Certain Investigations on the Application of Machine learning Algorithms and Deep Learning Architectures for Covid -19 Diagnosis

R.S.Soundariya[*], R.M.Tharsanee[**], B.Vishnupriya[***], R.Ashwathi[****], M.Nivaashini[*****]

[*]Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Sathyamangalam, Tamilnadu, India.

[**]Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Sathyamangalam, Tamilnadu, India.

[***]Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Sathyamangalam, Tamilnadu, India.

[****]Department of Civil Engineering, Bannari Amman Institute of Technology, Sathyamangalam, Tamilnadu, India.

[*****]Department of Computer Science and Engineering, KPR Institute of Engineering and Technology, Sathyamangalam, Tamilnadu, India

[*] Email ID: rssoundariya@gmail.com ; Corresponding Author.

ABSTRACT

Corona virus disease (Covid - 19) has started to promptly spread worldwide from April 2020 till date, leading to massive death and loss of lives of people across various countries. In accordance to the advices of WHO, presently the diagnosis is implemented by Reverse Transcription Polymerase Chain Reaction (RT- PCR) testing, that incurs four to eight hours' time to process test samples and adds 48 hours to categorize whether the samples are positive or negative. It is obvious that laboratory tests are time consuming and hence a speedy and prompt diagnosis of the disease is extremely needed. This can be attained through several Artificial Intelligence methodologies for prior diagnosis and tracing of corona diagnosis. Those methodologies are summarized into three categories: (i) Predicting the pandemic

spread using mathematical models (ii) Empirical analysis using machine learning models to forecast the global corona transition by considering susceptible, infected and recovered rate. (iii) Utilizing deep learning architectures for corona diagnosis using the input data in the form of X-ray images and CT scan images. When X-ray and CT scan images are taken into account, supplementary data like medical signs, patient history and laboratory test results can also be considered while training the learning model and to advance the testing efficacy. Thus the proposed investigation summaries the several mathematical models, machine learning algorithms and deep learning frameworks that can be executed on the datasets to forecast the traces of COVID-19 and detect the risk factors of coronavirus.

Keyword: Covid – 19 diagnosis; Machine Learning; Covid – 19 datasets; Mathematical models; Deep Learning;

INTRODUCTION

In today's scenario Artificial Intelligence (AI) and its applications like Machine Learning (ML) and Deep Learning (DL) have been greatly employed in cracking infinite problems stretching from agriculture to healthcare and various sectors. During the global outburst, when clinicians and biomedical scientists were in the hectic discovery of appropriate vaccines and drugs to stop the pandemic, AI researchers had begun their investigation for Covid-19 diagnosis and prevention, as a measure to combat against the coronavirus. The remarkable applications of AI in healthcare comprise early diagnosis of disease, foreseeing risk factors of a disease and AI supported medical systems. The methodology behind such learning techniques is to create a model to take in and learn from its experience with no programming. DL being the enlarged version of ML, whose models are designed for automatic feature learning and expertise in handling enormous datasets with ideal outcomes.

The proposed survey is focused on mathematical models, machine learning algorithms and deep learning architectures that can be used for Covid − 19 diagnoses. The survey is organised as follows (1) explains the appropriate mathematical models that can be helpful to track the growth of coronavirus, (2) outlines the supervised and unsupervised algorithms that supports in diagnoses, (3) discuss about the deep learning architectures that aids in corona diagnosis, (4) discuss about the available datasets (5) summarizes the survey and explains the proposed AI architecture along with its comparison on various performance metrics.

MATHEMATICAL MODELS

Mathematical models play a vital role in realizing the spreading pattern of a disease, resulting in prediction outcomes by considering the possible factors for the pandemic spread, distribution strategy, and number of infected and recovered cases. These models are helpful in linking the biological development of virus and the developing dynamics of the virus among the inhabitants. The most common models like differential equations, regression models, and statistical models are widely utilized for disease diagnosis (Siettos CI & Russo L, 2013).

Probability can be said as the likelihood of an occurrence of an event, which can be utilized for disease diagnosis. Based on the disease probability, the physicians can take decisions regarding the spread of the disease. Statistical models like regression, time series analysis plays a significant role in disease diagnosis. Time series analysis is used to discover the trend or pattern of the data over certain intervals. Particularly, ARIMA (Autoregressive Integrated Moving Average) model can be used to investigate the linear/non-linear correlation of the dependent attributes / variables for improved understanding of the data. Susceptible Infected Recovered (SIR) models are traditional compartmental models in epidemiology that widely used for diagnosis of infectious diseases.

Table 1 Mathematical models for Covid diagnosis

Reference	Mathematical model used	Inference	Parameters considered
Arti M.K , 2020	Time series forecasting method	A mathematical model to detect the spread of COVID-19 was proposed which assumes a tree based structure with certain constraints	Parameters: social distancing and lock down Constraints: quarantined people and travel history of both infected and uninfected people
Li et al., 2020	Brown linear exponential smoothing, ARIMAX models, Multiple Time Series analysis	The results of exponential smoothing and ARIMAX model showed that early signs of virus transmission can be predicted in short term	-
Anwar et al., 2020	NSFD scheme (nonstandard finite difference) and Runge-Kutta fourth order method	Mathematically proved that the quarantine of the diseased person can lessen the danger of forthcoming corona virus spread	isolation class, recovery rate, mortality; infection rate
Ankarali et al., 2020	Logarithm cubic models	Validation techniques like logistic, power, quadratic, exponential, growth curves and cubic were applied to the logarithmic values, out of which cubic model outperforms the best with successful results	collective cases and aggregate death toll
He et al., 2020	Susceptible Exposed Infected Recovered (SEIR) model	A modified SEIR model was proposed to concentrate on additional parameters. And were estimated using Particle Swarm Optimization (PSO)	quarantine people, hospitalized and the one who are infectious with and without infection
Chen et al., 2020	Statistical analysis, Univariate analysis and LASSO regression.	The combination of semantic features, radiological features can be effectively utilized for Covid diagnosis	clinical and radiological features

The above table clearly states that statistical and mathematical models can be used for Covid diagnosis by considering certain factors like travel history of infected and uninfected people, quarantine people, confirmed cases and death rate, human to human contact. Additional parameters like climatic conditions, genetic dissimilarities can also be considered to improve the prediction. Also, the above models had considered medium sized data, like part of country or selected countries. These models when combined with machine learning

models, the number of parameters to be considered for diagnosis and size of data can be increased, thereby strengthening the prediction rate of the diagnostic models.

MACHINE LEARNING MODELS

Machine learning models can be classified into major categories: A model that can handle labelled data is called as Supervised learning, i.e. The model is made to learn from the features that has a label against itself. In contrary, unsupervised learning model by itself trains to develop meaningful insights from the available data by exploiting the properties that are similar between the features that are present in the input data. The algorithms that follow supervised learning technique are intended to solve regression as well as classification kind of problems. On the other hand, the algorithms under the unsupervised learning category are suitable for association and clustering problems. This study on the diagnosis of COVID-19 falls under the regression/classification problem type. The proposed analysis can be considered as a classification problem when the input set of features is used to classify between the COVID positive and negative cases. Similarly, for the regression problem the model may be utilized to foretell the virus transmission in the forthcoming days, count of the deaths/recovered individuals. The standard machine learning algorithms deployed to different datasets relevant to COVID-19 are discussed in the following section.

Support Vector Machine (SVM), the most popular supervised learning algorithm, can be utilized to categorize diseased and healthy individuals. The ideology of SVM algorithm is to find the most suitable hyperplane or decision boundary that classifies the class labels. Naïve Bayes (NB) is another standard supervised algorithm that can be widely used for disease prediction and classification. It works on the basis of Bayes Theorem that calculates the probability of the input data points associated with each of the class labels and the class label with maximum probability will be identified as the appropriate class label. Depending on the nature of input data points, Naïve Bayes can be considered for any of its three forms:

Gaussian NB, Multinomial NB and Bernoulli NB. These models are easy to implement and highly suitable for large data sets also, these models are identified to overtake extremely popular classification models. K Nearest Neighbor (KNN) algorithm can be suitable for problems related to classification and regression that calculates similarity measures for all the input data points. Decision trees, a category of supervised learning algorithms that forecasts simple decision rules in the form of tree structures. Decision trees have been effectively applied in clinical diagnosis to classify medical patients by their disease based on the health records.

Table 2 ML algorithms for Covid diagnosis

Reference	Inference	Classification algorithm
Khanday et al., 2020	Data (Text) from clinical reports were collected and categorized into 4 different classes: SARS, ARDS, Covid and combination of ARDS and Covid.	ML algorithms : logistic regression, multinomial NB, SVM and decision trees ensemble techniques : bagging and boosting
Elaziz et al., 2020	The diagnosis involved two datasets with high resolution images and the research was focused on orthogonal moments, a parallel framework was implemented in order to speed up the computation.	Feature extraction: FrMEMs (Fractional Multichannel Exponent Moments) Classification : KNN
Yao et al., 2020	Identified 28 features were as Covid related biomarkers for severely ill individuals	SVM, KNN, random forest and AdaBoost
Good man et al., 2020	Ensemble learning model was designed , that can be utilized as a screening tool in hospitals to spot Covid, when PCR testing is not possible	SVM, Random forest, logistic regression, Neural network stochastic gradient descent, XGBoost, and ADABoost.

The above discussed ML algorithms had shown a greater performance in classifying between Covid and Non-Covid samples. Yet there are certain limitations as ML lags in performance when massive datasets are considered, accuracy and performance are affected as it could not handle such datasets. Also, ML algorithms always require a human to feed the

knowledge for extracting and selecting the features. When a new input sample is fed to the ML model, it can classify based on the trained input samples and only the features in the training samples will be considered, it cannot think out of the trained samples. Such shortcomings can be eliminated using deep learning frameworks.

DEEP LEARNING ARCHITECTURES

Deep Learning (DL) is another broader discipline of AI and subset of ML, having a layered architecture for performing representation learning. The parallel competencies of human brain and distributed neuron architecture are amended in DL models and henceforth these architectures are Deep Learning models. The various deep learning architectures include Recurrent Neural Network (RNN), Convolution Neural Network (CNN), Long Short Term Memory networks (LSTM) and Restricted Boltzmann machines (RBM) .The following section explains the utilization of the above mentioned deep learning architectures on various real time Covid -19 datasets.

CNN is widely utilized for image related applications. They are the extension of MLPs, since MLP had limitations like dimensionality and scaling related issues for high resolution images, CNN came into existence. RNN is another DL architecture that can handle progressive data by accepting the present input and previously received inputs from the other layers, due to its internal memory. Hence it can memorize (save) the output of a specific layer and feed the saved output back to the input in order to predict the output of the layer.

In order to regulate the spread of Covid-19 pandemic, (Shuai et al., 2020) proposed a deep learning method to figure out the graphical features of Covid-19 for clinical diagnosis. This paper emphasized the use of transfer learning neural network model for accurate prediction of Covid-19 by the use of radio graphical features. The test results proved that the manual prediction using eye recognition was less accurate compared to the deep learning model used. A probabilistic classifier was proposed in (Ahmed et al., 2020), which classifies and predicts

the existence of Covid-19 by using composite hybrid feature extraction and stack hybrid classification system. The hybrid model is developed by combining the traditional statistical methods and available ML tools like weka for extracting the features from the input CT images. An automated CT image analysis tool designed in (Ophir Gozes, 2020) demonstrates the utilization of 2-Dimensional and 3-Dimensional deep learning models developed by combining existing AI models and clinical knowledge for Covid-19 detection and tracking. CT images were analyzed using deep learning algorithms to differentiate between patients affected by Coronavirus and those who are not affected by the disease.

In another work CT scans were investigated using deep learning algorithms by employing VB-Net neural networks to identify the infected portions of lungs in Covid-19 patients. The system was validated with data from 249 COVID-19 patients for training phase and 300 patients in testing phase. The quantitative estimation of the system proved to be highly accurate in performing automatic delineation of infected regions in CT scan images (Fei et al., 2020).

Authors in (Mucahid et al., 2020) applied machine learning methods to classify the CT images of abdominal portions to illustrate early stage detection of Covid -19. 150 CT images were taken into consideration for the formation of datasets into four categories based on the size of the patches. Performance analysis was carried out based on metrics such as precision, F-score metrics, specificity, sensitivity. Classification performance evaluation showed better accuracy for Grey-Level Size Zone Matrix feature extraction method.

A category of feed forward networks called as Extreme Learning Machine shortly said as ELM can also be used for Covid diagnosis. A different combination of DL techniques like LSTM (Long Short Term Memory) networks, GAN (Generative Adversarial networks), RNN and ELM was utilized for Covid diagnosis. These neural networks were helpful for choosing

the suitable forecast models with Covid preferred parameters using different categories of dataset (M. Jamshidi et al., 2020).

DL architectures are extremely good in handling huge volume of data and they can automatically learn features from the input samples. Due to its parallel structure, neural networks can perform the computations quicker as well as the results can be more accurate, irrespective of the dataset size. Also the failures of any nodes in any layer will not affect the entire computation; these are reasons to prefer DL architectures over traditional ML algorithms.

DATASETS FOR COVID – 19 DIAGNOSES

In order to implement any scientific methodology effectively, the core component that is required is data. The data that is used in the research works can be made available either as closed source or open source. When the data is as an open source, it becomes more usable, transparent, and easily verifiable. The data that is applied for COVID-19 diagnosis is available open source for use by any researcher to mitigate the process of virus detection more efficiently. When Artificial Intelligence techniques such as ML and DL are combined together along with the open source data sets, it becomes a powerful resource to find concrete solutions for detection of COVID-19 that can be implemented in healthcare centers all over the world. Since the disease is considered as a global outbreak, the data are collected across various countries and available in different repositories. The efficacy of the datasets lies in monitoring the geographical growth and spread of the disease and thereby helping healthcare workers in the prediction and prevention of the pandemic. The datasets are collected and available in different patterns like the country-wise record of corona patients, X-ray and CT images of the affected and social media. All such datasets are summarized and interpreted in the form of numeric and image datasets.

Table 3 Datasets for Covid diagnosis

Nature of the data	Reference	Dataset source
Textual Data	Dong et al., 2020	The portal developed by the researchers of John Hopkins University, were intended to interactive and provided real time data regarding the number of COVID positive cases, number of deaths that has occurred and count of the people who have recovered based on the geographic regions.
	Dey et al., 2020	Social media data
	Kleinberg et al., 2020	RWWD (Real World Worry Dataset) consisted of the details regarding the feedbacks given by people at UK during the lockdowns in the first week of April. The emotions rendered by the individuals were recorded as a dataset consisting of 5000 records in textual format.
	Banda et al., 2020	The dataset was built using the tweets posted by people all around the world counting to 424 million.
Imagery Data	Cohen et al., 2020	Collection of 125 images comprising of CT Scans and X-rays taken from different publication by other authors and online resources.
	Zhao et al., 2020	Dataset comprising of CT scans of about 275 patients collected from bioRxiv and medRxiv preprints

The packages used in Machine Learning and Deep Learning are most often open source and easy to implement. When research based on closed sources are implemented, it can only lead to developing patents regarding the innovations made but not provide the necessary solution required to tackle the pandemic situation. Therefore, open source datasets are key factors to hasten up the process of COVID-19 diagnosis by incorporating Artificial Intelligence techniques such as ML and DL for effective data analysis.

INFERENCE

The Regression Model is one of the well-known practices for prediction and to evaluate the relationship between a different set of variable, which include a variety of models like Linear, Multiple Linear Regression, Polynomial and Logistic Regression Model. The Logistic regression model is well suited for categorical data, as the prediction of COVID will have a continuous data; Linear Regression Model is a best approach. The future time prediction of

the COVID pandemic, understanding that the evolution will not be linear but an exponential, instead of applying the simple linear regression model, log transformation is required. To be more accurate instead of modelling the prediction using basic logistic regression and linear regression model, ARIMA model are well suited for Covid diagnosis in terms of numeric data. A simple linear regression technique can be applied to track the spread of Covid cases across countries. In deep learning, an important advantage is the ability to reuse the knowledge gained from one set of problems to other related problems. Most often, the training data of the real world applications will not have the similar feature space as the data that will be used in the testing phase. Different domains may be experienced during training and classification of the dataset. In such scenarios, transfer learning may be employed to improve the performance of the deep learning models. This technique involves using the same weights of one or more than one layers that were applied in a pre-trained network model to upcoming new models. The weights may be kept fixed or fine-tuned according to the requirement. This type of learning framework helps to reduce the time taken to train the deep learning model and thus results in a minimized generalization error.
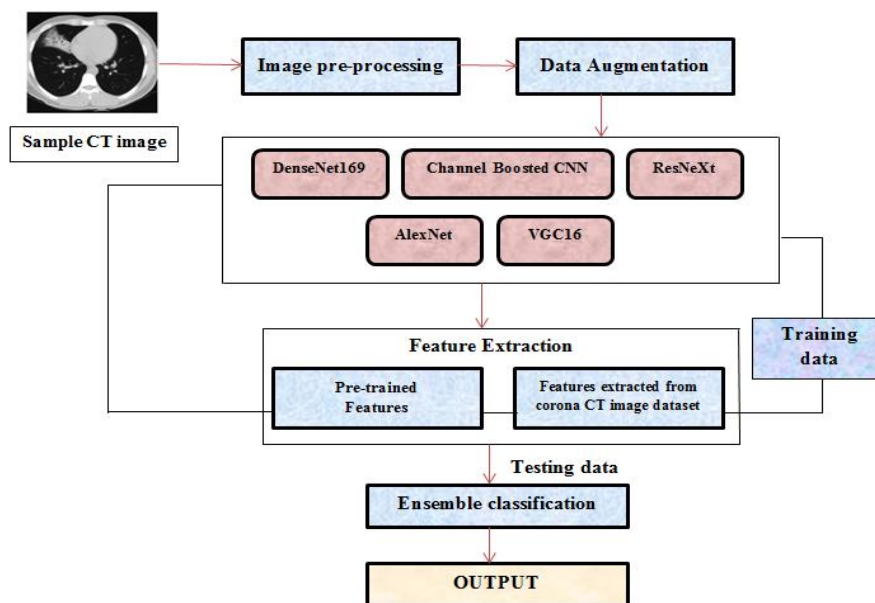
Figure 1 Proposed architecture for Covid-19 diagnosis

The proposed model can be validated against the models discussed in the literature. As discussed in many articles, Ensemble classifier outperforms the other models by producing better accuracy. The proposed ensemble learning model can be suitable for data processing of COVID-19 by procuring the essential radiological attributes of the CT images. Future research on this domain may be emphasized on implementing the proposed model with smaller datasets and can be extended to training the neural networks for larger dataset by incorporating more data augmentation techniques to achieve improved accuracy

CONCLUSION

In the present study, existing Convolutional Neural Network (CNN) models such as ResNeXt, Channel Boosted CNN, DenseNet, AlexNet and VGG 16 were repurposed to support in recognizing the presence of COVID-19 before they reach bulk scale. The dataset considered in the study encompasses of Computer Tomography (CT) images taken from 275 healthy individuals and 195 COVID-19 samples collected from 216 affected individuals. The proposed artificial intelligence based approach using deep learning models classifies COVID-19 affected cases by analyzing CT images and provides rapid detection of COVID-19 in a shorter time span. Ensemble classifier employed in this study is assumed to predict the presence of infection with a greater accuracy compared to the other similar works in the literature.

AI based diagnosis perspectives for Covid – 19 are openly explored in the proposed study. The reasonable supervised ML algorithms and deep learning framework that are most appropriate for corona diagnosis are evidently recognized and clarified with proper results. The analysis justifies the prominence of deciding an architecture that is not only excellent in feature learning but also scalable to enormous datasets. The proposed investigation clearly justifies the fact that AI related algorithms can fasten the Covid diagnosis process, but there are certain limitations that exist in dealing with real experiments and hence this level is still

under research and the novel techniques are yet to be identified. Though the predictive models can certainly offer excellent accuracy, the diagnosis process should accompanied by clinical experts.

REFERENCES

Siettos CI, Russo L., 2013. Mathematical modeling of infectious disease dynamics. Virulence. 2013;4(4):295-306. doi:10.4161/viru.24041

Arti M.K, 2020. Modeling and Predictions for COVID 19 Spread in India.

Li Y, Wang B, Peng R, Zhou C, Zhan Y, Liu Z, et al., 2020. Mathematical Modeling and Epidemic Prediction of COVID-19 and Its Significance to Epidemic Prevention and Control Measures. Ann Infect Dis Epidemiol. 5(1): 1052. ISSN: 2475-5664.

Anwar Zeb, Ebraheem Alzahrani, Vedat Suat Erturk, Gul Zaman,, 2020. Mathematical Model for Coronavirus Disease 2019 (COVID-19) Containing Isolation Class, BioMed Research International, Article ID 3452402, 7 pages.

Ankarali H, Ankarali S, Caskurlu H, et al., 2020. A Statistical Modeling of the Course of COVID-19 (SARS-CoV-2) Outbreak: A Comparative Analysis. Asia Pacific Journal of Public Health;32(4):157-160.

He, S., Peng, Y. & Sun, K., 2020. SEIR modeling of the COVID-19 and its dynamics. Nonlinear Dyn 101, 1667–1680

Chen X, Tang Y, Mo Y, et al., 2020. A diagnostic model for coronavirus disease 2019 (COVID-19) based on radiological semantic and clinical features: a multi-center study. 2020). Eur Radiol.; 30(9):

Khanday AMUD, Rabani ST, Khan QR, Rouf N, Mohi Ud Din M, 2020. Machine learning based approaches for detecting COVID-19 using clinical text data. Int J Inf Technol. 2020; 1-9.

Elaziz MA, Hosny KM, Salah A, Darwish MM, Lu S, Sahlol AT, 2020. New machine learning method for image-based diagnosis of COVID-19. PLoS ONE 15(6): e0235187. https://doi.org/ 10.1371/journal.pone.0235187

Yao, Haochen and Zhang, Nan and Zhang, Ruochi and Duan, Meiyu and Xie, Tianqi and Pan, Jiahui and Peng, Ejun and Huang, Juanjuan and Zhang, Yingli and Xu, Xiaoming and Xu, Hong and Zhou, Fengfeng and Wang, Guoqing, 2020. Severity Detection for the Coronavirus Disease 2019 (COVID-19) Patients Using a Machine Learning Model Based on the Blood and Urine Tests, Frontiers in Cell and Developmental Biology, 8, 683.

Goodman-Meza D, Rudas A, Chiang JN, et al. A machine learning algorithm to increase COVID-19 inpatient diagnostic capacity. PLoS One. 2020;15(9):e0239474. Published 2020 Sep 22. doi:10.1371/journal.pone.0239474

Shuai Wang1, Bo Kang, 2020. A deep learning algorithm using CT images to screen for CoronaVirus Disease (COVID-19). doi: https://doi.org/10.1101/2020.02.14.20023028.

Ahmed Abdullah Farid, Gamal Ibrahim Selim, 2020. A Novel Approach of CT Images Feature Analysis and Prediction to Screen for CoronaVirus Disease (COVID-19). International Journal of Scientific & Engineering Research. 1141-1149.

Ophir Gozes 2020. Rapid AI Development Cycle for the Coronavirus (COVID-19) Pandemic: Initial Results for Automated Detection & Patient Monitoring using Deep Learning CT Image Analysis. Radiology: Artificial Intelligence.

Fei Shan, Yaozong Gao, Jun Wang, Weiya Shi, Nannan Shi, Miaofei Han, Zhong Xue, Dinggang Shen, Yuxin Shi , 2020. Lung Infection Quantification of COVID-19 in CT Images with Deep Learning. Computer Vision and Pattern Recognition (cs.CV) ; Image and Video Processing (eess.iv) ; Quantitative methods (q-bio-QM). arXiv:2003.04655 [cs.CV]

Mucahid Barstugan,Umut Ozkaya , 2020. Coronavirus (COVID-19) Classification using CT Images by Machine Learning Methods. Computer Vision and Pattern Recognition; Machine Learning (cs.LG) ; Image and Video Processing (eess.iv) ; Machine Learning (stat.ML) arXiv:2003.09424 [cs.CV]

M. Jamshidi et al., 2020. Artificial Intelligence and COVID-19: Deep Learning Approaches for Diagnosis and Treatment," in IEEE Access, vol. 8, pp. 109581-109595.

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

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

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

Juan M. Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, and Gerardo Chowell, 2020. A large-scale covid-19 twitter chatter dataset for open scientific research – an international collaboration.

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

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