysis	https://github.com/huaiyutian/COVID-19_TCM-50d_China
[49]	COVID-19 city wise case analysis	COVID-19 statistics	NA	https://github.com/cheongsa/Coronavirus-COVID-19-statistics-in-China
[50]	Correcting under-reported cases	Reported case and world demographics	Statistical	https://www.kaggle.com/lachmann12/correcting-under-reported-covid-19-case-numbers
[51]	Mobility-transmission analysis	Mobility and epidemiological data	Statistical	https://github.com/Emergent-Epidemics/covid19_npi_china
[8]	Reporting China cases	Location and epidemiological data	NA	https://github.com/beoutbreakprepared/nCoV2019/tree/master/latest_data
[53]	US county level data	348 socioeconomic parameters	proposed ML for epidemiological analysis	https://github.com/JieYingWu/COVID-19_US_County-level_Summaries
[54]	Mobility-transmission analysis exported from China	epidemiological dataset	Statistical	http://www.mdpi.com/2077-0383/9/2/601/s1
[55]	Effect of non-pharmaceutical interventions on COVID-19 in China	Location and epidemiological data	NA	https://github.com/wpgp/BEARmod
[56]	Effect of non-pharmaceutical interventions on COVID-19 in Europe	Location and epidemiological data	semi-mechanistic Bayesian hierarchical model	https://github.com/ImperialCollegeLondon/covid19model/releases/tag/v1.0
[57]	Measuring emotions	Textual data	statistical analysis (correlation and regression)	https://github.com/ben-aaron188/covid19worry
[58]	Conversation dynamics	Tweets	NA	https://github.com/echen102/COVID-19-TweetIDs
[5]	Perception and policies	Tweets	Proposed NLP, data mining	https://github.com/lopezbec/COVID19_Tweets_Dataset
[59]	Societal issues	Tweets (arabic)	NA	https://github.com/SarahAlqurashi/COVID-19-Arabic-Tweets-Dataset
[60]	Fake new identification	Instagram posts	NA	https://github.com/kooshazarei/COVID-19-InstaPostIDs
[61]	COVID-19 symptoms identification	Tweets	Data mining	https://sarkerlab.org/covid_sm_data_bundle/
[6]	Collecting published articles on COVID-19	Published articles	Proposed data extraction, retrieval mining	https://www.semanticscholar.org/cord19/download
[62]	Analyzing published articles on COVID-19	Published articles	Statistical analysis	https://static-content.springer.com/esm/art%3A10.1186%2Fs40249-020-00646-x/MediaObjects/40249_2020_646_MOESM1_ESM.docx
[12]	Systematic review of COVID-19 diagnosis articles	Published articles	CHARM and PROBAST tools	https://osf.io/ehc47/
COVID Scholar	NLP based search portal	Published articles	NLP	https://covidscholar.org

county levels, **(b)** predicting transmission factors and incubation period, **(c)** estimated effect of existing preventive measures on COVID-19 infections and transmissions, **(d)** using natural language processing (NLP) to analyze public sentiments from social media, **(e)** applying NLP on scholarly articles to automatically infer scientific findings, **(f)** identifying key health (obesity, air pollution, etc) and social risk factors, **(g)** identifying demographics at more risk of infection from existing cases, and **(h)** ethical and social consideration of analyzing patients data. Multiple challenges are being hosted on Kaggle [34], [35], [36] and elsewhere on the Internet [37]. We detail the challenges in the following subsection.

5.1 Challenges to medical data

Most of the researchers studying image-based diagnoses of COVID-19 have emphasized that further accuracy is required for application of their methods in clinical practices. Contact-less work-flows need to be developed for AI-assisted COVID-19 screening and detection to keep medics safe from the infected patients [18, 68]. Moreover, researchers have also emphasized that the primary source of COVID-19 diagnosis remains the RT-PCR test and medical imaging services aim to aid the current shortage of test kits as a secondary diagnosis method [26, 69]. Furthermore, a patient with RT-PCR test positive can have a normal chest CT scan at the time of admission, and changes may occur only after more than two days of onset of symptoms [69]. Therefore, further analysis is required to establish a correlation between radiographics and PCR tests [70].

Data sets are available for most of the research directions in biomedical imaging. However, these data sets are limited in size for the application of deep learning techniques. Researchers have emphasized that larger data sets are required for deep learning algorithms to provide better insights and accuracy in diagnosis [40]. Detecting/screening COVID-19 from cough using ML techniques has indicated promising results [7, 71]. The accuracy of the study is hindered due to the small data set of COVID-19 cough samples. Several researchers are gathering cough based data and have made appeals for contribution from public [38] [39].

5.2 Challenges to textual data

Three reported studies collected scholarly articles related to COVID-19 [62, 6, 12]. However, the application of NLP is proposed in these works. The inference of scientific facts from published scholarly articles remains a challenge yet to be addressed in reference to COVID-19. The only resource available in this direction is the COVIDScholar developed by Berkeley Lab for semantic analysis of COVID-19 research. Similarly, data sets have been curated from social media platforms [59, 5]. The human emotions and psychology in the pandemic, sentiments regarding lock-downs, and other non-pharmaceutical interventions are yet to be investigated thoroughly. Another research direction is related to social distancing in the pandemic. Given the preferred social distance across multiple countries and the open-source data, what can be its effect on COVID-19? [72]. Moreover, data set curation is also needed to provide an update on preferred social distances during the COVID-19 lock-down initiatives.

Researcher [73] collected evidence of reduced air pollution in urban cities after implementation of lock-downs while comparing February 2019 and February 2020 air quality index (AQI). The study compared AQI of six cities and observed that the cities under lock-down in the evaluation time showed decreased air pollutants. The data was collected from World Air Quality Index Project and illustrated in the article. However, the study was conducted early when the COVID-19 was not declared as a pandemic and the actual effect of lock-downs on AQI could not be measured.

5.3 Privacy issues

Privacy issues are a concern regarding user mobility, medical, and social data. Privacy concerns are further escalated due to open-source nature of data. Google, Baidu, and SafeGraph have been identified as sources for mobility data in this survey. Users have concerns about large-scale governmental surveillance in case such data from applications is shared with a third party.

The automated contact tracing application initiated by several governments in the wake of COVID-19 transmissions also demands consideration of user privacy issues [74]. Automated contact tracing application monitor the user-user interactions with the help of Bluetooth communications. The population at risk can be identified if one user is diagnosed as COVID-19 positive from his user-user interactions automatically saved by the contact tracing application [75]. The

[34] https://www.kaggle.com/covid19
[35] https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge/tasks
[36] https://www.kaggle.com/roche-data-science-coalition/uncover/tasks
[37] https://www.covid19challenge.eu/
[38] https://www.covid-19-sounds.org/en/
[39] https://coughvid.epfl.ch/

contact tracing applications can be utilized for large-scale surveillance as user data is updated in a central repository frequently. It is yet to be debated the compliance of contact tracing applications with country-level health and privacy laws. Similarly, patient privacy concerns need to be addressed on the country level based on health and privacy laws and social norms [76].

Public hatred and discrimination have also been reported against COVID-19 patients and health workers [77]. The situation demands complete anonymity of medical and mobility data to avoid any discrimination generating from data shared for academic purposes. All of the open-source data either medical or textual regarding COVID-19 patients should be anonymized before sharing.

5.4 Misuse of Social Media

Although the digital technologies have significantly aided the combat against COVID-19, it has also provided ground for vulnerabilities that can be exploited in terms of social behaviors [78]. Fake news/misinformation sharing on social media platforms [20], racist hatred [21], propaganda (against 5G technologies and governments) [22], and online financial scams [23] are few forms of digital platform exploitation in COVID-19 pandemic. Fake news and rumors have been spread about lock-downs policies, over-crowded places, and death cases on social media platforms. Fake-news identification is already a popularly debated topic among the social and computer science community [79]. Existing NLP techniques for fake news identification need to be applied on COVID-19 social media data-sets for evaluation of proposed works in the current pandemic [80, 21]. The social media platforms also need to be analyzed for human perceptions and sentiments regarding specific ethnicity (for example, sinophobia) and lock-down policies [81]. The propaganda that 5G networks are responsible for COVID-19 spread has also received widespread attention on social media. With the society heavily relying on online shopping and banking transactions in the pandemic, increased number of online scamming and hacking activities have been reported [23].

6 Conclusion

This article provided a comprehensive review of COVID-19 open-source data sets. The review was organized on the type of data and application. Textual and medical image data formed the main data types. The application of open-source data set were COVID-19 diagnosis, infection estimation, mobility and demographic correlations, socio-economic analysis, and sentiment analysis. We found that although scientific research works on COVID-19 are growing exponentially, there is room for open-source data curation and extraction in multiple directions such as expanding of existing CT scan data sets for application of deep learning and compilation of cough samples. We compared the listed works on their openness, application, and ML/statistical methods. Moreover, we provided a discussion of future research directions and challenges concerning COVID-19 open-source data sets. Further work is required on **(a)** the curation of data set for cough based COVID-19 diagnosis, **(b)** expanding CT scan and X-ray data sets for higher accuracy of deep learning techniques, **(c)** establishment of privacy-preserving mechanisms for patient data, user mobility, and contact tracing, **(d)** contact-less diagnosis based on biomedical images to protect front-line health workers from infection, **(e)** sentiment analysis and fake new identification from social media for policy making, and **(f)** semantic analysis for automated fact-finding from scholarly articles to list a few. We advocate that the works listed in this survey based on open-source data and code are the way forward to extendable, transparent, and verifiable scientific research on COVID-19.

References

[1] World Health Organization. Coronavirus disease 2019 (covid-19): situation report, 102. 2020.

[2] Matt J Keeling, T Deirdre Hollingsworth, and Jonathan M Read. The efficacy of contact tracing for the containment of the 2019 novel coronavirus (covid-19). *medRxiv*, 2020.

[3] Stefano Boccaletti, William Ditto, Gabriel Mindlin, and Abdon Atangana. Modeling and forecasting of epidemic spreading: The case of covid-19 and beyond. *Chaos, Solitons, and Fractals*, 2020.

[4] Adam J Kucharski, Timothy W Russell, Charlie Diamond, Yang Liu, John Edmunds, Sebastian Funk, Rosalind M Eggo, Fiona Sun, Mark Jit, James D Munday, et al. Early dynamics of transmission and control of covid-19: a mathematical modelling study. *The lancet infectious diseases*, 2020.

[5] Christian E Lopez, Malolan Vasu, and Caleb Gallemore. Understanding the perception of covid-19 policies by mining a multilanguage twitter dataset. *arXiv preprint arXiv:2003.10359*, 2020.

[6] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William. Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan,

Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. Cord-19: The covid-19 open research dataset. 2020.

[7] Ali Imran, Iryna Posokhova, Haneya N Qureshi, Usama Masood, Sajid Riaz, Kamran Ali, Charles N John, and Muhammad Nabeel. Ai4covid-19: Ai enabled preliminary diagnosis for covid-19 from cough samples via an app. *arXiv preprint arXiv:2004.01275*, 2020.

[8] Bo Xu, Moritz UG Kraemer, and Data Curation Group. Open access epidemiological data from the covid-19 outbreak. *The Lancet. Infectious Diseases*, 2020.

[9] John Scott Frazer, Amelia Shard, and James Herdman. Involvement of the open-source community in combating the worldwide covid-19 pandemic: a review. *Journal of Medical Engineering & Technology*, pages 1–8, 2020.

[10] Ahmad Alimadadi, Sachin Aryal, Ishan Manandhar, Patricia B Munroe, Bina Joe, and Xi Cheng. Artificial intelligence and machine learning to fight covid-19, 2020.

[11] Joseph Paul Cohen, Paul Morrison, and Lan Dao. Covid-19 image data collection. *arXiv preprint arXiv:2003.11597*, 2020.

[12] Laure Wynants, Ben Van Calster, Marc MJ Bonten, Gary S Collins, Thomas PA Debray, Maarten De Vos, Maria C Haller, Georg Heinze, Karel GM Moons, Richard D Riley, et al. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. *bmj*, 369, 2020.

[13] Kun Lan, Dan-tong Wang, Simon Fong, Lian-sheng Liu, Kelvin KL Wong, and Nilanjan Dey. A survey of data mining and deep learning in bioinformatics. *Journal of medical systems*, 42(8):139, 2018.

[14] Mohammad Ali Humayun, Ibrahim A Hameed, Syed Muslim Shah, Sohaib Hassan Khan, Irfan Zafar, Saad Bin Ahmed, and Junaid Shuja. Regularized urdu speech recognition with semi-supervised deep learning. *Applied Sciences*, 9(9):1956, 2019.

[15] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[16] EJ Yates, LC Yates, and H Harvey. Machine learning "red dot": open-source, cloud, deep convolutional neural networks in chest radiograph binary normality classification. *Clinical radiology*, 73(9):827–831, 2018.

[17] Arni SR Srinivasa Rao and Jose A Vazquez. Identification of covid-19 can be quicker through artificial intelligence framework using a mobile phone-based survey in the populations when cities/towns are under quarantine. *Infection Control & Hospital Epidemiology*, pages 1–18, 2020.

[18] F. Shi, J. Wang, J. Shi, Z. Wu, Q. Wang, Z. Tang, K. He, Y. Shi, and D. Shen. Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for covid-19. *IEEE Reviews in Biomedical Engineering*, pages 1–1, 2020.

[19] Roman Kalkreuth and Paul Kaufmann. Covid-19: A survey on public medical imaging data resources. *arXiv preprint arXiv:2004.04569*, 2020.

[20] Gordon Pennycook, Jonathon McPhetres, Yunhao Zhang, and David Rand. Fighting covid-19 misinformation on social media: Experimental evidence for a scalable accuracy nudge intervention. 2020.

[21] Kazuki Shimizu. 2019-ncov, fake news, and racism. *The Lancet*, 395(10225):685–686, 2020.

[22] Inayat Ali. The covid-19 pandemic: Making sense of rumor and fear: Op-ed. *Medical Anthropology*, pages 1–4, 2020.

[23] Elisabeth Beaunoyer, Sophie Dupéré, and Matthieu J Guitton. Covid-19 and digital inequalities: Reciprocal impacts and mitigation strategies. *Computers in Human Behavior*, page 106424, 2020.

[24] Joseph Paul Cohen, Paul Bertin, and Vincent Frappier. Chester: A web delivered locally computed chest x-ray disease prediction system. *arXiv preprint arXiv:1901.11210*, 2019.

[25] Jinyu Zhao, Yichen Zhang, Xuehai He, and Pengtao Xie. Covid-ct-dataset: A ct scan dataset about covid-19. *arXiv preprint arXiv:2003.13865*, 2020.

[26] Shuai Wang, Bo Kang, Jinlu Ma, Xianjun Zeng, Mingming Xiao, Jia Guo, Mengjiao Cai, Jingyi Yang, Yaodong Li, Xiangfei Meng, et al. A deep learning algorithm using ct images to screen for corona virus disease (covid-19). *medRxiv*, 2020.

[27] Fei Shan+, Yaozong Gao+, Jun Wang, Weiya Shi, Nannan Shi, Miaofei Han, Zhong Xue, Dinggang Shen, and Yuxin Shi. Lung infection quantification of covid-19 in ct images with deep learning. *arXiv preprint arXiv:2003.04655*, 2020.

[28] Ma Jun, Ge Cheng, Wang Yixin, An Xingle, Gao Jiantao, Yu Ziqi, Zhang Minqing, Liu Xin, Deng Xueyuan, Cao Shucheng, Wei Hao, Mei Sen, Yang Xiaoyu, Nie Ziwei, Li Chen, Tian Lu, Zhu Yuntao, Zhu Qiongjie, Dong Guoqiang, and He Jian. COVID-19 CT Lung and Infection Segmentation Dataset, April 2020.

[29] V Rajinikanth, Nilanjan Dey, Alex Noel Joseph Raj, Aboul Ella Hassanien, KC Santosh, and N Raja. Harmony-search and otsu based system for coronavirus disease (covid-19) detection using lung ct scan images. *arXiv preprint arXiv:2004.03431*, 2020.

[30] Ioannis Apostolopoulos, Sokratis Aznaouridis, and Mpesiana Tzani. Extracting possibly representative covid-19 biomarkers from x-ray images with deep learning approach and image data related to pulmonary diseases. *arXiv preprint arXiv:2004.00338*, 2020.

[31] Linda Wang and Alexander Wong. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest radiography images. *arXiv preprint arXiv:2003.09871*, 2020.

[32] Zhong Qiu Lin, Mohammad Javad Shafiee, Stanislav Bochkarev, Michael St Jules, Xiao Yu Wang, and Alexander Wong. Explaining with impact: A machine-centric strategy to quantify the performance of explainability algorithms. *arXiv preprint arXiv:1910.07387*, 2019.

[33] Ezz El-Din Hemdan, Marwa A Shouman, and Mohamed Esmail Karar. Covidx-net: A framework of deep learning classifiers to diagnose covid-19 in x-ray images. *arXiv preprint arXiv:2003.11055*, 2020.

[34] Ioannis D Apostolopoulos and Tzani A Mpesiana. Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks. *Physical and Engineering Sciences in Medicine*, page 1, 2020.

[35] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018.

[36] Ali Narin, Ceren Kaya, and Ziynet Pamuk. Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks. *arXiv preprint arXiv:2003.10849*, 2020.

[37] Prabira Kumar Sethy and Santi Kumari Behera. Detection of coronavirus (covid-19) based on deep features and support vector machine. 5, 2020.

[38] Parnian Afshar, Shahin Heidarian, Farnoosh Naderkhani, Anastasia Oikonomou, Konstantinos N Plataniotis, and Arash Mohammadi. Covid-caps: A capsule network-based framework for identification of covid-19 cases from x-ray images. *arXiv preprint arXiv:2004.02696*, 2020.

[39] Muhammad EH Chowdhury, Tawsifur Rahman, Amith Khandakar, Rashid Mazhar, Muhammad Abdul Kadir, Zaid Bin Mahbub, Khandakar R Islam, Muhammad Salman Khan, Atif Iqbal, Nasser Al-Emadi, et al. Can ai help in screening viral and covid-19 pneumonia? *arXiv preprint arXiv:2003.13145*, 2020.

[40] Saddam Hussain and Asifullah Khan. Coronavirus disease analysis using chest x-ray images and a novel deep convolutional neural network, 04 2020.

[41] Huazhu Fu, Deng-Ping Fan, Geng Chen, and Tao Zhou. COVID-19 Imaging-based AI Research Collection. `https://github.com/HzFu/COVID19_imaging_AI_paper_list`.

[42] Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*, 2020.

[43] Hiroshi Nishiura, Natalie M Linton, and Andrei R Akhmetzhanov. Serial interval of novel coronavirus (covid-19) infections. *International journal of infectious diseases*, 2020.

[44] Lauren Tindale, Michelle Coombe, Jessica E Stockdale, Emma Garlock, Wing Yin Venus Lau, Manu Saraswat, Yen-Hsiang Brian Lee, Louxin Zhang, Dongxuan Chen, Jacco Wallinga, et al. Transmission interval estimates suggest pre-symptomatic spread of covid-19. *medRxiv*, 2020.

[45] Domenico Benvenuto, Marta Giovanetti, Lazzaro Vassallo, Silvia Angeletti, and Massimo Ciccozzi. Application of the arima model on the covid-2019 epidemic dataset. *Data in brief*, page 105340, 2020.

[46] Zhanwei Du, Xiaoke Xu, Ye Wu, Lin Wang, Benjamin J Cowling, and Lauren Ancel Meyers. The serial interval of covid-19 from publicly reported confirmed cases. *medRxiv*, 2020.

[47] Samrat Kumar Dey, Md Mahbubur Rahman, Umme Raihan Siddiqi, and Arpita Howlader. Analyzing the epidemiological outbreak of covid-19: A visual exploratory data analysis (eda) approach. *Journal of Medical Virology*, 2020.

[48] Huaiyu Tian, Yonghong Liu, Yidan Li, Chieh-Hsi Wu, Bin Chen, Moritz UG Kraemer, Bingying Li, Jun Cai, Bo Xu, Qiqi Yang, et al. An investigation of transmission control measures during the first 50 days of the covid-19 epidemic in china. *Science*, 2020.

[49] Wenyuan Liu, Peter Tsung-Wen Yen, and Siew Ann Cheong. Coronavirus disease 2019 (covid-19) outbreak in china, spatial temporal dataset. *arXiv preprint arXiv:2003.11716*, 2020.

[50] Alexander Lachmann. Correcting under-reported covid-19 case numbers. *medRxiv*, 2020.

[51] Moritz UG Kraemer, Chia-Hung Yang, Bernardo Gutierrez, Chieh-Hsi Wu, Brennan Klein, David M Pigott, Louis du Plessis, Nuno R Faria, Ruoran Li, William P Hanage, et al. The effect of human mobility and control measures on the covid-19 epidemic in china. *Science*, 2020.

[52] Bo Xu, Bernardo Gutierrez, Sumiko Mekaru, Kara Sewalk, Lauren Goodwin, Alyssa Loskill, Emily L Cohn, Yulin Hswen, Sarah C Hill, Maria M Cobo, et al. Epidemiological data from the covid-19 outbreak, real-time case information. *Scientific data*, 7(1):1–6, 2020.

[53] Benjamin D Killeen, Jie Ying Wu, Kinjal Shah, Anna Zapaishchykova, Philipp Nikutta, Aniruddha Tamhane, Shreya Chakraborty, Jinchi Wei, Tiger Gao, Mareike Thies, et al. A county-level dataset for informing the united states' response to covid-19. *arXiv preprint arXiv:2004.00756*, 2020.

[54] Asami Anzai, Tetsuro Kobayashi, Natalie M Linton, Ryo Kinoshita, Katsuma Hayashi, Ayako Suzuki, Yichi Yang, Sung-mok Jung, Takeshi Miyama, Andrei R Akhmetzhanov, et al. Assessing the impact of reduced travel on exportation dynamics of novel coronavirus infection (covid-19). *Journal of clinical medicine*, 9(2):601, 2020.

[55] Shengjie Lai, Nick W Ruktanonchai, Liangcai Zhou, Olivia Prosper, Wei Luo, Jessica R Floyd, Amy Wesolowski, Chi Zhang, Xiangjun Du, Hongjie Yu, et al. Effect of non-pharmaceutical interventions for containing the covid-19 outbreak: an observational and modelling study. *medRxiv*, 2020.

[56] Seth Flaxman, Swapnil Mishra, Axel Gandy, H Unwin, H Coupland, T Mellan, H Zhu, T Berah, J Eaton, P Perez Guzman, et al. Report 13: Estimating the number of infections and the impact of non-pharmaceutical interventions on covid-19 in 11 european countries. 2020.

[57] Bennett Kleinberg, Isabelle van der Vegt, and Maximilian Mozes. Measuring emotions in the covid-19 real world worry dataset. *arXiv preprint arXiv:2004.04225*, 2020.

[58] Emily Chen, Kristina Lerman, and Emilio Ferrara. Covid-19: The first public coronavirus twitter dataset. *arXiv preprint arXiv:2003.07372*, 2020.

[59] Sarah Alqurashi, Ahmad Alhindi, and Eisa Alanazi. Large arabic twitter dataset on covid-19. *arXiv preprint arXiv:2004.04315*, 2020.

[60] Koosha Zarei, Reza Farahbakhsh, Noel Crespi, and Gareth Tyson. A first instagram dataset on covid-19. *arXiv preprint arXiv:2004.12226*, 2020.

[61] Abeed Sarker, Sahithi Lakamana, Whitney Hogg-Bremer, Angel Xie, Mohammed Ali Al-Garadi, and Yuan-Chi Yang. Self-reported covid-19 symptoms on twitter: An analysis and a research resource. *medRxiv*, 2020.

[62] Sasmita Poudel Adhikari, Sha Meng, Yu-Ju Wu, Yu-Ping Mao, Rui-Xue Ye, Qing-Zhi Wang, Chang Sun, Sean Sylvia, Scott Rozelle, Hein Raat, et al. Epidemiology, causes, clinical manifestation and diagnosis, prevention and control of coronavirus disease (covid-19) during the early outbreak period: a scoping review. *Infectious diseases of poverty*, 9(1):1–12, 2020.

[63] Hilary Arksey and Lisa O'Malley. Scoping studies: towards a methodological framework. *International journal of social research methodology*, 8(1):19–32, 2005.

[64] Robert F Wolff, Karel GM Moons, Richard D Riley, Penny F Whiting, Marie Westwood, Gary S Collins, Johannes B Reitsma, Jos Kleijnen, and Sue Mallett. Probast: a tool to assess the risk of bias and applicability of prediction model studies. *Annals of internal medicine*, 170(1):51–58, 2019.

[65] Mingzhi Li, Pinggui Lei, Bingliang Zeng, Zongliang Li, Peng Yu, Bing Fan, Chuanhong Wang, Zicong Li, Jian Zhou, Shaobo Hu, et al. Coronavirus disease (covid-19): spectrum of ct findings and temporal progression of the disease. *Academic radiology*, 2020.

[66] Lin Li, Lixin Qin, Zeguo Xu, Youbing Yin, Xin Wang, Bin Kong, Junjie Bai, Yi Lu, Zhenghan Fang, Qi Song, et al. Artificial intelligence distinguishes covid-19 from community acquired pneumonia on chest ct. *Radiology*, page 200905, 2020.

[67] Ophir Gozes, Maayan Frid-Adar, Hayit Greenspan, Patrick D Browning, Huangqi Zhang, Wenbin Ji, Adam Bernheim, and Eliot Siegel. Rapid ai development cycle for the coronavirus (covid-19) pandemic: Initial results for automated detection & patient monitoring using deep learning ct image analysis. *arXiv preprint arXiv:2003.05037*, 2020.

[68] Becky McCall. Covid-19 and artificial intelligence: protecting health-care workers and curbing the spread. *The Lancet Digital Health*, 2(4):e166–e167, 2020.

[69] Wenjie Yang and Fuhua Yan. Patients with rt-pcr confirmed covid-19 and normal chest ct. *Radiology*, page 200702, 2020.

[70] Tao Ai, Zhenlu Yang, Hongyan Hou, Chenao Zhan, Chong Chen, Wenzhi Lv, Qian Tao, Ziyong Sun, and Liming Xia. Correlation of chest ct and rt-pcr testing in coronavirus disease 2019 (covid-19) in china: a report of 1014 cases. *Radiology*, page 200642, 2020.

[71] Björn W Schuller, Dagmar M Schuller, Kun Qian, Juan Liu, Huaiyuan Zheng, and Xiao Li. Covid-19 and computer audition: An overview on what speech & sound analysis could contribute in the sars-cov-2 corona crisis. *arXiv preprint arXiv:2003.11117*, 2020.

[72] Agnieszka Sorokowska, Piotr Sorokowski, Peter Hilpert, Katarzyna Cantarero, Tomasz Frackowiak, Khodabakhsh Ahmadi, Ahmad M Alghraibeh, Richmond Aryeetey, Anna Bertoni, Karim Bettache, et al. Preferred interpersonal distances: a global comparison. *Journal of Cross-Cultural Psychology*, 48(4):577–592, 2017.

[73] Marc Cadotte. Early evidence that covid-19 government policies reduce urban air pollution. 2020.

[74] Justin Chan, Shyam Gollakota, Eric Horvitz, Joseph Jaeger, Sham Kakade, Tadayoshi Kohno, John Langford, Jonathan Larson, Sudheesh Singanamalla, Jacob Sunshine, et al. Pact: Privacy sensitive protocols and mechanisms for mobile contact tracing. *arXiv preprint arXiv:2004.03544*, 2020.

[75] Joel Hellewell, Sam Abbott, Amy Gimma, Nikos I Bosse, Christopher I Jarvis, Timothy W Russell, James D Munday, Adam J Kucharski, W John Edmunds, Fiona Sun, et al. Feasibility of controlling covid-19 outbreaks by isolation of cases and contacts. *The Lancet Global Health*, 2020.

[76] Hyunghoon Cho, Daphne Ippolito, and Yun William Yu. Contact tracing mobile apps for covid-19: Privacy considerations and related trade-offs. *arXiv preprint arXiv:2003.11511*, 2020.

[77] Jun He, Leshui He, Wen Zhou, Xuanhua Nie, and Ming He. Discrimination and social exclusion in the outbreak of covid-19. *International Journal of Environmental Research and Public Health*, 17(8):2933, 2020.

[78] Michael Chary, Nicholas Genes, Christophe Giraud-Carrier, Carl Hanson, Lewis S Nelson, and Alex F Manini. Epidemiology from tweets: estimating misuse of prescription opioids in the usa from social media. *Journal of Medical Toxicology*, 13(4):278–286, 2017.

[79] Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(3):1–42, 2019.

[80] Graeme D Smith, Fowie Ng, and William Ho Cheung Li. Covid-19: Emerging compassion, courage and resilience in the face of misinformation and adversity. *Journal of Clinical Nursing*, 29(9-10):1425, 2020.

[81] Delan Devakumar, Geordan Shannon, Sunil S Bhopal, and Ibrahim Abubakar. Racism and discrimination in covid-19 responses. *The Lancet*, 395(10231):1194, 2020.